

## ANSWERS TO THE ASTRONOMY & ASTROPHYSICS PHD QUALIFICATION EXAMINATION

J.D. EMBERSON

*Draft version June 12, 2012*

### ABSTRACT

This document contains the answers to the set of one hundred questions contained in the Astronomy & Astrophysics PhD Qualification Examination test bank. Click on the following links to jump to the different sections and questions covered in this document:

- Cosmology: Questions 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, and 19.
- Extragalactic Astronomy: Questions 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, and 18.
- Galactic Astronomy: Questions 1, 2, 3, 4, 5/6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, and 18.
- Stars and Planets: Questions 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, and 22.
- Math and General Physics: Questions 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, and 23.

**QUESTION 1**

**What is recombination? At what temperature did it occur? How does this relate to the ionization potential of Hydrogen?**

## QUESTION 1

**What is recombination? At what temperature did it occur? How does this relate to the ionization potential of Hydrogen?**

Before moving further to discuss the physics of recombination, it is important to distinguish among three closely related (but not identical) moments in the history of the universe. First, the epoch of **recombination**, is the time at which the baryonic component of the universe goes from being ionized to being neutral. Rigorously, this is defined as the instant at which the number density of ions is equal to the number density of neutral atoms. Second, the epoch of **photon decoupling**, is the time at which the rate at which photons scatter from electrons becomes smaller than the Hubble parameter. When photons decouple, they cease to interact with the electrons, and the universe becomes transparent. Third, the epoch of **last scattering** is the time at which a typical CMB photon underwent its last scattering from an electron (Ryden 2002, pg. 186).

### RECOMBINATION

To a good approximation we can assume that hydrogen is the only neutral element to form during recombination. Although helium was present at this time, its inclusion is merely a complicating factor since all the relevant physics are pertinent to hydrogen. With this approximation, the degree to which the baryonic content of the universe is ionized can be expressed as the fractional ionization  $X$ , defined as

$$X \equiv \frac{n_p}{n_p + n_H} = \frac{n_p}{n_{\text{bary}}} = \frac{n_e}{n_{\text{bary}}}, \quad (1)$$

where subscript  $H$  denotes hydrogen in neutral form and subscript  $p$  denotes a free proton, or equivalent, ionized hydrogen. The last identity in equation (1) comes from the requirement of charge neutrality in the universe, that is  $n_e = n_p$  (Ryden 2002, pg. 187).

One useful consequence of assuming that hydrogen is the only element is that there is now a single relevant energy scale in the problem: the ionization energy of hydrogen,  $Q = 13.6$  eV. A photon with an energy  $h\nu > Q$  is capable of photoionizing a hydrogen atom:



The reverse reaction describes the process of radiative recombination where a free electron combines with a free proton to form neutral hydrogen H, with any excess energy carried away by a photon. In a universe containing protons, electrons, and photons, the fractional ionization  $X$  will depend on the balance between photoionization and radiative recombination (Ryden 2002, pg. 187).

Motivated by reaction (2) it is easy to make a crude approximation of the recombination temperature. Recombination, one could argue, must take place when the mean energy per photon of the CMB falls below the ionization energy of hydrogen,  $Q = 13.6$  eV. When this happens, the average CMB photon is no longer able to photoionize hydrogen. Since the mean CMB photon energy is  $\sim 2.7kT$ , this line of argument would indicate a recombination temperature of

$$T_{\text{rec}} \sim \frac{Q}{2.7k} \sim 6 \times 10^4 \text{ K}. \quad (3)$$

However, this crude approximation is a little *too* crude to be useful. In particular, it ignores the exponential tail of the blackbody spectrum trailing off at high energies. Although extremely high energy photons make up only a tiny fraction of the CMB photons, the total number of CMB photons is enormous; the baryon-to-photon ratio is  $\eta = n_{\text{bary}}/n_\gamma = 5.5 \times 10^{-10}$ . Hence, vast swarms of photons that surround every newly formed hydrogen atom greatly increase the probability that the atom will collide with a photon from the high-energy tail of the blackbody spectrum, and be photoionized (Ryden 2002, pg. 190).

Thus, we expect the recombination temperature to depend on  $\eta$  as well as on  $Q$ . The exact calculation of  $X$  as a function of time will therefore require a smattering of statistical mechanics. We will start by again focusing on reaction (2). While the photons are still coupled to the baryonic component, this reaction will be in statistical equilibrium, with the photoionization rate balancing the radiative recombination rate. When a reaction is in statistical equilibrium at a temperature  $T$ , the number density  $n_x$  of particles with mass  $m_x$  is given by the Maxwell Boltzmann equation

$$n_x = g_x \left( \frac{m_x k T}{2\pi\hbar^2} \right)^{3/2} \exp \left( -\frac{m_x c^2}{k T} \right), \quad (4)$$

as long as the particles are non-relativistic, with  $kT \ll m_x c^2$ . Here  $g_x$  is the statistical weight of the particle; for instance, electrons, protons, and neutrons (and their anti-particles as well) all have a statistical weight  $g_x = 2$ , corresponding to their two possible spin states. From the Maxwell Boltzmann equation for H,  $p$ , and  $e^-$ , we can construct an equation which relates the number densities of these particles:

$$\frac{n_H}{n_p n_e} = \frac{g_H}{g_p g_e} \left( \frac{m_H}{m_p m_e} \right)^{3/2} \left( \frac{k T}{2\pi\hbar^2} \right)^{-3/2} \exp \left[ \frac{(m_p + m_e - m_H)c^2}{k T} \right] \Rightarrow \frac{n_H}{n_p n_e} = \left( \frac{m_e k T}{2\pi\hbar^2} \right)^{-3/2} \exp \left( \frac{Q}{k T} \right). \quad (5)$$

The latter result arises by noting that  $n_p = n_e = 2$  and  $n_H = 4$ , setting  $m_H \approx m_p$ , and using the definition that  $Q = (m_p + m_e - m_H)c^2$ . Equation (5) is called the **Saha equation** (Ryden 2002, pg. 191).

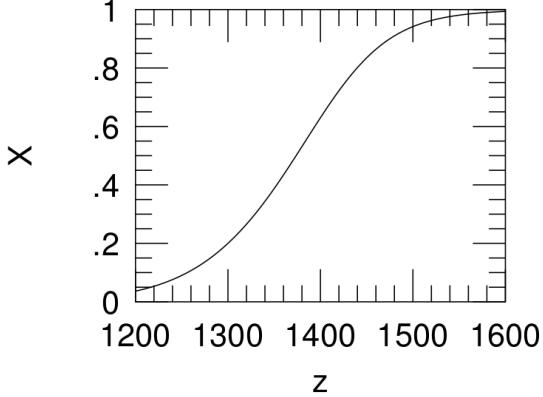


FIG. 1.— The fractional ionization  $X$  as a function of redshift during the epoch of recombination. Image taken from Ryden (2002).

The next step is to convert the Saha equation into a relation among  $X$ ,  $T$ , and  $\eta$ . From the definition of  $X$  in equation (1) we have that  $n_H = (1-X)n_p/X$ , and from charge neutrality, we can write equation (5) as

$$\frac{1-X}{X} = n_p \left( \frac{m_e k T}{2\pi \hbar^2} \right)^{-3/2} \exp\left(\frac{Q}{kT}\right). \quad (6)$$

To eliminate  $n_p$  from this equation we recall that  $\eta \equiv n_{\text{gary}}/n_\gamma$  so that  $\eta = n_p/Xn_\gamma$  for an ionized fraction of  $X$ . Since the photons have a blackbody spectrum, for which

$$n_\gamma = 0.243 \left( \frac{kT}{\hbar c} \right)^3, \quad \text{so that} \quad n_p = 0.243X\eta \left( \frac{kT}{\hbar c} \right)^3. \quad (7)$$

Substituting equation (7) into equation (6) yields

$$\frac{1-X}{X} = 3.84\eta \left( \frac{kT}{m_e c^2} \right)^{3/2} \exp\left(\frac{Q}{kT}\right). \quad (8)$$

If we define the moment of recombination as the exact instant when  $X = 1/2$ , then equation (8) can be solved to yield

$$kT_{\text{rec}} = 0.323 \text{ eV} = \frac{Q}{42} \rightarrow T_{\text{nuc}} \approx 3740 \text{ K}. \quad (9)$$

Because of the exponential dependence of the right-hand side of equation (8), the exact value of  $\eta$  doesn't strongly affect the value of  $T_{\text{rec}}$ . The redshift at which this event occurs is  $z_{\text{rec}} = 1370$ , when the age of the universe was roughly 0.24 Myr. Recombination was not an instantaneous process; however, as shown in Figure 1, it proceeded fairly rapidly (Ryden 2002, pg. 192).

#### PHOTON DECOUPLING

When the universe was fully ionized, photons interacted primarily with electrons, and the main interaction mechanism was Thomson scattering:



The scattering interaction is accompanied by a transfer of energy and momentum between the photon and electron. The cross-section for Thomson scattering is  $\sigma_e = 6.65 \times 10^{-29} \text{ m}^2$ . The rate at which this reaction proceeds is simply the photon speed ( $c$ ) divided by its mean free path:

$$\Gamma = \frac{c}{\lambda} = cn_e\sigma_e. \quad (11)$$

The photons remain coupled to the electrons as long as their scattering rate  $\Gamma > H$ ; this is equivalent to saying that their mean free path is shorter than the Hubble distance  $c/H$ . As long as photons scatter frequently from electrons, the photons remain in thermal equilibrium with the electrons (and, indirectly, with the protons due to Coulomb interactions). The photons, electrons, and protons, as long as they remain in thermal equilibrium, all have the same temperature  $T$ . When the photon scattering rate  $\Gamma$  drops below  $H$ , then the electrons are being diluted by expansion more rapidly than the photons can interact with them. The photons then decouple from the electrons and the universe becomes transparent. Afterward, the baryonic component of the universe is no longer compelled to have the same temperature as the photons (Ryden 2002, pg. 189).

If hydrogen remained ionized (and note the qualifying *if*), then photons would have remained coupled to the electrons and protons until a relatively recent time. Taking into account the transition from a radiation-dominated to a matter-dominated universe, and the resulting change in the expansion rate, we can compute that if hydrogen had remained fully ionized, then decoupling would have taken place at  $z \approx 42$  with  $T_\gamma \approx 120 \text{ K}$ . However, at such a low temperature, the CMB photons are too low in energy to keep the hydrogen ionized. Thus, the decoupling of photons is not a gradual process, caused by the

TABLE 1  
EVENTS IN THE EARLY UNIVERSE.

Event	$z$	$T_\gamma$ (K)	$t$ (Myr)
Matter-radiation Equality	3175	9000	0.05
Recombination	1370	3740	0.24
Photon Decoupling	1100	3000	0.35
Last Scattering	1100	3000	0.35

continuous lowering of free electron density as the universe expands. Rather, it is a relatively sudden process, caused by the abrupt plummeting of free electron density during the epoch of recombination, as electrons combine with protons to form hydrogen atoms (Ryden 2002, pg. 189).

Since the number density of free electrons drops rapidly during the epoch of recombination, the time of photon decoupling comes soon after the time of recombination. The rate of photon scattering, when the hydrogen is partially ionized, is

$$\Gamma(z) = n_e(z)\sigma_e c = X(z)(1+z)^3 n_{\text{bary},0} \sigma_e c. \quad (12)$$

While recombination is taking place, the universe is matter-dominated, so the Hubble parameter is given by the relation

$$\frac{H^2}{H_0^2} = \Omega_{\text{m},0}(1+z)^3. \quad (13)$$

The redshift of photon decoupling is found by setting equating (12) and (13), which yields  $z_{\text{dec}} = 1130$ . In truth, the exact redshift of photon decoupling is somewhat smaller than this value. The Saha equation assumes that the reaction (2) is in equilibrium. However, when  $\Gamma$  starts to drop below  $H$ , the photoionization reaction is no longer in equilibrium. As a consequence, at redshifts smaller than  $\sim 1200$ , the fractional ionization  $X$  is larger than would be predicted by the Saha equation, and the decoupling of photons is therefore delayed. More detailed calculations give  $z_{\text{dec}} = 1100$ , when the universe was 0.35 Myr in age.

The epoch of photon decoupling marked an important change in the state of the universe. Before photon decoupling, there existed a single photon-baryon fluid, consisting of photons, electrons, and protons coupled together. Since the photons traveled about at the speed of light, kicking the electrons before them as they went, they tended to smooth out any density fluctuations in the photon-baryon fluid smaller than the horizon. After photon decoupling, however, the photon-baryon fluid became a pair of gases, one of photons and the other of neutral hydrogen. Although the two gases coexisted spatially, they were no longer coupled together. Thus, instead of being kicked to and fro by the photons, the hydrogen gas was free to collapse under its own self-gravity (and the added gravitational attraction of the dark matter). Thus, when we observe the CMB we are looking backward in time to an important epoch when the baryons, free from the harassment of photons, were free to collapse gravitationally, forming the first structures in the universe (Ryden 2002, pg. 196).

Note, however, that the baryon temperature remained nearly equal to the CMB temperature up until  $z \sim 300$  due to residual ionization that allowed an exchange of energy between matter and radiation via Compton diffusion. For  $z \lesssim 300$ , the thermal interaction between baryons and radiation becomes insignificant, so that the matter component cools adiabatically with  $T_{\text{mat}} \propto a^{-2}$  (Coles & Lucchin 2002, pg. 196). This power-law relation arises since, for adiabatically cooling gas,  $T \propto \rho^{\gamma-1}$ , and since the majority of baryons are in the form of monatomic hydrogen  $\gamma = 5/3$ , the relation  $\rho \propto a^{-3}$ , implies that  $T \propto a^{-2}$ .

#### LAST SCATTERING

When we examine the CMB with our microwave antennas, the photons we collect have been traveling straight toward us since the last time they scattered from a free electron. During a brief time interval  $t \rightarrow t + dt$ , the probability that a photon undergoes a scattering is  $dP = \Gamma(t)dt$ , where  $\Gamma(t)$  is the scattering rate at time  $t$ . Thus, if we detect a CMB photon at time  $t_0$ , the expected number of scatterings it has undergone since an earlier time  $t$  is

$$\tau(t) = \int_t^{t_0} \Gamma(t)dt, \quad (14)$$

where  $\tau$  is the optical depth. The time  $t$  for which  $\tau = 1$  is the time of last scattering, and represents the time which has elapsed since a typical CMB photon last scattered from a free electron (Ryden 2002, pg. 194).

Integrating equation (14) over  $z$  while using equations (12) and (13) will give us  $z_{\text{ls}}$ . As it turns out, the last scattering of a typical CMB photon occurs after reaction (2) falls out of equilibrium, so the Saha equation doesn't strictly apply. To sufficient accuracy for our purposes, we can state that the redshift of last scattering was comparable to the redshift of photon decoupling:  $z_{\text{ls}} \approx z_{\text{dec}} \approx 1100$ . Obviously, not all the CMB photons underwent their last scattering simultaneously, so the "last scattering surface" is more like a "last scattering layer" with some depth (Ryden 2002, pg. 195).

The relevant times of various events around the time of recombination are shown in Table 1. For purposes of comparison, the table also contains the time of **matter-radiation equality**, emphasizing the fact that recombination, photon decoupling, and last scattering took place when the universe was matter-dominated. The redshift at which matter-radiation equality takes place,  $z_{\text{eq}}$ , is important for the generation of large-scale structure and development of CMB anisotropies, because perturbations grow at different rates in the two different eras. The epoch is found by setting

$$\frac{\rho_{\text{rad}}(t)}{\rho_{\text{mat}}(t)} = \frac{\rho_{\text{rad},0}}{\rho_{\text{mat},0}} \frac{1}{a(t)} \Rightarrow z_{\text{eq}} = \frac{\Omega_{\text{mat}}}{\Omega_{\text{rad}}} - 1, \quad (15)$$

which, with  $\Omega_m = 0.27$  and  $\Omega_r = 8.5 \times 10^{-5}$ , yields  $z_{\text{eq}} = 3175$  (Dodelson 2003, pg. 51).

**QUESTION 2**

The universe is said to be “flat”, or, close to flat. What are the properties of a flat universe and what evidence do we have for it?

## QUESTION 2

The universe is said to be “flat”, or, close to flat. What are the properties of a flat universe and what evidence do we have for it?

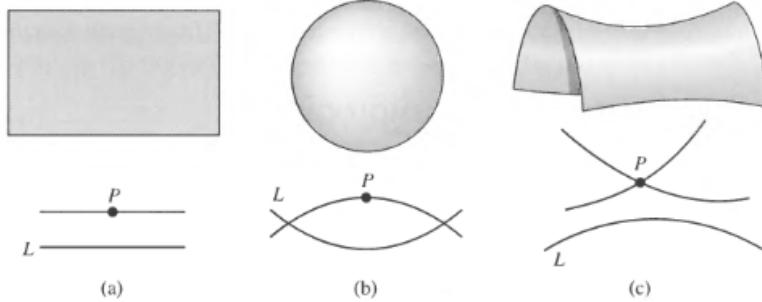


FIG. 2.— The parallel postulate, illustrated for three alternative geometries: (a) Euclidean or flat, (b) elliptic or closed, and (c) hyperbolic or open. Image taken from Carroll & Ostlie (2007).

### GEOMETRY

The appearance of objects at cosmological distances is affected by the curvature of the spacetime through which the light travels on its way to Earth. There are three options for the overall geometry of the universe: **Euclidean**, **elliptic**, or **hyperbolic**. These three geometries are mathematically independent and self-consistent, and are distinguished based on how they handle parallel lines. In particular, the parallel postulate for each geometry is:

- Euclidean: Given, in a plane, a line L and a point P not on L, then through P there exists *one and only one line* parallel to L.
- Elliptic: Given, in a plane, a line L and a point P not on L, then through P there exists *no line* parallel to L.
- Hyperbolic: Given, in a plane, a line L and a point P not on L, then through P there exist *at least two lines* parallel to L.

These statements are shown schematically in Figure 2. As we can see, Euclidean geometry describes a plane where the interior angles of triangles add up to  $180^\circ$  and the circumference of a circle is  $2\pi r$ . Elliptic geometry, on the other hand, describes the surface of a sphere where two lines that btw start out perpendicular to the sphere’s equator meet at its poles. In this case, the angles of a triangle add up to *more* than  $180^\circ$  and the circumference of a circle is *less* than  $2\pi r$ . Finally, we see that hyperbolic geometry can be applied to describe the surface of a saddle. For this geometry, the angles of a triangle add up to *less* than  $180^\circ$  and the circumference of a circle *exceeds*  $2\pi r$ . Since all three geometries are equally valid from a mathematical viewpoint, that which describes the spatial structure of the universe must be determined empirically. Cosmologically, the three geometries are often referred to as **flat** (Euclidean  $\kappa = 0$  and  $\Omega_0 = 1$ ), **closed** (Elliptic with  $\kappa = 1$  and  $\Omega > 1$ ), and **open** (Hyperbolic with  $\kappa = -1$  and  $\Omega_0 < 1$ ) (Carroll & Ostlie 2007, pg. 1185).

### FRIEDMANN EQUATION

The metric describing the notion of distance within these three geometries is known as the **Robertson-Walker (RW) metric** which is valid for a universe obeying the cosmological principle (i.e. in a homogenous and isotropic universe, although the curvature of space may change with time, the curvature must have the same value everywhere at a given time since the Big Bang). The RW metric is usually expressed as

$$ds^2 = -c^2 dt^2 + a(t)^2 [dr^2 + S_\kappa(r)^2 d\Omega^2], \quad (16)$$

where  $d\Omega^2 \equiv d\theta^2 + \sin^2\theta d\phi^2$ ,  $a(t)$  is the **scale factor** describing how distances grow or decrease in time (normalized so that  $a(t_0) = 1$ ), and  $S_\kappa(r)$  depends on the geometry in question. In particular, we have that

$$S_\kappa(r) = \begin{cases} R \sin(r/R) & \kappa = +1 \\ r & \kappa = 0 \\ R \sinh(r/R) & \kappa = -1 \end{cases} \quad (17)$$

Here  $\kappa$  is the **curvature constant** and if the space is curved ( $\kappa \neq 0$ ) then the quantity  $R$ , which has dimensions of length, is the radius of curvature. Note that in the limit  $r \ll R$ ,  $S_\kappa \approx r$ , regardless of the value of  $\kappa$ . When space is either flat or negatively curved (open),  $S_\kappa$  increases monotonically with  $r$  and  $S_\kappa \rightarrow \infty$  as  $r \rightarrow \infty$ . By contrast, when space is positively curved (closed),  $S_\kappa$  increases to a maximum of  $S_{\max} = R$  at  $r/R = \pi/2$  then decreases again to 0 at  $r/R = \pi$  (Ryden 2002, pg. 46).

The RW metric simply describes the geometric distance between points in spacetime and is derived independently of general relativity. When general relativity is invoked, the Christoffel symbols for equation (16) can be computed, and Einstein's field equations can be used to derive the **Friedmann Equation**. This equation describes the dynamic evolution of the universe in the form of a differential equation for the scale factor and is often written as

$$H^2(t) = \frac{8\pi G}{3} \rho(t) - \frac{\kappa}{a^2} = H_0^2 \left[ \frac{\rho(t)}{\rho_{\text{crit},0}} - \frac{\kappa}{a^2 H_0^2} \right], \quad (18)$$

where  $\rho_{\text{crit},0} \equiv 3H_0^2/8\pi G$ ,  $H(t) \equiv \dot{a}/a$  is the **Hubble parameter**, and  $H_0 = H(t_0)$  is the **Hubble constant** (Ryden 2002, pg. 63). The form of equation (18) suggests that the evolving expansion or contraction of the universe will depend on how  $\rho(t)$  changes in time. For an isotropic universe described by a fluid of density  $\rho$  and pressure  $P$ , the energy-momentum tensor will be  $T_\nu^\mu = \text{diag}(-\rho, P, P, P)$ . If gravity and velocities were negligible in this universe the pressure and energy density will evolve according to the **continuity equation**,  $\partial\rho/\partial t = 0$ , and the **Euler equation**,  $\partial P/\partial x^\mu = 0$ . This can be promoted to a 4-component conservation equation for the energy-momentum tensor:  $\partial T_\nu^\mu/\partial x^\mu = 0$ . In an expanding universe, however, the conservation criterion must be modified and instead asserts that the covariant derivative of  $T_\nu^\mu$  must vanish. This leads to the following conservation law:

$$\frac{\partial\rho}{\partial t} + \frac{\dot{a}}{a} [3\rho + 3P] = 0 \Leftrightarrow \frac{1}{a^3} \frac{\partial[\rho a^3]}{\partial t} = -3 \frac{\dot{a}}{a} P. \quad (19)$$

This conservation law can be applied immediately to glean information about the scaling of both matter and radiation with expansion. Matter has effectively zero pressure so that

$$\frac{\partial[\rho_m a^3]}{\partial t} = 0 \rightarrow \rho_m \propto a^{-3}. \quad (20)$$

Intuitively this makes sense since the energy of one nonrelativistic matter particle is simply its rest mass, which remains constant in time. The energy density of a collection of particles is then the rest mass energy times their number density, where the latter is inversely proportional to volume and thus scales like  $a^{-3}$ . For radiation,  $P = \rho/3$ , which leads to

$$\frac{\partial\rho_r}{\partial t} + \frac{\dot{a}}{a} 4\rho_r = a^{-4} \frac{\partial[\rho_r a^4]}{\partial t} = 0 \rightarrow \rho_r \propto a^{-4}. \quad (21)$$

Again, we can interpret this easily by noting that the wavelength of radiation will scale with  $a(t)$ . This means that some photon with  $\lambda_0$  today would have had a smaller wavelength of  $\lambda_0/a$  at earlier times, and therefore an energy larger by a factor of  $1/a$ . The total energy density of radiation, which scales as the energy density of each photon times their number density, will then increase by a factor of  $a^{-4}$  (Dodelson 2003, pg. 38).

#### MEASURING GEOMETRY

Measuring distances in an expanding universe is tricky business. The two simplest definitions of distance are the **comoving distance**,  $d_c$ , which remains fixed in time and the **physical distance**,  $d_p$ , which grows simply because of expansion. Frequently, neither of these two measures accurately describes the process of interest so we commonly refer to other types of distances (Dodelson 2003, pg. 34).

One way of inferring distances in astronomy is to measure the flux from some object of known luminosity. In this case we can define the **luminosity distance** to be

$$d_L \equiv \left( \frac{L}{4\pi F} \right)^{1/2}. \quad (22)$$

The function  $d_L$  is called a “distance” because its dimensionality is that of a distance and because it is what the proper distance to a standard candle would be if the universe were static and Euclidean. Suppose, though, that we are in a universe described by the RW metric in equation (16). We will place ourselves at the origin where at the present moment,  $t = t_0$ , we see light that was emitted by a standard candle at comoving coordinate location  $(r, \theta, \phi)$ . The photons which were emitted at time  $t_e$  are, at the present moment, spread over a sphere of radius  $d_p(t_0) = r$  and proper surface area  $A_p(t_0)$ . If space is flat, then the proper area of the sphere is given by the Euclidean relation  $A_p(t_0) = 4\pi d_p(t_0)^2 = 4\pi r^2$ . More generally, however,

$$A_p(t_0) = 4\pi S_\kappa(r)^2, \quad (23)$$

where  $S_\kappa(r)$  was defined in equation (17). When space is positively curved (i.e. closed),  $A_p < 4\pi r^2$ , and the photons are spread over a *smaller* area than they would be in flat space; the opposite is true for a negatively curved (i.e. open) universe. This means that objects at some fixed comoving distance will appear brighter (fainter) in a closed (open) universe compared to a flat space, meaning that this can be used to constrain the overall geometry of the universe. In addition to these geometric effects, which would apply even in a static universe, the expansion of the universe causes the observed flux of light from a standard candle at redshift  $z$  to be decreased by a factor of  $a^2$ . First, the expansion of the universe causes the energy of each photon from the standard candle to decrease. Second, thanks to the expansion of the universe, the time between photon detections will be greater. If two photons are emitted in the same direction separated by a time interval  $\delta t_e$ , the proper distance between them will initially be  $c(\delta t_e)$ ; by the time we detect the photons at time  $t_0$ , the proper distance between them will be stretched to  $c(\delta t_e)(1+z)$ , and we

will detect them separated by a time interval  $\delta t_0 = \delta t_e(1+z)$ . The net result is that in an expanding, spatially curved universe, the luminosity distance to be used in equation (22) is  $d_L = S_\kappa(1+z)$  (Ryden 2002, pg. 134).

The available evidence indicates that our universe is nearly flat, with a radius of curvature  $R_0$  which is larger than the current horizon distance. Objects with finite redshift are at proper distances smaller than the horizon distance, and hence smaller than the radius of curvature. Thus, it is safe to make the approximation  $r \ll R_0$ , implying  $S_\kappa(r) \approx r$ . With our assumption that space is very close to being flat, the relation between the luminosity distance and the current proper distance becomes very simple:  $d_L = r(1+z) = d_p(1+z)$ . Hence, even if space is perfectly flat, if you estimate the distance to a standard candle by using a naive inverse square law, you will overestimate the actual proper distance by a factor of  $(1+z)$  (Ryden 2002, pg. 135).

Suppose now that we have a **standard yardstick** instead of a standard candle; a standard yardstick is an object whose proper length  $l$  is known. Imagine that we see a yardstick of proper length  $l$  aligned perpendicular to our line of sight and we measure an angular distance  $\delta\theta$  between its ends and a redshift  $z$  for the light it emits. If  $\delta\theta \ll 1$  and  $l$  is known we can compute the distance to the yardstick as

$$d_A = \frac{l}{\delta\theta}, \quad (24)$$

known as the **angular-diameter distance**. The angular-diameter distance is equal to the proper distance to the yardstick if the universe is static and Euclidean. In general, though, if the universe is expanding or curved, the angular-diameter distance will not be equal to the current proper distance. We will again place ourselves at the origin and suppose that the yardstick is at a comoving coordinate distance  $r$ . At a time  $t_e$ , the yardstick emitted the light which we observe at time  $t_0$ . The comoving coordinates of the two ends of the yardstick, at the time the light was emitted, were  $(r, \theta_1, \phi)$  and  $(r, \theta_2, \phi)$ . As the light from the yardstick moves toward the origin, it travels along geodesics with  $\theta = \text{constant}$  and  $\phi = \text{constant}$ . Thus, the angular size which we measure for the yardstick will be  $\delta\theta = \theta_2 - \theta_1$ . The distance  $ds$  between the two ends of the yardstick, measured at the time  $t_e$  when the light was emitted, can be found from the Robertson-Walker metric as  $ds = a(t_e)S_\kappa(r)\delta\theta$ . For a standard yardstick we set  $ds = l$  and thus find that

$$\delta\theta = \frac{l(1+z)}{S_\kappa(r)}, \quad (25)$$

so that the angular-diameter distance to be used in equation (24) is  $d_A = S_\kappa(r)/(1+z)$ . From equation (25) we see that the angle subtended by a standard yardstick at some known redshift is larger (smaller) in a closed (open) compared to a flat universe (Ryden 2002, pg. 137). Physically this occurs because gravitational lensing by the background magnifies or demagnifies the object we are looking at. A common yardstick to use is the sound horizon, which is the maximum length that the acoustic oscillations could have traversed since the time of the big bang to recombination. Note thought that increasing the distance to the surface of last scattering also decreases the angular extent of acoustic features in the CMB. This distance depends mainly on the expansion rate and hence the matter content of the universe. Because the same quantity enters length scales such as the sound horizon, this dependence nearly cancels out leaving the much stronger dependence on curvature (Hu et al. 1997).

#### MEASURING COSMOLOGICAL PARAMETERS

The parameter  $w_i$  is defined such that  $\rho_i \propto a^{-3(1+w_i)}$  so that  $w_m = 0$ ,  $w_r = 1/3$  and  $w_\Lambda = -1$ . If we use this relation in equation (18) it is easy to show that the Hubble parameter evolves with redshift as

$$H(z) = H_0 \sqrt{(1+z)^4 \Omega_r + (1+z)^3 \Omega_m + (1+z)^{3(1+w)} \Omega_\Lambda + (1+z)^2 \Omega_\kappa}, \quad (26)$$

where  $\Omega_\kappa \equiv -\kappa/a_0^2 H_0^2$ . This equation tells us that we can use the Friedmann equation to determine  $a(t)$  for some model of the universe. This argument also works in reverse; if we can determine  $a(t)$  from observations we could use that knowledge to determine  $\Omega_i$  for each model component (Ryden 2002, pg. 126).

Since determining the exact functional form of  $a(t)$  is difficult, it is useful, instead, to consider keeping the first three terms in a Taylor expansion for  $a(t)$  around the present epoch:

$$a(t) \approx a(t_0) + \frac{da}{dt} \Big|_{t=t_0} (t-t_0) + \frac{1}{2} \frac{d^2a}{dt^2} \Big|_{t=t_0} (t-t_0)^2, \quad (27)$$

which, after dividing through by  $a(t_0)$  and noting that  $a(t_0) = 1$ , can be rewritten as

$$a(t) \approx 1 + H_0(t-t_0) - \frac{1}{2} q_0 H_0^2 (t-t_0)^2. \quad (28)$$

Here  $H_0$  is the familiar Hubble constant and  $q_0$  is a dimensionless number known as the **deceleration parameter**, defined as

$$q_0 \equiv - \left( \frac{\dot{a}}{a^2} \right)_{t=t_0}. \quad (29)$$

A positive value of  $q_0$  corresponds to  $\ddot{a} < 0$  meaning that the universe's expansion is decelerating, whereas a negative value of  $q_0$  corresponds to  $\ddot{a} < 0$  meaning that the relative velocity between any two points is increasing with time. The results of this Taylor expansion are simply a mathematical description of how the universe expands at times  $t \sim t_0$  and does not invoke any of the physics contained in the parameters  $H_0$  and  $q_0$ . Of course, as we have seen above, we can theoretically predict how

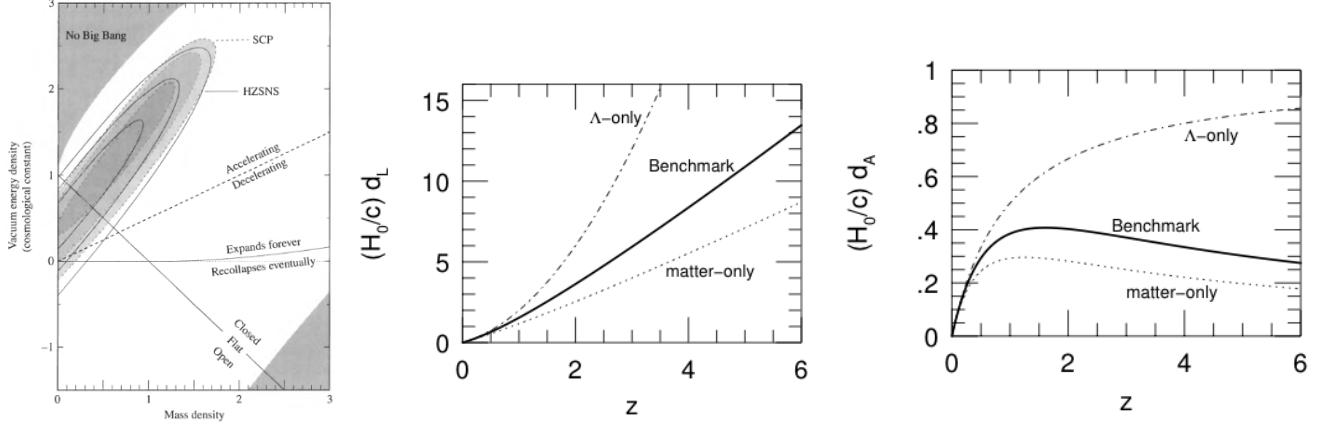


FIG. 3.— (left) The location of the most probable values of  $\Omega_m$  and  $\Omega_\Lambda$  for high- $z$  SNe. The results of two separate results (SCP) and (HZSNS) are superimposed. Image taken from Carroll & Ostlie (2007). (centre) The luminosity distance of a standard candle with observed redshift  $z$ . The bold solid line gives the result for the Benchmark Model, the dot-dash line for a flat, lambda-only universe, and the dotted line for a flat, matter-only universe. (right) The angular-diameter distance for a standard yardstick with observed redshift  $z$ . Images taken from Ryden (2002).

these parameters depend on the specific cosmology. In particular, for a universe containing radiation, matter, and a cosmological constant, we have that

$$q_0 = \Omega_r + \frac{1}{2}\Omega_m - \Omega_\Lambda. \quad (30)$$

Such a universe will be currently accelerating outward if  $\Omega_\Lambda > \Omega_r + \Omega_m/2$ ; our universe is something like  $q_0 = -0.6$  (Ryden 2002, pg. 129).

The proper distance  $d_p(t)$  between two points is defined as the length of the spatial geodesic between the points when the scale factor is fixed at the value  $a(t)$ . If we observe, at time  $t_0$ , light that was emitted by a distant galaxy at time  $t_e$ , the current proper distance to the galaxy is

$$d_p(t_0) = c \int_{t_e}^{t_0} \frac{dt}{a(t)}. \quad (31)$$

Using the approximation in equation (28) this can be rewritten as

$$d_p(t_0) = \frac{c}{H_0} z \left[ 1 - \frac{1+q_0}{2} z \right]. \quad (32)$$

The linear Hubble relation  $d_p \propto z$  thus only holds true in the limit that  $z \ll 2/(1+q_0)$  (Ryden 2002, pg. 130).

For a flat universe we know that  $d_L = d_p(1+z)$ , which from equation (32), implies that at low redshift

$$d_L \approx \frac{c}{H_0} z \left( 1 + \frac{1-q_0}{2} z \right), \quad (33)$$

so that as  $z \rightarrow 0$ ,  $d_L \rightarrow cz/H_0$ . If we construct a redshift-magnitude diagram comparing distance modulus  $m-M$  versus luminosity distance  $d_L$  we can determine  $H_0$  and  $q_0$  and subsequently  $\Omega_m$  and  $\Omega_\Lambda$  (radiation contributes negligibly at the current epoch). In particular, at large enough redshifts that the linear approximation that  $d_L \propto z$  breaks down, the divergence from a straight line allows for the determination of  $q_0$ . In the mid-1990's measurements made of Type 1a SNe at cosmological distances showed that they were *dimmer* than expected for a universe with  $\Omega_\Lambda = 0$ . The solution is that  $\Omega_\Lambda > 0$  implying universal expansion so that the SNe are actually further away than originally believed (Carroll & Ostlie 2007, pg. 1214).

Figure 3 shows the location on the  $\Omega_m - \Omega_\Lambda$  plane the most likely set of values that are consistent with high- $z$  SNe results. In this plot the  $\Omega_m - \Omega_\Lambda$  line separates open and closed universes while from equation (30) the line  $\Omega_m - 2\Omega_\Lambda$  divides accelerating and decelerating universes. Also shown in Figure 3 are the evolution in  $d_L$  and  $d_A$  with redshift for different spatially flat geometries. The luminosity distance is larger for a  $\Lambda$ -dominated universe since accelerated expansion will drive standard candles farther away. The situation is more complicated for the angular-diameter distance. Note that in all cases, as  $z \rightarrow 0$ ,  $d_L \approx d_A \approx d_p \approx (c/H_0)z$ . However, the state of affairs is very different in the limit  $z \rightarrow \infty$ . In models with a finite horizon size,  $d_p \rightarrow d_{\text{hor}}$  in this limit whereas  $d_L \rightarrow \infty$  and  $d_A \rightarrow 0$ . Hence, in model universes other than the lambda-only model, the angular-diameter distance  $d_A$  has a maximum for standard yardsticks at some critical redshift  $z_c$ . This means that if the universe were full of glow-in-the-dark yardsticks, all of the same size  $l$ , their angular size  $\delta\theta$  would decrease with redshift out to  $z = z_c$ , but then would increase at larger redshifts. The sky would be full of big, faint, redshifted yardsticks (Ryden 2002, pg. 140).

**QUESTION 3**

**Outline the development of the Cold Dark Matter spectrum of density fluctuations from the early universe to the current epoch.**

### QUESTION 3

**Outline the development of the Cold Dark Matter spectrum of density fluctuations from the early universe to the current epoch.**

The standard model of cosmology is based on the **cosmological principle**, the assumption of a spatially homogenous and isotropic Universe. Of course, the assumption of homogeneity is justified only on large scales because observations show us that our Universe is inhomogeneous on small scales – otherwise no galaxies or stars would exist. The largest gravitationally bound structures we observe in the universe (i.e. the Great Wall of galaxies) have linear dimensions on the order of 100 Mpc/h in addition to large-scale voids on the order of 50 Mpc/h, which are nearly spherical regions containing virtually no (bright) galaxies. Hence, the Universe seems to be basically homogeneous if averaged over scales of  $R \sim 200$  Mpc/h. This “homogeneity scale” needs to be compared to the **Hubble radius**,  $R_H \equiv c/H_0 \approx 3000$  Mpc/h. This yields  $(R/R_H)^3 \sim 10^{-4}$  so that we are justified in assuming a homogenous cosmology when considering the mean history of the universe (Schneider 2002, pg. 277).

On small scales the universe is inhomogeneous as evidenced by galaxy surveys and anisotropies in the CMB. The smallness of the latter ( $\Delta T/T \sim 10^{-5}$ ) suggests that the density inhomogeneities at the recombination redshift of  $z \approx 1000$  had very small amplitudes. Today, the amplitude of density inhomogeneities are considerably larger. For the purpose of studying the evolution of density fluctuations, it is useful to define the **relative density contrast** as

$$\delta(\mathbf{r}, t) \equiv \frac{\rho(\mathbf{r}, t)}{\bar{\rho}(t)} - 1, \quad (34)$$

where  $\bar{\rho}(t)$  denotes the mean cosmic matter density in the universe at time  $t$ . From the definition of  $\delta$ , we see that  $\delta \geq 1$  because  $\rho \geq 0$ . The dynamics of the cosmic Hubble expansion is controlled by the gravitational field of the average matter density  $\bar{\rho}(t)$  whereas the density fluctuations  $\Delta\rho(\mathbf{r}, t) = \rho(\mathbf{r}, t) - \bar{\rho}(t)$  generate an additional gravitational field (Schneider 2002, pg. 278).

We will be considering regions for which  $\Delta\rho > 0$ , and hence  $\delta > 0$ , so that the gravitational field in this region is stronger than the cosmic average. An overdense region produces a stronger gravitational field than that corresponding to the mean Hubble expansion. By this additional self-gravity, the overdense region will expand more slowly than the average Hubble expansion. Because of the delayed expansion, the density in this region will also decrease more slowly than in the cosmic mean,  $\bar{\rho}(t) = a^{-3}\rho_0$ , and hence the density contrast in this region will increase. As a consequence, the relative density will increase, which again produces an even stronger gravitational field, and so on. It is obvious that this situation is unstable. Of course, the argument also works the other way round: in an under-dense region with  $\delta < 0$ , the gravitational field generated is weaker than in the cosmic mean, therefore the self-gravity is weaker than that which corresponds to the Hubble expansion. This implies that the expansion is decelerated less than in the cosmic mean, the underdense region expands faster than the Hubble expansion, and thus the local density will decrease more quickly than the mean density of the universe. In this way, the density contrast decreases so that  $\delta$  becomes more negative over the course of time. This process describes how **gravitational instability** leads to an increase of density fluctuations over the course of time (Schneider 2002, pg. 278).

#### LINEAR PERTURBATION THEORY

We first will examine the growth of density perturbations with a concentration on length scales that are substantially smaller than the Hubble radius. On these scales, structure growth can be described in the framework of the Newtonian theory of gravity. The effects of spacetime curvature and thus of General Relativity need to be accounted for only for density perturbations on length-scales comparable to, or larger than the Hubble radius. Since the Poisson equation, which specifies the relation between matter density and the gravitational potential, is linear, the effects of the homogeneous matter distribution and of density fluctuations can be considered separately. The gravitational field of the total matter distribution is then the sum of the average matter distribution and that of the density fluctuations. In addition, we will assume that the matter in the Universe consists only of dust (i.e., pressure-free matter) with density and velocity fields  $\rho(\mathbf{r}, t)$  and  $\mathbf{v}(\mathbf{r}, t)$  respectively (Schneider 2002, pg. 279).

The behaviour of this fluid is described by the **continuity equation**,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, \quad (35)$$

which expresses the fact that matter is conserved: the density decreases if the fluid has a diverging velocity field (thus, if particles are moving away from each other). Furthermore, the **Euler equation** applies,

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} = -\frac{\Delta P}{\rho} - \nabla \Phi, \quad (36)$$

which describes the conservation of momentum and the behaviour of the fluid under the influence of forces. The left-hand side of equation (36) is the time derivative of the velocity as would be measured by an observer moving with the flow, because  $\partial \mathbf{v} / \partial t$  is the derivative at a fixed point in space, whereas the total left-hand side of equation (36) is the time derivative of the velocity measured along the flow lines. The latter is affected by the pressure gradient and the gravitational field  $\Phi$ , the latter satisfying the Poisson equation

$$\nabla^2 \Phi = 4\pi G \rho. \quad (37)$$

Since we are considering dust, the pressure vanishes so that  $P = 0$ . These three equations for the description of a self-gravitating fluid can in general not be solved analytically. However, we will show that a special, cosmologically relevant exact solution can

be found, and that by linearization of the system of equations approximate solutions can be constructed for  $|\delta| \ll 1$  (Schneider 2002, pg. 279).

The special case is to consider the **Hubble flow** where  $\mathbf{v}(\mathbf{r}, t) = H(t)\mathbf{r}$  is the solution to the above set of equations for a homogeneous distribution of matter in an RW metric. As long as the density contrast  $|\delta \ll 1|$ , the deviations of the velocity field from the Hubble expansion will be small. We expect that in this case, physically relevant solution of the above equations are those which deviate only slightly from the homogeneous case. It is convenient to consider the problem in comoving coordinates where  $\mathbf{r} = a(t)\mathbf{x}$ . Likewise, the velocity field is written in the form

$$\mathbf{v}(\mathbf{r}, t) = \frac{\dot{a}}{a}\mathbf{r} + \mathbf{u}(\mathbf{x}, t), \quad (38)$$

where the first term represents the homogeneous Hubble expansion and the second represents the peculiar velocity  $\mathbf{u}$ . With this convention, equations (35) through (37) can be rewritten in terms of spatial derivatives with respect to  $\mathbf{x}$  and all variables like  $\rho$  and  $\delta$  are in reference to physical values evaluated at comoving coordinates (Schneider 2002, pg. 280).

The next step is to then linearize the equations so that we can look for approximate solutions which describe only small deviations from this homogeneous solution. This process involves considering only first-order terms in the small parameters  $\delta$  and  $\mathbf{u}$ . After this linearization, it is possible to eliminate  $\Phi$  and  $\mathbf{u}$  from the equations and obtain a second-order differential equation for the density contrast:

$$\frac{\partial^2 \delta}{\partial t^2} + \frac{2\dot{a}}{a} \frac{\partial \delta}{\partial t} = 4\pi G \bar{\rho}(t) \delta. \quad (39)$$

This equation contains neither partial derivatives with respect to spatial coordinates, nor coefficients that depend on  $\mathbf{x}$ . Therefore, equation (39) has solutions of the form

$$\delta(\mathbf{x}, t) = D(t) \tilde{\delta}(\mathbf{x}), \quad (40)$$

that is, the spatial and temporal dependences factorize in these solutions. Here  $\tilde{\delta}(\mathbf{x})$  is an arbitrary function of the spatial coordinate, and  $D(t)$  must satisfy

$$\ddot{D} + \frac{2\dot{a}}{a} \dot{D} - 4\pi G \bar{\rho}(t) D = 0. \quad (41)$$

This differential equation has two linearly independent solutions, with one increasing with time and the other decreasing. If, at some early time, both functional dependencies were present, the increasing solution will dominate at later times, whereas the solution decreasing with  $t$  will become irrelevant. Therefore, we consider only the increasing solution,  $D_+(t)$ , normalized so that it is unity at the current epoch. Then the density contrast becomes

$$\delta(\mathbf{x}, t) = D_+(t) \delta_0(\mathbf{x}). \quad (42)$$

This mathematical consideration allows us to draw immediately a number of conclusions. First, equation (42) indicates that in linear perturbation theory the spatial shape of the density fluctuations is frozen in comoving coordinates, only their amplitude increases. The growth factor  $D_+(t)$  of the amplitude follows a simple differential equation that is easily solvable for any cosmological model. In fact, one can show that for arbitrary values of the density parameter in matter and vacuum energy, the growth factor has the form

$$D_+(t) \propto \frac{H(a)}{H_0} \int_0^a \frac{da'}{\left[\Omega_m/a' + \Omega_\Lambda/a'^2 - (\Omega_m + \Omega_\Lambda - 1)\right]^{3/2}}. \quad (43)$$

where the factor of proportionality is determined from the condition  $D_+(t_0) = 1$  (Schneider 2002, pg. 281).

In accordance with  $D_+(t_0) = 1$ ,  $\delta_0(x)$  would be the distribution of density fluctuations today if the evolution was indeed linear until the present epoch. Therefore,  $\delta_0(x)$  is denoted as the linearly extrapolated density fluctuation field. However, the linear approximation breaks down if  $|\delta|$  is no longer  $\ll 1$ . In this case, the terms that have been neglected in the above derivations are no longer small and have to be included. The problem then becomes considerably more difficult and defies analytical treatment. Instead one needs, in general, to rely on numerical procedures for analyzing the growth of density perturbations. Furthermore, for large density perturbations, objects like galaxies begin to form and the fluid approximation breaks down. Also, it should be noted that the above results are only valid for  $z \leq z_{eq}$  so that radiation can be safely neglected (Schneider 2002, pg. 281).

Note that for the special case of an **Einstein-de Sitter model** where  $\Omega_m = 1$  and  $\Omega_\Lambda = 0$ , equation (41) can be solved explicitly. In this case, from equation (26)  $\dot{a} \propto a^{-1/2}$  so that  $a(t) = (t/t_0)^{2/3}$ , and we solve the equation for  $D$  by assuming  $D \propto t^q$ . This positive growth factor becomes  $D_+(t) = a(t)$  so that fluctuations grow with the scale factor. For different cosmological parameters this is not the case, but the qualitative behaviour is quite similar. In particular, fluctuations were able to grow by a factor  $\sim 1000$  from the epoch of recombination to today (Schneider 2002, pg. 281).

At the present epoch,  $\delta \gg 1$  certainly on scales of clusters of galaxies ( $\sim 2$  Mpc) and  $\delta 1$  on scales of superclusters ( $\sim 10$  Mpc). Hence, from the result above that  $D_+ \sim a$ , we would expect  $\delta \gtrsim 10^{-3}$  at  $z = 1000$  for these structures to be able to grow to non-linear structures at the current epoch. For this reason, we should also expect CMB fluctuations to be of comparable magnitude,  $\Delta T/T \sim 10^{-3}$ ; however we find  $\Delta T/T \sim 10^{-5}$ . The corresponding density fluctuations therefore cannot have grown sufficiently strongly up to today to form non-linear structures. The contradiction can be resolved by the dominance of dark matter. Since photons interact with baryonic matter only, the CMB anisotropies basically provide (at least on angular scales below  $\sim 1^\circ$ ) information on the density contrast of baryons. Dark matter may have had a higher density contrast at recombination and may have formed potential wells, into which the baryons then “fall” after recombination (Schneider 2002, pg. 282).

### DESCRIPTION OF DENSITY FLUCTUATIONS

Two universes are considered equivalent if their density fields  $\delta$  have the same statistical properties. One may then imagine considering a large (statistical) ensemble of universes whose density fields all have the same statistical properties, but for which the individual functions  $\delta(\mathbf{x})$  are all different. This statistical ensemble is called a **random field**, and any individual distribution with the respective statistical properties is called a **realization of the random field** (Schneider 2002, pg. 282).

One way to describe the statistical clustering of matter in the universe is to compute the **correlation function**  $\xi(r)$  which is typically used in reference to galaxy surveys. In particular, this function provides the normalized probability of finding two galaxies residing a distance  $r$  from each other. This function is simply determined by averaging over the density products for a large number of pairs of points (i.e galaxies) with given separation  $r$ .  $\xi(r)$  is more accurately labelled as the two-point correlation since it describes the statistical separation of pairs of points. Correlations of higher order may also be defined, leading to general  $n$ -point correlation functions. These are more difficult to determine from observation, though. It can be shown that the statistical properties of a random field are fully specified by the set of all  $n$ -point correlations (Schneider 2002, pg. 284).

An alternative (and equivalent) description of the statistical properties of a random field, and thus of the structure of the Universe, is the **power spectrum**  $P(k)$ . Roughly speaking, the power spectrum  $P(k)$  describes the level of structure as a function of the length-scale  $L = 2\pi/k$ ; the larger  $P(k)$ , the larger the amplitude of the fluctuations on a length-scale  $2\pi/k$ . Here,  $k$  is a **wave number**. The power spectrum and the correlation function  $\xi(r)$  are actually related through a Fourier transform. In general, knowing the power spectrum is not sufficient to unambiguously describe the statistical properties of any random field – in the same way as the correlation function only provides an incomplete characterization. However, random fields do exist, so-called **Gaussian random fields**, which are uniquely characterized by  $P(k)$ . Such Gaussian random fields play an important role in cosmology because it is assumed that at very early epochs, the density field obeyed Gaussian statistics (Schneider 2002, pg. 285).

### TRANSFER FUNCTION

Both  $P(k)$  and  $\xi(r)$  depend on redshift since the density field of the universe evolves over time. Note that  $P(k,t)$  and  $\xi(r,t)$  are linearly related through a Fourier transform and  $\xi(r)$  quadratically depends on  $\delta$ . Thus, within the scope of the validity of equation (42)

$$\xi(x,t) = D_+^2(t)\xi(x,t_0) \quad \text{and} \quad P(k,t) = D_+^2(t)P(k,t_0) \equiv D_+^2(t)P_0(k), \quad (44)$$

where  $k$  is a comoving wave number. We shall stress once again that these relations are valid only in the framework of Newtonian, linear perturbation theory in the matter dominated era of the universe. This result tells us that knowledge of  $P_0(k)$  is sufficient to obtain the power spectrum  $P(k,t)$  at any time, again within the framework of linear perturbation theory (Schneider 2002, pg. 285).

It now becomes useful to try and understand the initial power spectrum  $P_0(k)$ . At early times, the expansion of the universe follows a power-law,  $a(t) \propto t^{1/2}$ , in the radiation-dominated era. At that time, no natural length-scale existed in the universe to which one might compare a wavelength. The only mathematical function that depends on a length but does not contain any characteristic scale is a power law; hence for very early times one should expect

$$P(k) \propto k^{n_s}. \quad (45)$$

Based on scaling relations it was argued that  $n_s$  should equal unity; such a power spectrum is referred to as the *Harrison-Zel'dovich spectrum* (Schneider 2002, pg. 285).

It now becomes important to step away from the simplifications we have been making. In particular, the evolution of perturbations in the radiation-dominated era proceeds differently from the matter-dominated era we have been considering. For this reason, we introduce a correction term

$$P_0(k) = Ak^{n_s}T^2(k), \quad (46)$$

where  $T(k)$  is called the **transfer function**; it can be computed for any cosmological model if the matter content of the universe is specified. Of course, equation (46) is only valid in the regime of linear perturbation theory and is often used, for example, in generating the initial conditions of cosmological simulations (Schneider 2002, pg. 286).

In linear perturbation theory, fluctuations grow on all scales, or for all wave numbers, independent of each other. This is valid for both the Newtonian and GR cases as long as the fluctuation amplitudes are small. Therefore, the behaviour on any (comoving) length-scale can be investigated independently of the other scales. At very early times, perturbations with a comoving scale  $L$  are larger than the (comoving) horizon, and only for  $z < z_{\text{enter}}(L)$  does the horizon become larger than the considered scale  $L$ . Here,  $z_{\text{enter}}(L)$  is defined as the redshift at which the (comoving) horizon equals the (comoving) length-scale  $L$ ,

$$\eta(z_{\text{enter}}(L)) = L \quad \text{where} \quad \eta(z) \equiv c \int_0^t \frac{dt}{a} = c \int_0^a \frac{da}{a^2 H(a)}. \quad (47)$$

It is common to say that at  $z_{\text{enter}}(L)$  the perturbation under consideration “enters the horizon”, whereas actually the process is the opposite – the horizon outgrows the perturbation. Relativistic perturbation theory shows that density fluctuations of scale  $L$  grow as long as  $L > \eta$ , namely  $\propto a^2$  if radiation dominates (thus, if  $z > z_{\text{eq}}$ ), or  $\propto a$  if matter dominates (thus, if  $z < z_{\text{eq}}$ ). Free-streaming particles or pressure gradients cannot impede the growth on scales larger than the horizon length because, according to the definition of the horizon, physical interactions cannot extend to scales larger than the horizon size (Schneider 2002, pg. 287).

The behaviour of the growth of a density perturbation on a scale  $L$  for  $z < z_{\text{enter}}(L)$  depends on  $z_{\text{enter}}$  itself. If a perturbation enters the horizon in the radiation-dominated phase,  $z_{\text{eq}} \leq z_{\text{enter}}(L)$ , the fluctuation cannot grow during the epoch  $z_{\text{eq}} \leq z \leq z_{\text{enter}}(L)$ . In this period, the energy density in the Universe is dominated by radiation, and the resulting expansion rate prevents an efficient

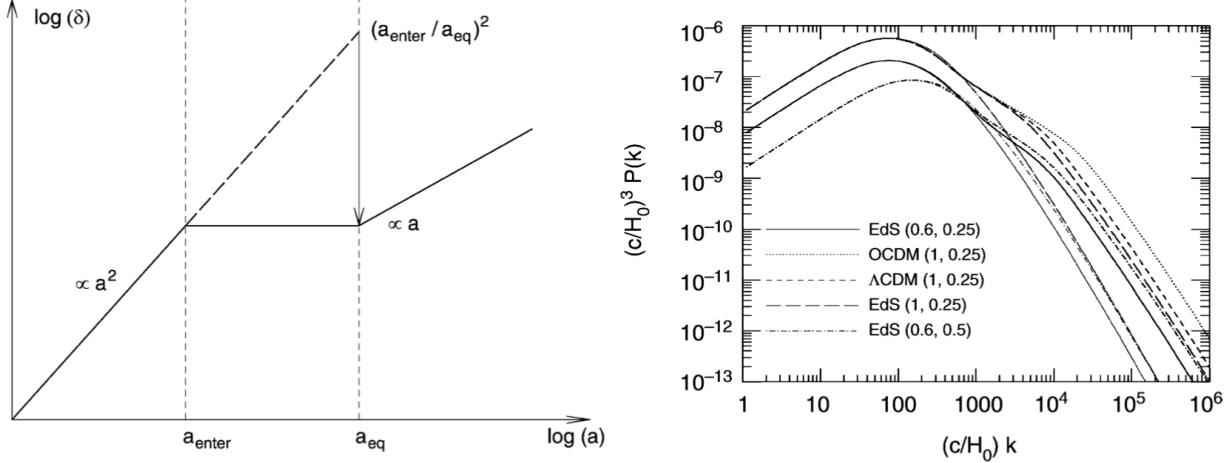


FIG. 4.— (left) A density perturbation that enters the horizon during the radiation-dominated epoch of the Universe ceases to grow until matter starts to dominate the energy content of the Universe. In comparison to a perturbation that enters the horizon later, during the matter-dominated epoch, the amplitude of the smaller perturbation is suppressed by a factor  $(a_{\text{eq}}/a_{\text{enter}})^2$ , which explains the qualitative behaviour of equation (49). (right) The current power spectrum of density fluctuations for CDM models. The various curves have different cosmological parameters: EdS:  $\Omega_m = 1, \Omega_\Lambda = 0$ ; OCDM:  $\Omega_m = 0.3, \Omega_\Lambda = 0$ ;  $\Lambda$ CDM:  $\Omega_m = 0.3, \Omega_\Lambda = 0.7$ . The values in parentheses specify  $(\sigma_8, \Gamma)$  where  $\sigma_8$  is the normalization of the power spectrum and  $\Gamma = \Omega_m h$  is the shape parameter. Images taken from Schneider (2002).

perturbation growth. At later epochs, when  $z \leq z_{\text{eq}}$ , the growth of density perturbation continues. If  $z_{\text{enter}}(L) \leq z_{\text{eq}}$ , thus if the perturbation enters the horizon during the matter-dominated epoch of the Universe, these perturbations will grow as described above with  $\delta \propto D_+(t)$ . This implies that a length-scale  $L_0$  is singled out, namely the one for which

$$z_{\text{eq}} = z_{\text{enter}}(L_0). \quad (48)$$

Density fluctuations with  $L > L_0$  enter the horizon after matter started to dominate the energy density of the Universe so their growth is not impeded by the radiation-dominated era. In contrast, density fluctuations with  $L < L_0$  enter the horizon at a time when radiation dominates and have to wait until matter-domination to grow in amplitude. The quantitative consideration of these effects allows us to compute the transfer function. In general, this needs to be done numerically, but very good approximations exist. Two limiting cases are easily treated analytically,

$$T(k) \approx \begin{cases} 1 & k \ll 1/L_0 \\ (kL_0)^{-2} & k \gg 1/L_0 \end{cases} \quad (49)$$

Figure 4 displays the qualitative description of the transfer function that we have just argued (Schneider 2002, pg. 287).

Figure 4 also shows  $P(k)$  for different cosmological models where thin lines show the power spectrum as derived from linear perturbation theory and bold lines include non-linear structure formation. The power spectra displayed all have a characteristic wave number at which the slope of  $P(k)$  changes; it is specified by  $\sim 2\pi/L_0$ . Since  $L_0$  depends on the horizon scale at the epoch of equality, it implicitly depends on  $\Omega_m$  (and to an even smaller degree  $\Omega_b$ ) so its location can be used as a cosmological probe (Schneider 2002, pg. 288).

#### NON-LINEAR EVOLUTION

When linear perturbation theory fails, the fluid mechanics equations can be expanded to higher order in  $\delta$  and  $\mathbf{u}$  to construct non-linear perturbation theories. However, these theories are often complicated and thus rely on the use of numerical simulations to study the non-linear dynamics of structure formation. However, there are a number of interesting cases that can be considered analytically; below we describe spherical collapse.

We consider a spherical region with a density enhancement of  $\rho(t) = [1 + \delta(t)]\bar{\rho}(t)$  that is spread homogeneously throughout its interior. At first the density perturbation will grow linearly with  $D_+(t)$  as long as  $\delta \ll 1$ ; after some time it will grow non-linearly. Let  $R_c$  be the initial comoving radius of the overdense sphere; as long as  $\delta \ll 1$  the comoving radius will change only marginally. Due to the enhanced gravitational force, the sphere will expand slightly more slowly than the Universe as a whole, which again will lead to an increase in its density contrast. This then decelerates the expansion rate even further, relative to the cosmic expansion rate. Indeed, the equations of motion for the radius of the sphere are identical to the Friedmann equations for the cosmic expansion, only with the sphere having an effective  $\Omega_m$  different from that of the mean Universe. If the initial density is sufficiently large, the expansion of the sphere will come to a halt, and the sphere will recollapse. Of course, the sphere will not collapse to a point, but will virialize due to small-scale inhomogeneities and gravitational fluctuations. This process occurs roughly on the dynamical time scale(i.e. the time it takes particles to fully cross the sphere). When the sphere is in virial equilibrium its average density will be

$$\langle \rho \rangle = (1 + \delta_{\text{vir}})\bar{\rho} \quad \text{where} \quad (1 + \delta_{\text{vir}}) \simeq 178\Omega_m^{-0.6}. \quad (50)$$

This relation forms the basis for the statement that the virialized region, e.g., of a cluster, is a sphere with an average density  $\sim 200$  times the critical density  $\rho_{\text{crit}}$  of the Universe at the epoch of collapse. Another conclusion from this consideration is that

a massive galaxy cluster with a virial radius of 1.5 Mpc must have formed from the collapse of a region that originally had a comoving radius of about six times this size, roughly 10 Mpc. Such a virialized mass concentration of dark matter is called a **dark matter halo** (Schneider 2002, pg. 290).

This description can be used to construct the Press-Schechter halo mass function. If the statistical properties of  $\Delta_0(\mathbf{x})$  are Gaussian – which is expected for a variety of reasons – the statistical properties of the fluctuation field  $\delta_0$  are solely defined by  $P(k)$ . If fluctuations reach some critical value  $\delta \geq \delta_{\min}$ , they will evolve to form DM halos through the process described above. We can then use  $P(k)$  to compute the number density  $n(M, z)$  of virialized DM halos as a function of mass  $M$  (the specific value of  $\delta$  is related to  $R$  and hence  $M$ ) and redshift  $z$  (Schneider 2002, pg. 291).

#### DENSITY DISTRIBUTION OF BARYONS

The evolution of density fluctuations of baryons differs from that of dark matter. The reason for this is essentially the interaction of baryons with photons: although matter dominates the Universe for  $z < z_{\text{eq}}$ , the density of baryons remains smaller than that of radiation for a long time, until after recombination begins. Since photons and baryons interact with each other by photon scattering on free electrons, which again are tightly coupled electromagnetically to protons and helium nuclei, and since radiation cannot fall into the potential wells of dark matter, baryons are hindered from doing so as well. Hence, the baryons are subject to radiation pressure. For this reason, the density distribution of baryons is initially much smoother than that of dark matter. Only after recombination does the interaction of baryons with photons cease to exist, and the baryons can fall into the potential wells of dark matter, i.e., some time later the distribution of baryons will closely resemble that of the dark matter (Schneider 2002, pg. 289).

#### NATURE OF THE DM

One thing that  $T(k)$  depends on is the nature of DM; specifically, whether it qualifies as CDM or HDM. These two kinds of dark matter differ in the thermal velocities of their constituents at the time of matter-radiation equality. The particles of CDM were non-relativistic at this time, whereas those of HDM had velocities of order  $c$ . If dark matter consists of weakly interacting elementary particles, the difference between CDM and HDM depends on the mass  $m$  of the particles. Assuming that the temperature of the DM particles is close to that of the universe, then a particle of mass  $m$  satisfying the relation,

$$mc^2 \gg k_B T(z_{\text{eq}}) \simeq k_B \times 2.73 \text{ K} (1 + z_{\text{eq}}) \sim 10 \text{ eV}, \quad (51)$$

indicates CDM and the opposite equality indicates HDM (Schneider 2002, pg. 286).

If dark matter consists of relativistic particles, these are not gravitationally bound in the potential well of a density concentration. In this case, they are able to move freely and to escape from the potential well, which in the end leads to its dissolution if these particles dominate the matter overdensity. From this, it follows immediately that for HDM small-scale density perturbations cannot form. For CDM this effect of **free-steaming** does not occur. This implies a clear difference must exist between HDM and CDM models as regards structure formation and evolution. In HDM models, small-scale fluctuations are washed out by free-streaming of relativistic particles, i.e., the power spectrum is completely suppressed for large  $k$ , which is expressed by the transfer function  $T(k)$  decreasing exponentially for large  $k$ . In the context of such a theory, very large structures will form first, and galaxies can form only later by fragmentation of large structures. However, this formation scenario is in clear contradiction with observations. For example, we observe galaxies and QSOs at  $z \sim 6$  so that small-scale structure is already present at times when the universe was less than 10% of its current age. In addition, the observed correlation function of galaxies, both in the local universe and at higher redshift, is incompatible with cosmological models in which the dark matter is composed mainly of HDM (Schneider 2002, pg. 286).

**QUESTION 4**

**State and explain three key pieces of evidence for the Big Bang theory of the origin of the Universe.**

#### QUESTION 4

State and explain three key pieces of evidence for the Big Bang theory of the origin of the Universe.

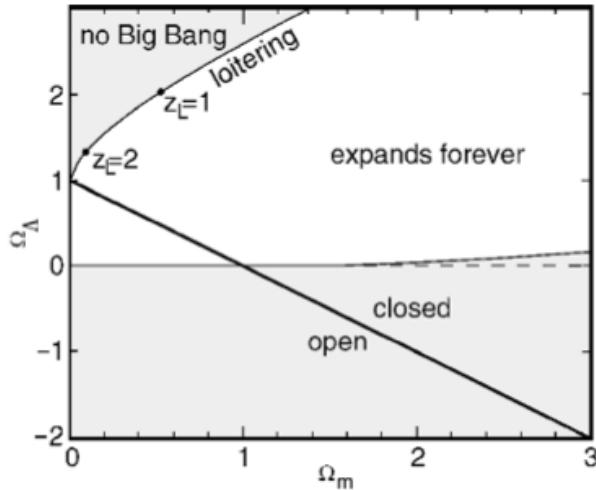


FIG. 5.— Classification of cosmological models. The straight solid line connects flat models and separates open ( $K < 0$ ) and closed ( $K > 0$ ) models. The nearly horizontal curve separates models that will expand forever from those that will recollapse in the distant future. Models in the upper left corner have an expansion history where  $a$  has never been close to zero and thus did not experience a Big Bang. In those models, a maximum redshift for sources exists, which is indicated for two cases. Since we know that  $\Omega_{\text{matter}} > 0.1$ , and sources at  $z > 6$  have been observed, these models can be excluded. Image taken from Schneider (2002).

#### HUBBLE'S LAW

Observations of distant galaxies and quasars show that these objects are redshifted. If the redshift is interpreted as a Doppler shift, the recessional velocity of the object can be calculated. The recessional velocities are a linear relationship with the distances known as Hubble's Law:

$$v = H_0 D, \quad (52)$$

where the Hubble constant  $H_0 \simeq 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$ . Note that a naive interpretation of Hubble's Law seemingly violates the cosmological principle. However, what we view from the MW is exactly what we would expect to see in a universe that is undergoing homogeneous and isotropic expansion. To see this, consider three galaxies that form some triangle. Homogeneous and uniform expansion means that the shape of the triangle is preserved as the galaxies move away from each other. From this, it is easy to see that the viewpoint from each galaxy will be equivalent; each will observe the same linear recession (or contraction) of galaxies with the same scale factor (Ryden 2002, pg. 19).

If galaxies are currently moving away from each other, this implies they were closer together in the past. Consider a pair of galaxies currently separated by a distance  $r$ , with a velocity  $v = H_0 r$  relative to each other. If there are no forces acting to accelerate or decelerate their relative motion, then their velocity is constant, and the time that has elapsed since they were in contact is

$$t_0 = \frac{r}{v} = \frac{r}{H_0 r} = \frac{1}{H_0} \simeq 14 \text{ Gyr}, \quad (53)$$

independent of the current separation  $r$ . This quantity is known as the **Hubble time**. If the relative velocities of galaxies have been constant in the past, then one Hubble time ago, all the galaxies in the universe were crammed together into a small volume. Thus, the observation of galactic redshifts lead naturally to a Big Bang model for the evolution of the universe. A Big Bang model may be broadly defined as a model in which the universe expands from an initially highly dense state to its current low-density state. The Hubble time is comparable to the ages computed for the oldest known stars in the universe. This rough equivalence is reassuring, though it should be noted that the actual age of the universe may not correspond exactly to  $t_0$ . Specifically, the age of the universe will depend on its matter and energy content since these influence the forces acting to speed up or slow down expansion in time. Just as the Hubble time provides a natural time scale for our universe, the **Hubble distance**,  $L_0 = c/H_0 \simeq 4.3 \text{ Gpc}$ , provides a natural distance scale. This will roughly equal the **horizon distance** (the greatest distance a photon can travel during the age of the universe), though its exact value again depends on the matter contents of the universe. Finally, note that this description ties in with **Olbers' Paradox**. If the universe is of finite age  $t_0$ , then the night sky can be dark, even if the universe is infinitely large, because light from distant galaxies has not yet had time to reach us. (Ryden 2002, pg. 20).

We can explore the expansion history of the universe in more detail by evaluating its matter content. In particular, we have previously seen that the Friedmann equation implies that the Hubble parameter evolves as

$$\left(\frac{\dot{a}}{a}\right)^2 = H^2(t) = H_0^2 \left[ \frac{\Omega_{\text{rad}}}{a(t)^4} + \frac{\Omega_{\text{matter}}}{a(t)^3} + \frac{\Omega_\Lambda}{a(t)^{3(1+w)}} + \frac{\Omega_\kappa}{a(t)^2} \right] \quad (54)$$

For small  $a$ , the first term dominates; for slightly larger  $a \gtrsim a_{\text{eq}}$  the dust (or matter) term dominates; for larger  $a$  the curvature term dominates so long as  $\Omega_\kappa \neq 0$ ; for very large  $a$  the cosmological constant term dominates as long as  $\Omega_\Lambda \neq 0$ . In general, the differential equation (54) cannot be solved analytically, though it is within bounds of numerical solutions. Nevertheless, we can analyze the qualitative behaviour of the function  $a(t)$  and thereby understand the essential aspects of the expansion history. From the Hubble law, we conclude that  $\dot{a}(t_0) > 0$ . Equation (54) shows that  $\dot{a}(t) > 0$  for all  $t$ , unless the right-hand side vanishes for some  $a$ ; the sign of  $\dot{a}$  can only switch when the right-hand side is zero. If  $H^2 = 0$  for a value of  $a > 1$ , the expansion will come to a halt and the Universe will recollapse afterwards. On the other hand, if  $H^2 = 0$  for a value  $a = a_{\min}$  with  $0 < a_{\min} < 1$ , then the sign of  $\dot{a}$  switches at  $a_{\min}$ . At this epoch, a collapsing Universe changes into an expanding one (Schneider 2002, pg. 152).

Which of these alternatives describes our Universe depends on the density parameters. In terms of whether or not a Big Bang existed, we have the following classification scheme:

- If  $\Lambda < 0$  then  $H^2 < 0$  for all  $a \leq 1$  so a Big Bang must have existed. The universe will eventually recollapse in on itself.
- If  $\Lambda = 0$ , then  $H^2 > 0$  for all  $a \leq 1$  so a Big Bang must have existed. The expansion behaviour for  $a > 1$  depends on the values of  $\Omega_{\text{matter}}$  and  $\Omega_\kappa$ .
- If  $\Lambda > 0$ , then it is in principle possible that  $H^2 = 0$  for  $a = a_{\min} < 1$ . However, this is only possible if  $\Omega_{\text{matter}}$  is sufficiently small and the value for  $a_{\min}$  will then depend on the values of  $\Omega_{\text{matter}}$  and  $\Omega_\Lambda$ . For instance, with  $\Omega_{\text{matter}} > 0.1$ , we have that  $a_{\min} > 0.3$  so that  $z_{\max} < 2$ . However, since we have observed quasars and galaxies with  $z > 6$  and the density parameter is known to be  $\Omega_{\text{matter}} > 0.1$ , such a model without a Big Bang can be excluded.

These cases are shown graphically in Figure 5. With the exception of the last case, which can be excluded, we come to the conclusion that  $a$  must have attained the value  $a = 0$  at some point in the past. At this instant the “size of the Universe” formally vanished. As  $a \rightarrow 0$ , both matter and radiation densities diverge so that the density in this state must have been singular. The epoch at which  $a = 0$  and the evolution away from this state is called the Big Bang (Schneider 2002, pg. 153).

#### COSMIC MICROWAVE BACKGROUND

Before we first discuss the CMB and its relation to the Big Bang, let’s learn some more about blackbody radiation. One way to make photons is to take a dense, opaque object – such as the filament of an incandescent lightbulb – and heat it up. If the object is opaque, then the protons, neutrons, electrons, and photons which it contains frequently interact, and attain thermal equilibrium. When a system is in thermal equilibrium, the density of photons in the system, as a function of photon energy, depends only on the temperature  $T$ . It doesn’t matter whether the system is a tungsten filament, or an ingot of steel, or a sphere of ionized hydrogen and helium. The energy density of photons in the frequency range  $\nu$  to  $\nu + d\nu$  is given by the **blackbody function**:

$$\varepsilon(\nu)d\nu = \frac{8\pi h}{c^3} \frac{\nu^3 d\nu}{\exp(h\nu/kT) - 1}. \quad (55)$$

When equation (55) is integrated over all  $\nu$ , to total energy density for blackbody radiation is  $\varepsilon = \alpha T^4$ , with  $\alpha$  a constant. In addition, the number density of blackbody photons can be computed from equation (55) as  $n = \beta T^3$ , with  $\beta$  a constant (Ryden 2002, pg. 25).

The discovery of the CMB showed that the universe is permeated by blackbody radiation with temperate, energy density, number density, and wavelength

$$T_0 \simeq 2.73 \text{ K}, \quad \varepsilon \simeq 0.25 \text{ eV cm}^{-3}, \quad n \simeq 400 \text{ cm}^{-3}, \quad \text{and} \quad \lambda \simeq 2 \text{ mm}. \quad (56)$$

The existence of the CMB is a very important cosmological clue. In particular, it is the clue which caused the Big Bang model for the universe to be favoured over the Steady State model. In a Steady State universe, the existence of blackbody radiation at 2.73 K is not easily explained. In a Big Bang universe, however, a cosmic background radiation arises naturally if the universe was initially very hot as well as being very dense. If mass is conserved in an expanding universe, then in the past, the universe was denser than it is now. Assume that the early dense universe was very hot ( $T \gg 10^4$  K or  $kT \gg 1$  eV). At such high temperatures, the baryonic matter in the universe was completely ionized, and the free electrons rendered the universe opaque. As we know above, a dense, hot, opaque body produces blackbody radiation. However, as the universe expanded, it cooled, and when the temperature dropped to  $T \sim 3000$  K, ions and electrons combined to form neutral atoms. When the universe no longer contained a significant number of free electrons, the blackbody photons started streaming freely through the universe, without further scattering off free electrons (Ryden 2002, pg. 28).

The blackbody radiation that fills the universe today can be explained as a relic of the time when the universe was sufficiently hot and dense to be opaque. However, at the time the universe became transparent, its temperature was  $T \sim 3000$  K; a factor of roughly 1000 times larger than today. The drop in temperature of the blackbody radiation is a direct consequence of the expansion of the universe. Consider a region of volume  $V$  which expands at the same rate as the universe, so that  $V \propto a(t)^3$ . The blackbody radiation in the volume can be thought as a photon gas with energy density  $\varepsilon = \alpha T^4$ . Moreover, since the photons in the volume have momentum as well as energy, the photon gas has a pressure; the pressure of a photon gas is  $P = \varepsilon/3$ . The photon gas within our imaginary box must follow the laws of thermodynamics; in particular, the boxful of photons must obey the first law

$$dQ = dE + PdV. \quad (57)$$

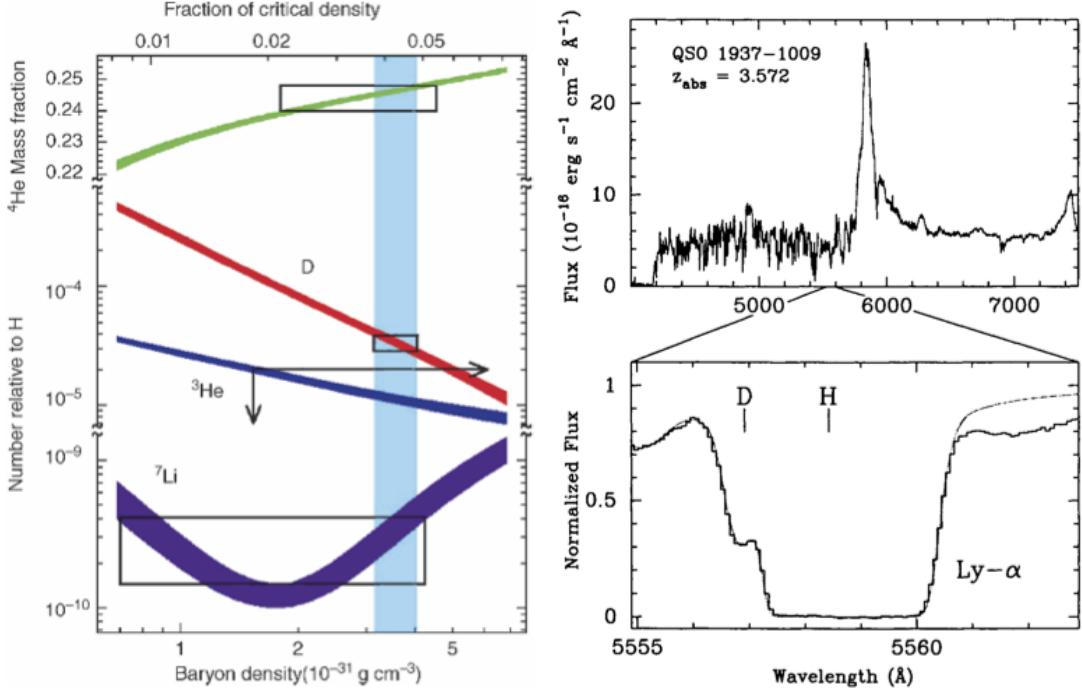


FIG. 6.— (left) BBN predictions of the primordial abundances of light elements as a function of today's baryon density ( $\rho_{b,0}$ , lower axis) and the corresponding density parameter  $\Omega_b$  where  $h = 0.65$  is assumed. The vertical extent of the rectangles marks the measured values of the abundances (top:  $\text{He}^4$ , centre: D, bottom:  $\text{Li}^7$ ). The horizontal extent results from the overlap of these intervals with curves computed from theoretical models. The ranges in  $\Omega_b$  that are allowed by these three species do overlap, as is indicated by the vertical strip. The deuterium measurements yield the most stringent constraints for  $\Omega_b$ . Spectrum from a distant quasar showing absorption by photons with rest wavelengths of  $\lambda = 1216 \text{ \AA}$ . Image taken from Schneider (2002). (right) Image taken from Dodelson (2003).

Since, in a homogeneous universe, there is no net flow of heat (everything is the same temperature, after all),  $dQ = 0$ . Thus, the first law of thermodynamics, applied to an expanding homogeneous universe, is transformed as

$$\frac{dE}{dt} = -P(t) \frac{dV}{dt} \Rightarrow \alpha \left( 4T^3 \frac{dT}{dt} V + T^4 \frac{dV}{dt} \right) = -\frac{1}{3} \alpha T^4 \frac{dV}{dt} \Rightarrow \frac{1}{T} \frac{dT}{dt} = -\frac{1}{3V} \frac{dV}{dt}, \quad (58)$$

where the second identity arises since for CMB photons  $E = \varepsilon V = \alpha T^4 V$  and  $P = \alpha T^4 / 3$ . Now, since  $V \propto a(t)^3$  as the box expands, this means that the rate in change of the photons' temperature is related to the rate of expansion of the universe by the relation

$$\frac{d}{dt}(\ln T) = -\frac{d}{dt}(\ln a), \quad (59)$$

implying the simple relation that  $T(t) \propto a(t)^{-1}$ ; the temperature of the CMB has dropped by a factor of roughly 1000 since the universe became transparent because the scale factor  $a(t)$  has increased by a factor of 1000 since then. Hence, at the time of recombination the CMB would have been a CNIRB, with a temperature slightly cooler than the surface of the star Betelgeuse (Ryden 2002, pg. 29).

#### BIG BANG NUCLEOSYNTHESIS

According to the Big Bang scenario, the universe was much hotter and denser at earlier times today. If we go far enough back in time when the temperature was on the order of  $T \sim \text{MeV}/k$ , there were no neutral atoms or even bound nuclei. The vast amounts of radiation in such a hot environment ensured that any atom or nucleus produced would immediately be destroyed by a high energy photon. As the universe cooled well below the binding energies of typical nuclei, light elements began to form. Knowing the conditions of the early universe and the relevant nuclear cross-sections, we can calculate the expected primordial abundances of all elements. Figure 6 shows the predictions of BBN for the light element abundances; boxes and arrows showing current estimates. These are consistent with the predictions, and this consistency test provides yet another ringing confirmation of the Big Bang (Dodelson 2003, pg. 9).

The light element measurements do more than just support the Big Bang model. Since the theoretical predictions depend on the density of protons and neutrons (i.e. **baryons**) at the time of nucleosynthesis, BBN gives us a way of measuring the baryon density in the universe. Since we know how these densities scale as the universe evolves (they fall as  $a(t)^{-3}$ ), we can turn the measurements of light element abundances into measures of the baryon density today. In particular, the measurement of primordial deuterium pins down the baryon density extremely accurately to only a few percent of the critical density;  $\Omega_b$  is at most 0.05 of the critical density. Since the total matter density is almost certainly higher, nucleosynthesis provides a compelling argument for nonbaryonic dark matter. The deuterium measurements are obtained by looking at the spectra of distant quasars,

at a time before much processing would have altered the primordial abundances (i.e. the gas in quasars should be metal-poor and thus mostly unaffected by stellar nucleosynthesis). The key absorption feature arises from the Ly $\alpha$  (transition from  $n = 2$  to  $n = 1$ ) absorption trough of H. As shown in Figure 6, the corresponding line for deuterium is (i) shifted somewhat (lower electron-nucleus reduced mass in D) and (ii) much less damped due to its lower abundance (Dodelson 2003, pg. 12).

#### *OTHER SUPPORTING EVIDENCE*

Another piece of evidence supporting the idea of the Big Bang scenario arises from observing the evolutionary properties of galaxies. A combination of observations and theory suggest that the first quasars and galaxies formed  $\sim 1$  Gyr after the Big Bang, and since then larger structures have been forming (i.e. the hierarchical model); distant galaxies are mainly small irregular galaxies. Moreover, a collection of quasars, which are believed to be powered by supermassive black holes, have been observed at  $z \leq 2$ . This suggests that there was a particular epoch when the conditions for supporting quasars were right (i.e. large quantities of dust and free gas near their cores). Additional supporting arguments can be made by noting that the ages of the oldest observed stellar objects correspond roughly with the Hubble time, and that dark matter – which is required to explain BBN – has been found to exist (Etsuko).

**QUESTION 5**

**Define and describe the “tired light hypothesis” and the “steady state universe” as alternatives to the Big Bang. How have they been disproved observationally?**

## QUESTION 5

**Define and describe the “tired light hypothesis” and the “steady state universe” as alternatives to the Big Bang. How have they been disproved observationally?**

### STEADY STATE THEORY

The **steady state theory** is based on the notion that the Universe is spatially homogeneous and isotropic on the large-scale and, moreover, it is invariant under time translations, that is, it is in a steady state (Peebles 1993, pg. 200). This theory accommodates Hubble’s Law by supposing that Hubble’s constant, defined as  $\dot{a}/a = H_o$ , is constant in time, thus implying that the universe expands exponentially as  $a \propto e^{H_o t}$ . In this theory there is no concept of a beginning nor end and therefore no Big Bang.

The steady state requires that the matter density of the Universe remain constant in time, despite an exponentially expanding universe. To allow for this it is proposed that a *creation field* provides a constant and steady creation of new matter (Partridge 2006, pg. 37). Galaxies continuously spawn out of the newly created matter, thereby filling in the gaps between already existing galaxies as the Universe expands. This process suggests that distant objects are on average the same as those we observe nearby. There are a number of observational tests, however, showing a clear presence of galactic evolution with redshift, seriously casting doubt on the steady state theory. In particular, it has been observed that at  $z \gtrsim 0.3$  galaxies in rich clusters tend to be bluer than the typical cluster member at low redshift. The interpretation is that the high-redshift cluster galaxies are undergoing a burst of star formation, consistent with what we would expect if we were observing them as they were when they were young. It is found that the rate of this occurrence of this trend is diminished by at least an order of magnitude at the present epoch than at  $z \sim 0.5$  (Peebles 1993, pg. 202).

Perhaps the most compelling argument against the steady state theory is the existence of the CMB. If the Universe is the same at all times, then it is hard to imagine a process responsible for the thermalized blackbody radiation we observe today. It would be reasonable to suppose that the radiation would be created by the creation field, but unreasonable to expect that the sum of its contribution over all redshifts adds up to precisely a blackbody spectrum. An attempt at remedying this is to suppose that the newly created radiation is absorbed and reradiated by intergalactic dust grains of a fixed temperature (Peebles 1993, pg. 203). This process is inadequate on two counts. First, the final arrangement is not completely blackbody and does not agree at all with the observed CMB. Second, such a process would require the Universe be extremely opaque to radio radiation inasmuch as the attenuation from  $z = 2$  would be roughly five orders of magnitude. This is, of course, in complete disagreement with our observations of radio galaxies at  $z > 2$  and with their apparently normal ratios of radio to optical luminosities (Peebles 1993, pg. 204).

### TIRED LIGHT HYPOTHESIS

The **tired light hypothesis** questions the assumption that the observed redshifts in the spectral lines of distant galaxies arises from the Doppler shift. Instead it proposes that photons moving through empty space lose energy at a rate

$$\frac{d\nu}{dl} = -H_o \nu, \quad (60)$$

where  $H_o$  is a constant and  $dl$  is the proper displacement along the path of the light in a static world (Peebles 1993, pg. 225). We see that this theory interprets the Hubble constant to describe the rate at which photons grow ‘tired’ by some process as they propagate through space. The redshifts we observe from distant galaxies is therefore not the result of expansion but rather the result of cosmological tiring of photons in a non-expanding Universe. Since the Universe is not expanding in this scenario we do not infer a Big Bang.

One immediate problem with this theory is the CMB. If we start out with a blackbody background in this Universe, then after a given time the tiring of photons will cause all photons to lose a given fraction of their energy, resulting in a distinctly non-thermal spectrum. One could attempt to remedy this by assuming some thermalizing process, but this would lead to the same problems experienced in the steady state model (Peebles 1993, pg. 225).

A more direct argument against the tired light hypothesis deals with observations of the surface brightness (SB) of a *standard source* at different redshifts. Suppose first that the tired light hypothesis is true. In this case, the flux from a standard source at redshift  $z$  decreases by a factor of  $(1+z)$  due to the tiring described in equation (60). Now we consider the case where the expansion of the universe is real. In this case, the flux from a standard source at redshift  $z$  decreases by two factors of  $(1+z)$ . The first factor is from the redshift due to expansion and the second factor arises from the stretching of the path length, diluting the reception rate of the photons at the detector (Ryden 2002, pg. 109). Another effect contributing to the SB in this case is the apparent area of the source. In an expanding universe the observed area of the source is increased by two factors of  $(1+z)$  due to the cosmological *aberration* for each of the two dimensions defining the area of the source (Sandage & Perelmuter 1991). Though the derivation of this aberration is much deeper than simple aberration, it can be thought classically in that the source appears larger since the photons we observe were emitted when the object was closer than it is today. Through all of this we see that in the tired light scenario the SB of a standard source decreases with redshift as  $(1+z)$  while in an expanding universe it decreases as  $(1+z)^4$ .

Observational tests of this type are difficult since there are no “standard” galaxies to compare with each other, one at high redshift and the other at small. Nevertheless, Sandage (2010) have used the HST to study this relation with elliptical galaxies in three remote clusters near  $z = 0.85$ . Their results are consistent with having an expanding universe and therefore dispel the tired light hypothesis.

**QUESTION 6**

Sketch a graph of recession speed vs. distance for galaxies out to and beyond the Hubble distance.

## QUESTION 6

**Sketch a graph of recession speed vs. distance for galaxies out to and beyond the Hubble distance.**

When we look at a galaxy at visible wavelengths, we are primarily detecting the light from the stars which the galaxy contains. Thus, when we take a galaxy's spectrum at visible wavelengths, it typically contains absorption lines created in the stars' relatively cool upper atmospheres<sup>1</sup>. Suppose we consider a particular absorption line whose wavelength, as measured in a laboratory here on Earth, is  $\lambda_{\text{em}}$ . The wavelength we measure for the same absorption line in a distant galaxy's spectrum,  $\lambda_{\text{ob}}$ , will not, in general, be the same. We say that the galaxy has a **redshift**  $z$ , given by the formula

$$z \equiv \frac{\lambda_{\text{ob}} - \lambda_{\text{em}}}{\lambda_{\text{em}}}. \quad (61)$$

Strictly speaking, when  $z < 0$ , this quantity is called a **blueshift**, though the vast majority of galaxies have  $z > 0$  (Ryden 2002, pg. 15).

In 1929 Edwin Hubble measured the distances to galaxies with known redshifts by using Cepheid variables as standard candles. Then, from a plot of redshift  $z$  versus distance  $r$ , he found the famous linear relation now known as **Hubble's Law**:

$$z = \frac{H_0}{c} r = \frac{r}{L_0}, \quad (62)$$

where  $H_0$  is called the Hubble constant and  $L_0 = c/H_0$  is the Hubble distance. Hubble interpreted the observed redshift of galaxies as being a Doppler shift due to their radial velocity away from Earth. Since the values of  $z$  in Hubble's analysis were all small ( $z < 0.04$ ), he was able to use the classical, nonrelativistic relation for the Doppler shift,  $z = v/c$ , where  $v$  is the galaxy's recessional speed. Interpreting the redshifts as Doppler shifts, Hubble's law takes the form

$$v = H_0 r. \quad (63)$$

Since the Hubble constant  $H_0$  can be found by dividing velocity by distance, it is customarily written in the rather baroque units of  $\text{km s}^{-1} \text{Mpc}^{-1}$  (Ryden 2002, pg. 17). Moreover, due to the uncertainties in measuring its true value, it is often quoted in the form

$$H_0 = 100h \text{ km s}^{-1} \text{ Mpc}^{-1}, \quad (64)$$

with modern measurements of nearby galaxies placing  $h = 0.72$  (Dodelson 2003, pg. 5).

### VELOCITY-DISTANCE RELATION

In an expanding homogeneous and isotropic space, let comoving markers  $A, B, C, \dots$  be equally spaced on a straight line, with a separating comoving distance  $d_0$  between them. Homogeneity requires that if  $B$  recedes from  $A$  at a velocity  $\alpha d_0$ , then also  $C$  simultaneously recedes from  $B$  at velocity  $\alpha d_0$ . Hence,  $C$  recedes from  $A$  at velocity  $2\alpha d_0$ ,  $D$  at velocity  $3\alpha d_0$ , and so on, thus illustrating the tape-measure nature of proper distance in uniform space, and the essential linearity of the **velocity-distance relation**:

$$v = H(t)r, \quad (65)$$

where  $H(t)$  is constant in space, but not necessarily in time (we have used this notation on purpose, but it should be stressed that there is no reason why  $H(t)$  here should be associated with  $H_0$  above). This constant can be identified using the notion of a scale factor,  $a(t)$ , which describes how a comoving coordinate  $x$  is stretched in time as  $r(t) = a(t)x$ . The velocity we would measure for the markers moving away from us at time  $t$  would thus be

$$v(r, t) = \frac{d}{dr} r(t) = \frac{da}{dr} x \equiv \frac{\dot{a}}{a} r \Rightarrow \text{thus } H(t) \equiv \frac{\dot{a}}{a}. \quad (66)$$

Hence, if we associate the redshifts observe by Hubble to correspond to cosmological expansion, rather than to Doppler shifts caused by peculiar velocities, we can relate equation (65) to equation (63), and thereby conclude that  $H_0$  is related to the current scale factor  $a_0$  through the relation  $H_0 = \dot{a}_0/a_0$ . We note that the spatially linear form of equation (65) arose from demanding homogeneity. The implications of this linear velocity-distance law seem startling at first since it implies that the recession velocity has no upper limit. In fact, for distances greater than the Hubble distance,  $L(t) = c/H(t)$ , equation (65) implies that galactic recession speeds should be greater than  $c$ . However, we must remember that when describing cosmological expansion, it is insufficient to invoke special relativity, and general relativity must be used instead. Moreover, as we know there is no contradiction with SR when superluminal motion occurs *outside the observer's inertial frame*. GR was specifically derived to be able to predict motion when global inertial frames were not available. Galaxies that are receding from us superluminally can be at rest locally (zero peculiar velocity) and motion in their local inertial frames (some small neighbourhood around them) remains well described by SR. They are in no sense catching up with photons. Rather, the galaxies and photons are both receding from us at recession velocities greater than the speed of light (Davis & Lineweaver 2004).

We should stress that the distance  $r$  used in equation (65) represents the *proper distance* as would be measured if we could instantaneously extend a tape measure to the distant galaxy *today*. As we have already noted, the linearity in that  $v \propto r$  is valid for all proper distances  $r$ . Thus, a plot of recession speed versus *proper* distance would be a straight line whose slope changes in time

<sup>1</sup> Galaxies containing AGN will also show emission lines from the hot gas in their nuclei.

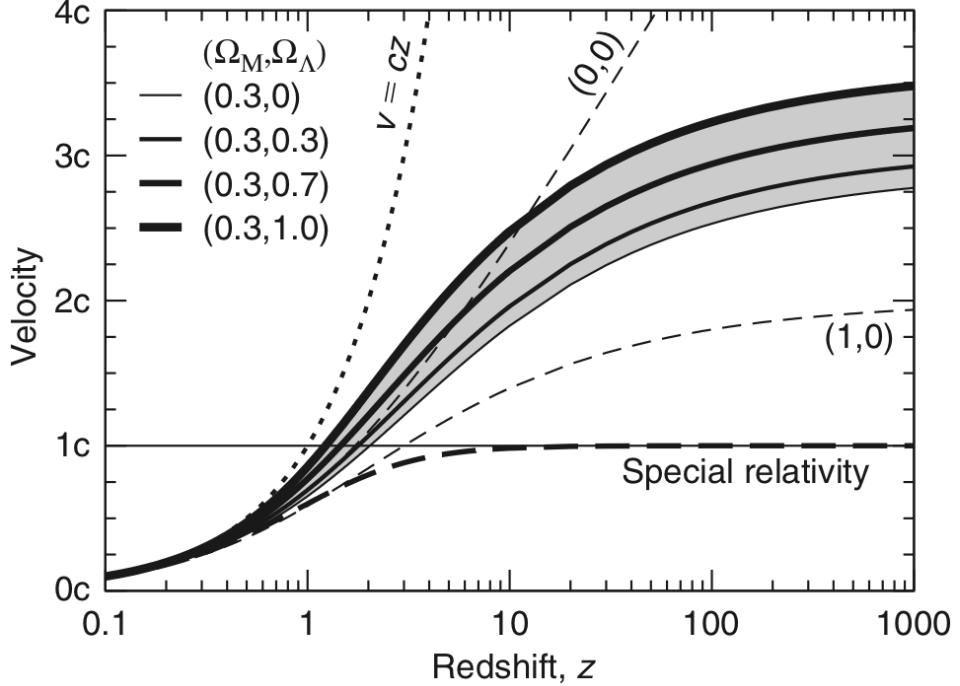


FIG. 7.— Recession velocity as a function of redshift for different cosmological models. The solid dark lines and grey shading show a range of FRW models as labelled in the legend. The recession velocity of all galaxies with  $z \gtrsim 1.5$  exceeds the speed of light in all viable cosmological models. The curve labelled  $v = cz$  is the low-redshift linear approximation, while the dark dashed curve is the SR result computed from equation (69). Image taken from Davis & Lineweaver (2004).

according to how  $H(t)$  evolves. This is *always* true and a common misconception that has arisen due to unclear notions about how to measure distance in an expanding space. This has probably arisen since proper distance is not practically measurable, though this should not be the reason for rejecting its conceptual utility (Harrison 1993).

If we wish to plot recession velocity versus some other distance measure, say something more observationally practical like redshift or luminosity distance, then we will in general *not* obtain a linear relationship. Strictly speaking, with the exception of cosmological models in which  $H(t)$  does not change in time, the linear relation in  $z$  given by equation (62) is only valid at low redshifts  $z \ll 1$ , and therefore distances  $r \ll L_0$ . Because the various observational “distances” (e.g., angular and luminosity distance) introduce differences of second and higher order in  $z$ , equation (62) in its validity ( $z \ll 1$ ) holds for most definitions of distance (Harrison 1993). The point is to stress that redshift,  $z$ , and proper distance,  $r$ , are not linearly related to each other, and that the velocity-distance relation is only strictly linear for the latter. To convert between the two distances it is easiest to think about what the comoving distance,  $x$ , is to an object at scale factor  $a$ . Well, in a time  $dt$ , light travels a coming distance  $dx = cdt/a$  and therefore

$$x = \int_{t(a)}^{t_0} \frac{cdt'}{a(t')} = \int_a^1 \frac{cda'}{a'^2 H(a')} \Rightarrow r = \int_0^z \frac{cdz'}{H(z')}. \quad (67)$$

The second identity is true as long as we are measuring at the current epoch (i.e.  $x = r$ ) (Dodelson 2003, pg. 34). Thus, using equation (67) and plugging into equation (65) yields the **velocity-redshift relation**:

$$v = cH_0 \int_0^z \frac{dz'}{H(z')}, \quad (68)$$

where we have explicitly evaluated for the current epoch.

Figure 7 shows  $v$  versus  $z$  obtained by numerically integrating equation (68) for different cosmological models. For comparison, the low-redshift linear approximation  $v = cz$  is plotted, and shows a much quicker divergence to infinite speeds. In addition, the result that would be obtained if we assumed that redshifts were produced by the peculiar velocities of galaxies is shown by solving the relativistic Doppler formula,

$$v = c \frac{(1+z)^2 - 1}{(1+z)^2 - 1}. \quad (69)$$

In this case, the maximum recessional speed is bound by  $c$  and a completely different physical interoperation of the universe is assumed. On the other hand, the correct formula given by equation (68), finds that recessional velocities  $v > c$  by  $z \gtrsim 1.5$  for all plausible cosmological models (Davis & Lineweaver 2004).

It may seem unusual that we are able to observe galaxies that are currently moving superluminally. However, it is important to stress that the Hubble sphere is *not* a cosmological horizon. Light that superluminal objects emit propagates toward us with a local peculiar velocity  $c$ , but since the recession velocity at that distance exceeds  $c$ , the total velocity of the light is away from

us. However, although the photons are in the superluminal region and therefore recede from us (in proper distance), the Hubble sphere also recedes. In deceleration universes,  $\dot{H}$  decreases as  $\dot{a}$  decreases, causing the Hubble sphere to expand. In accelerating universes,  $H$  also tends to decrease since  $\dot{a}$  increases more slowly than  $a$ . As long as the Hubble sphere recedes faster than the photons immediately outside it,  $\dot{L} > v - c$ , the photons end up in the subluminal region and approach us. Thus photons near the Hubble sphere that are receding slowly are overtaken by the more rapidly receding Hubble sphere. It is the **particle horizon**<sup>2</sup>, not the Hubble sphere, that marks the size of our observable universe because we cannot have received light from, or sent light to, anything beyond the particle horizon. Our effective particle horizon is the CMB, because we cannot use photons to see beyond the surface of last scattering (things like GW can though!) (Davis & Lineweaver 2004).

<sup>2</sup> The particle horizon is the distance that light could have travelled from time  $t = 0$  to now, and is evaluated by placing  $z = \infty$  in equation (67).

**QUESTION 7**

**What happened in the first 3 minutes after the Big Bang? Why is only He (and tiny traces of Li) synthesized in the Big Bang?**

## QUESTION 7

**What happened in the first 3 minutes after the Big Bang? Why is only He (and tiny traces of Li) synthesized in the Big Bang?**

### NEUTRINO DECOUPLING

A basic understanding of the interaction rates of neutrinos enables us to argue that neutrinos were once kept in equilibrium with the rest of the plasma. At late times, they lost contact with the plasma because their interactions are *weak*. Nonetheless, their distribution remained Fermi-Dirac (they are fermions) with their temperature simply falling as  $T_\nu \propto a^{-1}$ . The main task therefore is to relate the neutrino temperature to the photon temperature today. The trick part of this is that the annihilation of electrons and positrons when the cosmic temperature was on the order of the electron mass. Neutrinos lost contact with the cosmic plasma slightly before this annihilation so they did not inherit any of the associated energy. The photons, which did, are thus hotter than the neutrinos. A calculation which takes into account the conservation of entropy during electron-positron annihilation shows that

$$\frac{T_\nu}{T_\gamma} = \left( \frac{4}{11} \right)^{1/3}, \quad (70)$$

where the factor of 4/11 basically arises from the differing statistics and degeneracies of the photons, neutrinos, and electrons/positrons (Dodelson 2003, pg. 45). Note that pair annihilation of electrons and positrons occurred when the cosmic temperature was  $kT \sim 500$  keV, or  $t \approx 4$  seconds after the Big Bang.

In order for particles to maintain equilibrium with each other their mutual reactions must occur at a sufficient rate. The equilibrium state, specified by the temperature, continuously changes, so that the particle distribution needs to continually adjust to this changing equilibrium. This is possible only if the mean time between two reactions is much shorter than the time-scale on which equilibrium conditions change. The latter is given by the expansion. This means that the reaction rates (the number of reactions per particle per unit time) must be larger than the cosmic expansion rate  $H(t)$  in order for the particles to maintain equilibrium. The reaction rates  $\Gamma_{\text{weak}} \sim n\sigma$ , where both number density  $n$  and cross-section  $\sigma$  decrease in time:  $n$  due to expansion and  $\sigma$  for neutrinos is related to the weak interaction, thus falling with temperature. At sufficiently early times, the reaction rates were larger than the expansion rate, and thus neutrinos remained in equilibrium. Later, however, the reactions no longer took place fast enough to maintain equilibrium. From investigating the cross-section of weak interactions, it turns out that at  $T \lesssim 10^{10}$  K ( $t \approx 1$  second after the Big Bang) neutrinos were no longer in equilibrium. This process of decoupling from the other particles is also called **freeze-out**. After freeze-out, neutrinos moved freely throughout the universe, and at present have a temperature of 1.9 K and number density of  $100 \text{ cm}^{-3}$ . However, these neutrinos are currently undetectable because of their extremely low cross-section (Schneider 2002, pg. 162).

### PRIMORDIAL NUCLEOSYNTHESIS

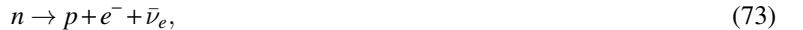
BBN occurs roughly within the first three minutes. During this period the universe was radiation-dominated so that the expansion followed the simple power-law form  $a(t) \propto t^{1/2}$ , and the temperature of blackbody photons evolved as

$$T(t) \approx 10^{10} \text{ K} \left( \frac{t}{1 \text{ s}} \right)^{-1/2} \Leftrightarrow kT(t) \approx 1 \text{ MeV} \left( \frac{t}{1 \text{ s}} \right)^{-1/2}. \quad (71)$$

The basic building blocks for nucleosynthesis are neutrons and protons. The rest energy of a neutron is greater than that of a proton by a factor

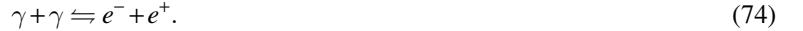
$$Q_n = m_n c^2 - m_p c^2 = 1.29 \text{ MeV}. \quad (72)$$

A free neutron is unstable, decaying via the reaction



with a decay timescale of  $\tau_n = 890$  s (at  $\tau_n$  its abundance drops by  $e^{-1}$ ). With a decay time of only fifteen minutes, the existence of a free neutron is as fleeting as fame; once the universe was several hours old, it contained essentially no free neutrons. However, a neutron which is bound into a stable atomic nucleus is preserved against decay. There are still neutrons around today, because they've been tied up in deuterium, helium, and other atoms (Ryden 2002, pg. 213).

Let's consider the state of the universe when its age is  $t = 0.1$  s. At that time, the temperature was  $T \approx 3 \times 10^{10}$  K, and the mean energy per photon was  $E_{\text{mean}} \approx 10$  MeV. This energy is much greater than the rest energy of a electron or positron, so there were positrons as well as electrons present at this time, created by pair production:



Additionally, at this time neutrons and protons were in equilibrium with each other, via the interactions



As long as neutrons and protons are kept in equilibrium by the reactions (75), their number density is given by the Maxwell-Boltzmann equation"

$$n_n = g_n \left( \frac{m_n k T}{2\pi\hbar^2} \right)^{3/2} \exp \left( -\frac{m_n c^2}{k T} \right) \quad \text{and} \quad n_p = g_p \left( \frac{m_p k T}{2\pi\hbar^2} \right)^{3/2} \exp \left( -\frac{m_p c^2}{k T} \right) \Rightarrow \frac{n_n}{n_p} = \left( \frac{m_n}{m_p} \right)^{3/2} \exp \left[ -\frac{(m_n - m_p)c^2}{k T} \right], \quad (76)$$

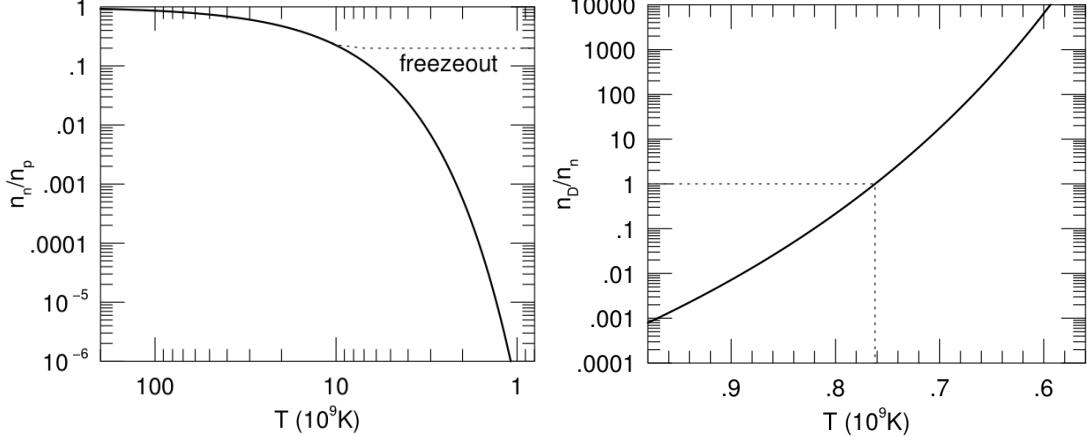


FIG. 8.— (left) Neutron-to-proton ratio in the early universe assuming equilibrium between the two species. The dotted line gives the ratio after proton-neutron (i.e. neutrino) freeze-out. (right) The deuterium-to-neutron ratio during the epoch of deuterium synthesis. Images taken from Ryden (2002).

since  $g_n = g_p = 2$ . This equation is simplified by discarding the first term in parenthesis since  $m_n \approx m_p$  and noting that the term in the exponential is simply the difference in rest energy, so that

$$\frac{n_n}{n_p} = \exp\left(-\frac{Q_n}{kT}\right) \quad (77)$$

Figure 8 plots this relation as a function of temperature. At temperatures  $kT \gg Q_n = 1.29$  MeV, corresponding to  $T \gg 2 \times 10^{10}$  K and  $t \ll 1$  s, the number of neutrons is nearly equal to the number of protons. However, as the temperature starts to drop below this limit, protons begin to be strongly favoured, and the neutron-to-proton ratio plummets exponentially (Ryden 2002, pg. 214).

If the neutrons and protons remained in equilibrium, then by the time the universe was six minutes old, there would be only one neutron for every million protons. However, neutrons and protons do not remain in equilibrium for nearly that long. The interactions which mediate between neutrons and protons in the early universe, shown in equation (75), involve the interaction of a baryon with a neutrino (or anti-neutrino). Neutrinos interact with baryons via the weak nuclear force, and we have already seen that they decouple from the cosmic plasma when  $T_{\text{freeze}} = 10^{10}$  K. Plugging this value into equation (76) we see that  $n_n/n_p \approx 0.2$  at neutrino decoupling. Hence, at times  $t_{\text{freeze}} < t \ll \tau_n$ , there was one neutron for every five protons in the universe (Ryden 2002, pg. 216).

It is the scarcity of neutrons relative to protons that explains why Big Bang Nucleosynthesis was so incomplete, leaving three-fourths of the baryons in the form of unfused protons. A neutron will fuse with a proton much more readily than a proton will fuse with another proton. The reason is that the former involves a strong interaction while the latter involves a weak interaction, and the cross-section for interactions involving the strong nuclear force are much larger than for those involving the weak nuclear force. In addition, proton-proton fusion demands that a larger Coulomb barrier be overcome. Note that proton-proton occurs in the Sun simply because the stellar interior is a stable environment and allows ample time for the reaction to complete. In the early universe, by strong contrast, the temperature and density drops so rapidly, that after less than one hour this process cannot occur. For the sake of completeness, neutron-neutron fusion, which is governed by the weak interaction, does not occur due to its low cross-section and low neutron abundance. For this reason, we state, as a first order approximation, that BBN proceeds until every free neutron is bonded into an atomic nucleus, with the leftover protons remaining solitary. In this approximation, we can compute the maximum possible value of  $Y$ , the fraction of the baryon mass in the form of  ${}^4\text{He}$ , by supposing that every neutron present after the proton-neutron freeze-out is incorporated into a  ${}^4\text{He}$  nucleus. Given a neutron-to-proton ratio of  $n_n/n_p = 1/5$ , we can consider a representative group of 2 neutrons and 10 protons. The 2 neutrons can fuse with 2 of the protons to form a single  ${}^4\text{He}$  nucleus. The remaining 8 protons, though, will remain unfused, and so

$$Y_{\max} = \frac{4}{12} = \frac{1}{3}. \quad (78)$$

Indeed, this value is larger than the observed  $Y = 0.24$ , indicating that we are on the right track (Ryden 2002, pg. 217).

Let's now move forward in time to  $t = 2$  seconds where, although neutrinos have frozen out, photons still remain strongly coupled to protons and neutrons. BBN takes place through a series of two-body reactions, building heavier nuclei step by step. The essential first step in BBN is the fusion of a proton and a neutron to form a deuterium nucleus:



When a proton and a neutron fuse, the energy released (and carried away by a gamma ray) is the binding energy of a deuterium nucleus

$$B_D = (m_n + m_p - m_D)c^2 = 2.22 \text{ MeV}. \quad (80)$$

Conversely, a photon with  $h\nu > B_D$  can photodissociate a deuterium nucleus into its component proton and neutron. Equation (79) is structurally equivalent to the recombination of hydrogen, and so can be thought of in a similar manner. As a result, around

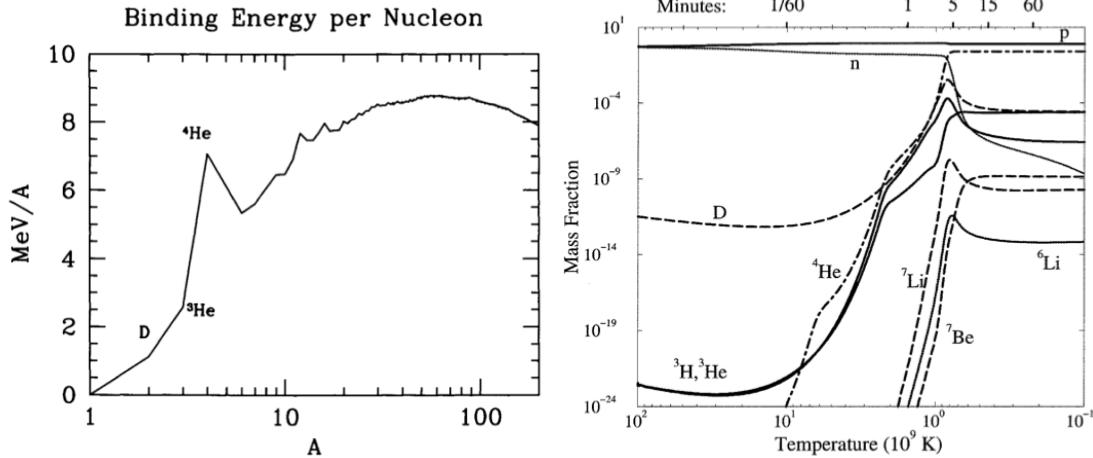


FIG. 9.— (left) Binding energy of nuclei as a function of mass number. Iron has the highest binding energy, but among the light elements  ${}^4\text{He}$  is a crucial local maximum. Nucleosynthesis in the early universe essentially stops at  ${}^4\text{He}$  because of the lack of tightly bound isotopes at  $A = 5 - 8$ . In the high-density environment of stars, three  ${}^4\text{He}$  nuclei fuse to form  ${}^{12}\text{C}$ , but the low baryon number precludes this process in the early universe. Image taken from Dodelson (2003). (right) The evolution of abundances of the light elements formed in BBN, as a function of temperature (lower axis) and cosmic time  $t$  (upper axis). The decrease in neutron abundance in the first  $\sim 3$  min is due to neutron decay. The density of deuterium increases steeply – linked to the steep decrease in neutron density – and reaches a maximum at  $t \sim 3$  min because then its density becomes sufficiently large for efficient formation of  ${}^4\text{He}$  to set in. Only a few deuterium nuclei do not find a reaction partner and remain, with a mass fraction of  $\sim 10^{-5}$ . Only a few other light nuclei are formed in the Big Bang, mainly  ${}^3\text{He}$  and  ${}^7\text{Li}$ . Image taken from Schneider (2002).

the time of deuterium synthesis, the relative numbers of free protons, free neutrons, and deuterium nuclei is given by an equation directly analogous to the Saha equation for ionization:

$$\frac{n_D}{n_p n_n} = \frac{g_D}{g_p g_n} \left( \frac{m_D}{m_p m_n} \right)^{3/2} \left( \frac{kT}{2\pi\hbar^2} \right)^{-3/2} \exp \left[ \frac{(m_p + m_n - m_D)c^2}{kT} \right] \Rightarrow \frac{n_D}{n_n} = 6n_p \left( \frac{m_n kT}{\pi\hbar^2} \right)^{-3/2} \exp \left( \frac{B_D}{kT} \right), \quad (81)$$

where the second relation arises from using equation (80),  $g_D = 3$ ,  $g_n = g_p = 2$ , and the approximation that  $m_n = m_p = m_D/2$ . Physically, equation (81) tells us that deuterium is favoured in the limit  $kT \rightarrow 0$ , and that free protons and neutrons are favoured in the limit  $kT \rightarrow \infty$  (Ryden 2002, pg. 219).

We can use equation (81) to write the deuterium-to-neutron ratio as a function of  $T$  and the baryon-to-photon ratio  $\eta$  if we make some simplifying assumptions. Even today, we know that  $\sim 75\%$  of all the baryons in the universe are in the form of unbound protons. Before the start of deuterium synthesis, 5 out of 6 baryons (or  $\sim 83\%$ ) were in the form of unbound protons. Thus, if we don't want to be fanatical about accuracy, we can write

$$n_p \approx 0.8n_{\text{bary}} = 0.8\eta n_\gamma = 0.8\eta \left[ 0.243 \left( \frac{kT}{hc} \right) \right] \quad \text{where } \eta = 5.5 \times 10^{-10}. \quad (82)$$

Substituting this result into equation (81) we find that

$$\frac{n_D}{n_n} \approx 6.5\eta \left( \frac{kT}{m_n c^2} \right)^{3/2} \exp \left( \frac{B_D}{kT} \right). \quad (83)$$

This function is plotted in Figure 8 and shows the temperature  $T_{\text{nuc}}$  at which we say that deuterium fusion takes place, defined such that  $n_D/n_n = 1$ . The result is that  $kT_{\text{nuc}} \approx 0.07$  MeV corresponding to an age of  $t_{\text{nuc}} \approx 200$  seconds. This time delay is not negligible compared to the neutron decay time; at this time the  $n_n/n_p \approx 0.15$ . This in turn lowers the maximum possible  ${}^4\text{He}$  fraction of equation (78) to  $Y_{\text{max}} \approx 0.27$  (Ryden 2002, pg. 222). Physically, this time delay is attributed to the requirement that high-energy photons capable of dissociating deuterium through the inverse of equation (79) dilute sufficiently enough that it can be stably made. The time delay is often called the **deuterium bottleneck** since it is the limiting factor in the eventual synthesis of helium-4.

Of course, the deuterium-to-neutron abundance will not remain indefinitely at the equilibrium value given by equation (83). Once a significant amount of deuterium forms, there are many possible nuclear reactions available. For instance, a deuterium nucleus can fuse with a proton to form  ${}^3\text{H}$ , or alternatively, can fuse with a neutron to form  ${}^3\text{H}$  (known as **tritium**):



Tritium is unstable; it spontaneously decays to  ${}^3\text{He}$ , but with a decay time of roughly 18 years, making it effectively stable during BBN. Nevertheless, neither a large amount of  ${}^3\text{H}$  nor  ${}^3\text{He}$  accrue, since soon after forming they are converted to  ${}^4\text{He}$  through reactions involving  $n$ ,  $p$ , and  $D$  nuclei. These reactions occur through the strong force, implying that they have large

cross-sections, occurring almost instantaneously. Once  ${}^4\text{He}$  is reached, however, the orderly march of nucleosynthesis to heavier and heavier nuclei reaches a roadblock. For such a light nucleus,  ${}^4\text{He}$  is exceptionally tightly bound, as illustrated in Figure 9. By contrast, there are no stable nuclei with  $A = 5$ . If you try to fuse a proton or neutron to  ${}^4\text{He}$ , it won't work;  ${}^5\text{He}$  and  ${}^5\text{Li}$  are not stable nuclei. Thus,  ${}^4\text{He}$  is resistant to fusion with protons and neutrons. Small amounts of  ${}^6\text{Li}$  and  ${}^7\text{Li}$ , the two stable isotopes of lithium, are made through reactions involving interactions between  ${}^4\text{He}$  and  $D$  and  ${}^3\text{H}$ . In addition, small amounts of  ${}^7\text{Be}$  are made by fusing  ${}^4\text{He}$  with  ${}^3\text{He}$ . The triple alpha process that converts  ${}^4\text{H}$  to  ${}^{12}\text{C}$  through the intermediate construction of  ${}^8\text{Be}$  is hindered since any  ${}^8\text{Be}$  formed will decay back to two  ${}^4\text{He}$  nuclei within  $10^{-16}$  seconds (Ryden 2002, pg. 223).

The bottom line is that once deuterium begins to be formed, fusion up to the tightly bound  ${}^4\text{He}$  nucleus proceeds very rapidly. Fusion of heavier nuclei occurs much less rapidly and determining precise values requires numerical work due to the complexity of reaction channels. The results of a typical BBN code is shown in Figure 9. Initially, at  $T \gg 10^9$  K, almost all the baryonic matter is in the form of free protons and neutrons. As the deuterium density climbs upward, however, the point is eventually reached where significant amounts of  ${}^3\text{H}$ ,  ${}^3\text{He}$ , and  ${}^4\text{He}$  are formed. By the time the temperature has dropped to  $T \sim 10^8$  K, at  $t \sim 10$  min, Big Bang Nucleosynthesis is essentially over. Nearly all the baryons are in the form of free protons or  ${}^4\text{He}$  nuclei. The small residue of free neutrons decays into protons. Small amounts of  $D$ ,  ${}^3\text{H}$ , and  ${}^3\text{He}$  are left over with  ${}^3\text{H}$  later decaying to  ${}^3\text{He}$ . Additionally, small amounts of  ${}^6\text{Li}$ ,  ${}^7\text{Li}$ , and  ${}^7\text{Be}$  are made, with the latter eventually being converted to  ${}^7\text{Li}$  through electron capture:  ${}^7\text{Be} + e^- \rightarrow {}^7\text{Li} + \nu_e$  (Ryden 2002, pg. 224).

#### COSMOLOGICAL PARAMETERS

The yields of  $D$ ,  ${}^3\text{He}$ ,  ${}^4\text{He}$ ,  ${}^6\text{Li}$ , and  ${}^7\text{Li}$  depend on various physical parameters. Most importantly, they depend on the baryon-to-photon ratio  $\eta$  since this controls the onset of deuterium burning. Broadly speaking, we know immediately that the baryon-to-photon ratio can't be as small as  $\eta \sim 10^{-12}$ . If it were, BBN would be extremely inefficient, and we would expect only tiny amounts of helium to be produced ( $Y < 0.01$ ). Conversely, we know that the baryon-to-photon ratio can't be as large as  $\eta \sim 10^{-7}$ . If it were, nucleosynthesis would have taken place very early (before neutrons had a chance to decay), the universe would be essentially deuterium-free, and  $Y$  would be near its maximum permissible value of  $Y_{\max} \approx 0.33$ . Pinning down the value of  $\eta$  more accurately requires making accurate observations of the primordial densities of the light elements; that is, the densities before nucleosynthesis in stars started to alter the chemical composition of the universe. In determining the value of  $\eta$ , it is most useful to determine the primordial abundance of deuterium. This is because the deuterium abundance is strongly dependent on  $\eta$  in the range of interest. As we mentioned in a previous question, this measurement boils down to looking at quasar spectra in order to evaluate the D/H ratio. In this process, we aren't looking for deuterium in the quasar itself, but using it as a flashlight to illuminate intervening galactic clouds. If an intergalactic gas cloud contains no detectable stars, and has very low levels of elements heavier than lithium, we can hope that its D/H value is close to the primordial value, and hasn't been driven downward by the effects of fusion within stars<sup>3</sup> (Ryden 2002, pg. 227).

#### WIMPS

One of the most promising candidates for DM are **weakly interacting massive particles** or **WIMPS**. From the considerations used above for neutrinos, we can make constraints on such a particle. In particular, if a WIMP is weakly interacting, it will decouple at the freeze-out temperature of  $\sim 10^{10}$  K determined from knowledge of the weak interaction cross-section. If its mass  $m_{\text{WIMP}}$  is smaller than the decoupling temperature ( $T \sim 1\text{ MeV}$ ), the WIMP was relativistic at the epoch of freeze-out and thus its current number density is the same as that of the neutrinos,  $n \sim 100 \text{ cm}^{-3}$ . Hence, we can compute the corresponding density parameter,

$$\Omega_{\text{WIMP}} h^2 = \frac{m_{\text{WIMP}}}{91.5 \text{ eV}}. \quad (85)$$

This equation is valid provided that  $m_{\text{WIMP}} \leq 1 \text{ MeV}$  and applies equally well to neutrinos. Since  $\Omega_m < 2$  certainly, we conclude from equation (85) that no stable weakly interacting particle can exist in the mass range of  $100 \text{ eV} \leq m \leq 1 \text{ MeV}$ . In particular, none of the three neutrinos can have a mass in this range, and neutrinos are thought to have upper mass bounds of  $m_\nu \leq 1 \text{ eV}$  (Schneider 2002, pg. 166).

On the other hand, if the WIMP is heavier than 1 MeV, it decouples at a time when it is already non-relativistic. Then the estimate of the number density, and with it equation (85), needs to be modified. In particular, if the WIMP mass exceeds that of the Z-boson ( $m_Z = 91 \text{ GeV}$ ), then

$$\Omega_{\text{WIMP}} h^2 \simeq \left( \frac{m_{\text{WIMP}}}{1 \text{ TeV}} \right)^2 \quad (86)$$

This means that a WIMP mass of  $m_{\text{WIMP}} \sim 100 \text{ Gyr}$  would provide a density of  $\Omega_{\text{WIMP}} \sim 0.3$ . The LHC should be able to detect such a particle if it really exists. In fact, arguably the most promising extension to the standard model of particle physics – the model of supersymmetry – predicts a stable particle with a mass of several hundred GeV, the **neutralino** (Schneider 2002, pg. 166).

<sup>3</sup> Deuterium fuses during the gravitational collapse of a molecular cloud.

**QUESTION 8**

**Explain how Supernovae (SNe of Type Ia in particular) are used in the measurements of cosmological parameters.**

### QUESTION 8

Explain how Supernovae (SNe of Type Ia in particular) are used in the measurements of cosmological parameters.

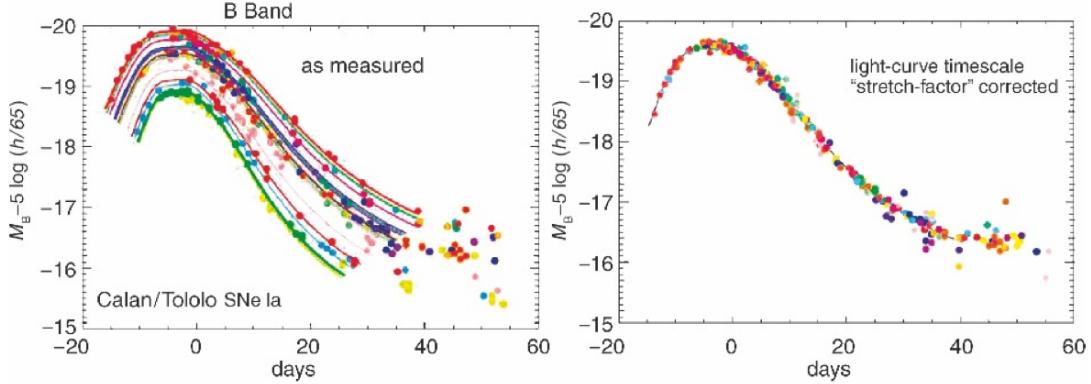


FIG. 10.— (left panel) B-band light curves of different SNe Ia. One sees that the shape of the light curves and the maximum luminosity of the SNeIa differ substantially among the sample. A transformation was found empirically with a single parameter described by the width of the light curve. By means of this transformation, the different light curves can all be made congruent (right panel). Image taken from Schneider (2002).

### STANDARD CANDLES

SNe make excellent long-range distance measures because of their intrinsic luminosities, which allow us to measure distances up to  $\sim 1$  Gpc. Of these, Type Ia SNe are the most homogeneous and therefore can be treated as standard candles (Charles). However, in reality they are not standard candles, since their maximum luminosity in the B band varies with a dispersion of about 0.4 mag. Fortunately, it turns out that there is a strong correlation between the luminosity and the shape of the light curve of SNe Ia. Those of higher maximum luminosity show a slower decline in the light curve, as measured from its maximum (Schneider 2002, pg. 324). This allows calibration between different SNe through one of two methods: the *multicolour light curve shapes* (MLCS) method, where light curves are fit to a family of parameterized curves to determine various properties, and the *stretch method*, which simply scales the fall-time of an observed SN Ia curve (the rise time is usually difficult to obtain) with that of a template light curve, and then scales the luminosity in the same manner; the latter method is shown in Figure 10. Not only do Type Ia SNe make excellent distance indicators, but detailed spectroscopy can be used to determine their line-of-sight velocities. In particular, the redshift of their emitted light is easily identified through the defining Si II absorption lines (Charles).

Below we describe how Type Ia SNe were used in relation to Hubble's Law and the expansion history of the universe. Before doing so, however, we first note that Type Ia SNe are not the only standard candles that can be used for this method. In particular, as long as some source with a known (and large) intrinsic luminosity exists, and a method is available for determining the redshift of its light, then that source can be used as a distance and recession indicator. For example, Type II-P SNe have also been used from the strong correlation existing between their plateau luminosity, and corresponding velocities obtained by locating their Fe II ( $\sim 5000$  Å) absorption lines.

### EXPANDING UNIVERSE

Comparing the maximum magnitude of the measured Type Ia SNe, or their distance modulus respectively, with that which would be expected for an empty universe ( $\Omega_{\text{mat}} = \Omega_\Lambda = 0$ ), one obtains a truly surprising result, as shown in Figure 11. Considering at first only the supernovae with  $z \lesssim 1$ , one finds that these are fainter than predicted even for an empty universe. It should be mentioned that the deceleration parameter  $q_0 = 0$  for an empty universe, so that  $\ddot{a} = 0$ . The luminosity distance in such a universe is therefore larger than in any other universe with a vanishing cosmological constant (since the inclusion of matter will act to decelerate the universe). The luminosity distance can only be increased by assuming that the universe expanded *more slowly* in the past than it does today, hence that *the expansion has accelerated over time*; this is only possible if  $\Omega_\Lambda > 0$ . This result is confirmed further by looking at higher redshift sources. For instance, in Figure 11 we see that the observed trend reverses for  $z \gtrsim 1$ , where Type Ia SNe become brighter than they would be in an empty universe. At these high redshifts the matter density dominates the universe, proceeding as  $a^{-3}$  in contrast to the constant vacuum energy. The corresponding constraints taken from fitting to the data are  $\Omega_{\text{mat}} = 0.27$  and  $\Omega_\Lambda = 0.73$  (Schneider 2002, pg. 327).

We can obtain a better physical understanding of these results by noting that in a flat universe (which, from separate reasoning, we have deduced is true for our universe), the luminosity distance,  $d_L$ , is related to the proper distance,  $d_p$ , of an object by the relation  $d_L = (1+z)d_p$ . This relation arises because an expanding universe causes the observed flux of light from a standard candle at redshift  $z$  to be decreased by a factor of  $(1+z)^2$ . The first factor arises from wavelength stretching and the second factor arises from the delay in photon detections induced by increased distance between photon packets emitted at different times. Since,  $d_L$  is inferred from the inverse-square law of flux decrement, we have that  $d_L^2 = (1+z)^2 d_p^2$ , and thus our relation above. In addition,  $d_p(t)$  is defined to as the length of the spatial geodesic between two points when the scale factor is fixed at the value  $a(t)$ . If we

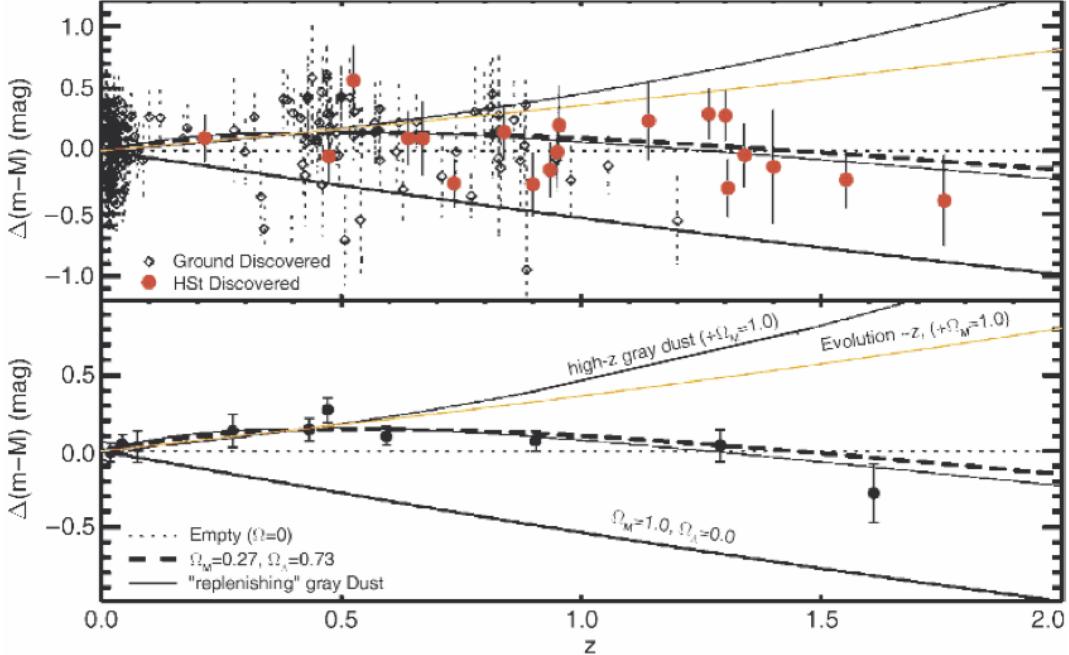


FIG. 11.— Difference between the maximum brightness of SNeIa and that expected in an empty universe ( $\Omega_{\text{mat}} = \Omega_{\Lambda} = 0$ ). Diamond symbols represent events that were detected from the ground, circles the ones discovered by the HST. In the top panel, the individual SNe Ia are presented, whereas in the bottom panel they are averaged in redshift bins. An empty universe would correspond to the dotted straight line,  $\Delta(m-M) = 0$ . The dashed curve corresponds to a cosmological model with  $\Omega_{\text{mat}} = 0.27$  and  $\Omega_{\Lambda} = 0.73$ . Image taken from Schneider (2002).

observe, at time  $t_0$ , light that was emitted by a distant galaxy at time  $t_e$ , the current proper distance to the galaxy is

$$d_p(t_0) = c \int_{t_e}^{t_0} \frac{dt}{a(t)}. \quad (87)$$

Using the Taylor approximation in equation (28) this can be rewritten as

$$d_p(t_0) = \frac{c}{H_0} z \left[ 1 - \frac{1+q_0}{2} z \right]. \quad (88)$$

Hence, if  $z \ll 1$ , this implies that

$$d_L \approx \frac{c}{H_0} z \left( 1 + \frac{1-q_0}{2} z \right), \quad (89)$$

We have discussed this earlier, in Question 2, using Ryden (2002) as a guide. From this discussion, we also learned that the current value of the deceleration parameter is

$$q_0 = \Omega_{\text{rad}} + \frac{1}{2} \Omega_{\text{mat}} - \Omega_{\Lambda} \approx \frac{1}{2} \Omega_{\text{mat}} - \Omega_{\Lambda}, \quad (90)$$

where the approximation arises since radiation has mostly diluted by the current epoch. We can now understand why  $d_L$  is larger in a  $\Lambda$ -dominated universe and thus why SNe appear fainter at  $z \lesssim 1$ . These objects appear fainter to us than if the universe were empty because, from equation (90), a cosmological constant increases expansion with time, so that light has taken more time to travel to us than in an empty universe, where the acceleration is constant in time. Conversely, at  $z \gtrsim 1$ , when the universe was matter-dominated, equation (90) tells us that the expansion would be slowing with time, so that light would reach detectors quicker than in an empty universe, thus appearing brighter. This turnaround will naturally occur at a higher redshift than matter- $\Lambda$  equality (roughly  $z = 0.4$ ) since sources from the matter-dominated era have to fight against the reversal of the  $\Lambda$ -dominated era (My logic; hopefully it is correct).

An equivalent description of these results is made through the use of a **redshift-magnitude diagram** which plots distance modulus,

$$\mu \equiv m - M = 5 \log \left( \frac{d_L}{10 \text{ pc}} \right), \quad (91)$$

against redshift. Through the use of equation (89), we can rewrite equation (91) as

$$m - M \simeq (42.38 - 5 \log(h)) + 5 \log(z) + 1.086(1 - q_0)z, \quad (92)$$

where the first term arises from the current value of  $H_0$ , and it is important to stress that this is only strictly valid for  $z \ll 1$  since it relied on a Taylor expansion of  $a(z)$ . Equation (92) tells us that a plot of  $(m-M)$  versus  $z$  will be at first linear, and then will curve upward. Accurately measuring this departure from a straight line allows the value of  $q_0$  to be determined, and consequently the values of  $\Omega_{\text{mat}}$  and  $\Omega_\Lambda$  (Carroll & Ostlie 2007, pg. 1213).

#### REJECTED ALTERNATE SCENARIOS

As we have mentioned, the above results indicate an expanding universe with a non-zero cosmological constant. Before this could be accepted, however, it was important to consider other scenarios that may explain why close SNe appear dimmer than expected. One possible scenario that was considered, was that about 20% of the light from distant SNe was absorbed at high  $z$  by some hypothetical “gray dust.” Or perhaps evolutionary effects were misleading the astronomers, since at high  $z$  we observed a younger generation of SNe, formed in a younger galactic environment where heavy elements were less abundant. However, both of these scenarios were rejected once the observational data had extended to higher redshift sources. In this case, the high- $z$  sampled appeared brighter than if the universe had expanded at a constant rate, explained by the matter-dominated deceleration phase of the early universe. These observations ruled out the possible gray dust and evolutionary effects (Carroll & Ostlie 2007, pg. 1213).

#### POSSIBLE SYSTEMATICS

There are a large number of possible systematic errors that could have affected the results described above; the main ones are:

- K-correction: The cosmological redshift affects the measurement of an object’s spectrum because these observations are usually made within a specific wavelength region. For example, observations made with the V-band at 550 nm can be affected as the cosmological redshift brings shorter-wavelength radiation into the V band. This effect can be corrected by adding a compensating term called the K-correction to equation (91) if the spectrum,  $I_\lambda$ , of the object is known (Carroll & Ostlie 2007, pg. 1214).
- UV Spread: Type Ia have a higher spread in the UV than in the optical or IR, which becomes problematic when cosmological expansion shifts the U band into the B band.
- Reddening: Intrinsic reddening of the SNe themselves and reddening due to dust should be handled separately, but in practice they are hard to deconvolve. Intrinsically fainter SNe Ia are redder than brighter ones, but the same effect occurs with dust extinction.
- Galactic Evolution: It is now well known that fainter SNe Ia tend to be embedded in older stellar populations; this translates to a  $\sim 12\%$  brightness increase for  $z = 1$  SNe Ia due to increased star formation.
- Curve Widths: The width of the light curve is larger for Type Ia SNe at higher redshift than it is for local objects. This arises from the cosmological expansion which delays light signals by a factor of  $(1+z)$  (Schneider 2002, pg. 326).

Despite this rather long list, the overall conclusion that  $\Omega_\Lambda > 0$  is highly robust in the face of any reasonable systematics (Charles).

**QUESTION 9**

**Rank the relative ages of the following universes, given an identical current-day Hubble constant for all of them: an accelerating universe, an open universe, a flat universe.**

### QUESTION 9

**Rank the relative ages of the following universes, given an identical current-day Hubble constant for all of them: an accelerating universe, an open universe, a flat universe.**

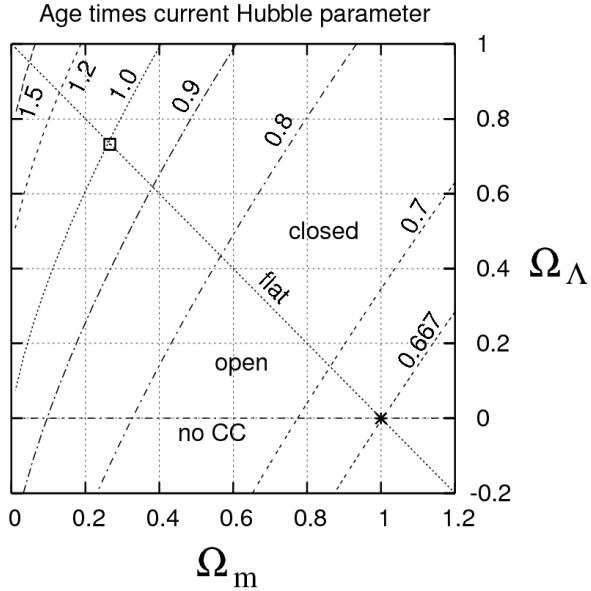


FIG. 12.— A plot showing the numerical value of the integral in equation (94) as a function of  $\Omega_m$  and  $\Omega_\Lambda$  with  $\Omega_{\text{rad}}$  held constant. The box in the upper left corner denotes the current estimates of these parameters from the WMAP 7-year data and the star in the bottom right corner denotes a matter-dominated, flat universe. This image, and the python code generating it, was taken from [http://en.wikipedia.org/wiki/Age\\_of\\_the\\_universe](http://en.wikipedia.org/wiki/Age_of_the_universe).

The Hubble constant is related to the scale factor through the relation  $H = \dot{a}/a$ , which can be easily rearranged to yield an expression describing the age of the universe:

$$t = \int_0^{a_0} \frac{da}{aH}. \quad (93)$$

This relation can be rewritten in terms of redshift  $z$  by noting that  $a/a_0 = 1/(1+z)$  so that  $da = -a_0/(1+z)^2 dz$  and therefore

$$t = \int_z^\infty \frac{dz}{(1+z)H(z)} = \frac{1}{H_0} \int_z^\infty \frac{dz}{(1+z)\sqrt{(1+z)^4\Omega_{\text{rad}} + (1+z)^3\Omega_m + \Omega_\Lambda + (1+z)^2\Omega_k}}. \quad (94)$$

In this expression  $\Omega_k$  is the curvature density defined as  $\Omega_k = -k/a_0 H_0^2 = 1 - \Omega_0$ . The integral in equation (94) is of order unity which allows a rough estimation of the universe to be the inverse of the present value of the Hubble constant. The integral then provides a correction factor that depends on the specific energy content of the universe (AST2401).

Figure 12 shows the value for the correction factor in equation (94) numerically integrated for different cosmologies. An accelerating universe is one with a positive value for  $\Omega_\Lambda$ . From the plot we can see that an accelerating universe is older for a fixed Hubble constant and fixed values for the other energy densities. To compare between an open and flat universe in the non-accelerating case we look at the line labeled “no CC” (no cosmological constant) in the plot. From this, we see that an open universe is older than a flat universe. The rankings of the three cosmologies in descending order of age is therefore an accelerating universe, an open universe, and a flat universe. However, I am not quite sure if this is exactly what the question is asking. Obviously, degeneracies exist between the age, curvature, and specific values of the density parameters, as evidenced by the contours in Figure 12. If, for example, they are asking this question in the context of Type Ia SNe results, then a different answer will be obtained; this will be discussed below when we talk about the eventual fate of the universe.

We can get a physical idea for this by considering the deceleration parameter  $q \equiv -\ddot{a}\dot{a}/\dot{a}^2$  which evaluates to 1, 0.5, -1, 0 for a universe dominated by  $\Omega_{\text{rad}}, \Omega_m, \Omega_\Lambda, \Omega_k$  respectively. This parameter describes the deceleration of the universal expansion where a positive sign indicates a decelerating universe (one for which  $\dot{a} < 0$ ). We see that the gravitational influences of matter and radiation cause the universe to decelerate whereas a cosmological constant causes it to increase and an empty universe is characterized by no acceleration. If we imagine rewinding the universe from its present day state then we will have that the galaxies in the universe accelerate (decelerate) towards each other in a matter-dominated ( $\Lambda$ -dominated) universe and so we reach the Big Bang quicker (slower). We can observe this by taking the different  $\Omega$ 's in equation (94) to be unity with the rest zero to see how the age of the universe depends on the different components. We find that the integral evaluates to 0.5, 2/3,  $\infty$ , 1 for  $\Omega_{\text{rad}}, \Omega_m, \Omega_\Lambda, \Omega_k$  respectively.

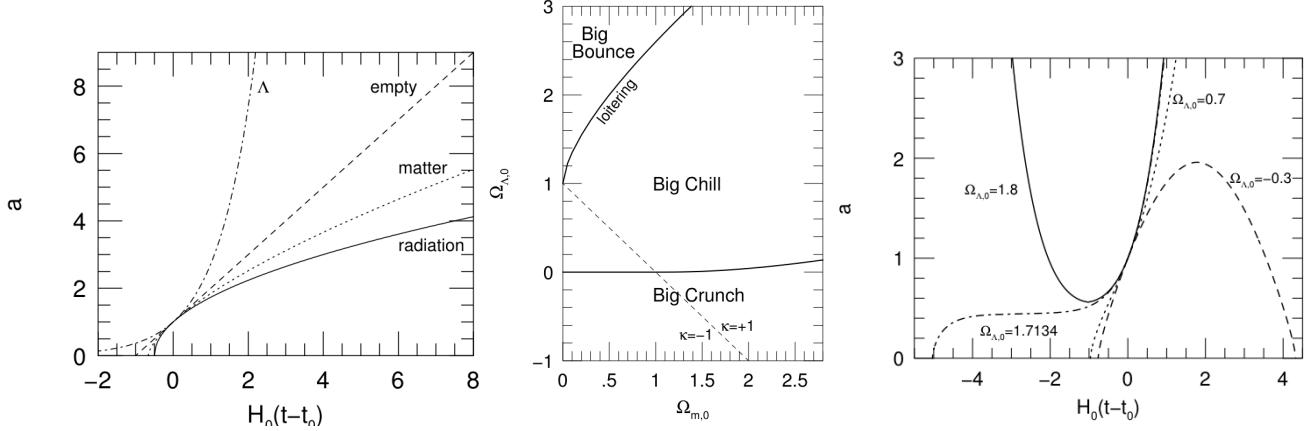


FIG. 13.—(left) The scale factor as a function of time for an expanding, empty universe (dashed), a flat, matter-dominated universe (dotted), a flat, radiation-dominated universe (solid), and a flat,  $\Lambda$ -dominated universe (dot-dash). (middle) The curvature and type of expansion for universes containing both matter and a cosmological constant. The dashed line indicates  $\kappa = 0$ ; models lying above this line are closed, and those lying below are open. (right) The scale factor  $a$  as a function of  $t$  in four different universes, each with  $\Omega_{\text{mat}} = 0.3$ . The dashed line shows a “Big Crunch” universe ( $\Omega_\Lambda = -0.3, \kappa = -1$ ). The dotted line shows a “Big Chill” universe ( $\Omega_\Lambda = 0.7, \kappa = 0$ ). The dot-dash line shows a loitering universe ( $\Omega_\Lambda = 1.7134, \kappa = +1$ ). The solid line shows a “Big Bounce” universe ( $\Omega_\Lambda = 1.8, \kappa = +1$ ). Images taken from Ryden (2002).

#### EXAMPLE UNIVERSES

In a spatially homogeneous and isotropic universe, the relation among the energy density  $\rho(t)$ , the pressure  $P(t)$ , and the scale factor  $a(t)$  is given by the Friedmann equation,

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2} \rho(t) - \frac{\kappa c^2}{R_0^2 a^2}, \quad (95)$$

the fluid equation,

$$\dot{\rho} + 3\frac{\dot{a}}{a}(\rho + P) = 0, \quad (96)$$

and the equation of state,

$$P = w\rho. \quad (97)$$

In principle, given the appropriate boundary conditions, we can solve these three equations for all times, past and future (Ryden 2002, pg. 79). For our purposes here it is useful to first think of single-component models of the universe in order to gain an intuition into how separate energy forms influence the expansion history of the universe.

First, let’s discuss an empty, curvature-only universe, that is one without matter, radiation, or other forms of mass-energy. For this universe, the Friedmann equation reduces to

$$\dot{a}^2 = -\frac{\kappa c^2}{R_0^2}. \quad (98)$$

One solution to this equation is  $\kappa = \dot{a} = 0$ . Evidently, an empty, *static*, spatially flat universe is a permissible solution to the Friedmann equation (i.e. Minkowski spacetime). Equation (98) also tells us that it is possible to have an empty, open universe with  $\kappa = -1$ . However, a positively curved (closed) empty universe is forbidden, since that would require an imaginary value of  $\dot{a}$ . For the negatively curved universe, we have that the universe must either be expanding or contracting, with  $\dot{a} = \pm c/R_0$ , which upon integration, suggests an evolution of the form  $a \propto t$ . In Newtonian terms, if there’s no gravitational force at work, then the relative velocity of any two points is constant, and thus the scale factor simply increases linearly with time in an empty universe (Ryden 2002, pg. 87).

Let’s now take the opposite approach and consider a universe for which there is no curvature, that is, a flat universe with  $\kappa = 0$ . In this case, the Friedmann equation for a single-component universe takes the simple form

$$\dot{a}^2 = \frac{8\pi G \rho_{\text{crit}}}{3c^2} a^{-(1+3w)}. \quad (99)$$

Using the guess that  $a(t) = (t/t_0)^q$ , we can solve equation (99), and then rearrange to find the current age of the universe:

$$t_0 = \frac{2}{3(1+w)} H_0^{-1}. \quad (100)$$

Hence, in a spatially flat universe, if  $w > -1/3$ , the universe is *younger* than the Hubble time. If  $w < -1/3$ , the universe is *older* than the Hubble time. In addition, we see that a flat matter-only universe will be *older* than a flat radiation-dominated universe. A caveat is that other solutions to equation (99). In particular, we know that in a  $\Lambda$ -dominated universe we will find exponential expansion of the form  $a(t) = e^{H_0(t-t_0)}$ . Thus, a spatially flat universe with nothing but a cosmological constant is exponentially

expanding, and has no beginning. We have seen this concept before when discussing the Steady State universe (Ryden 2002, pg. 98). Figure 13 shows the expansion history for the models we have discussed so far.

#### FATE OF THE UNIVERSE

To describe the eventual evolution of the universe we will think generally about one containing matter, curvature, and  $\Lambda$  (we will ignore radiation since it is unimportant at the current epoch). In this case, a wide range of behaviours is possible for the past and future evolution of  $a(t)$ . Start by writing down the Friedmann equation for a curved universe with both matter and a cosmological constant:

$$\frac{H^2}{H_0^2} = \frac{\Omega_{\text{mat}}}{a^3} + \frac{1 - \Omega_0 - \Omega_\Lambda}{a^2} + \Omega_\Lambda. \quad (101)$$

If  $\Omega_{\text{mat}} > 0$  and  $\Omega_\Lambda > 0$ , then both the first and last term on the right hand side of equation (101) are positive. However, if  $\Omega_{\text{mat}} + \Omega_\Lambda > 1$ , so that the universe is positively curved, then the central term on the right hand side is negative. As a result, for some choices of  $\Omega_{\text{mat}}$  and  $\Omega_\Lambda$ , the value of  $H^2$  will be positive for small values of  $a$  (where matter dominates) and for large values of  $a$  (where  $\Lambda$  dominates), but will be negative for intermediate values of  $a$  (where the curvature term dominates). Since negative values of  $H^2$  are unphysical, this means that these universes have a forbidden range of scale factors. Suppose such a universe starts out with  $a \gg 1$  and  $H < 0$ ; that is, it is contracting from a low-density,  $\Lambda$ -dominated state. As the universe contracts, however, the negative curvature term in equation (101) becomes dominant, causing the contraction to stop at a minimum scale factor  $a = a_{\min}$ , and then expand outward again in a “Big Bounce”. Thus, it is possible to have a universe which expands outward at late times, but which never had an initial Big Bang, with  $a = 0$  at  $t = 0$ . Another possibility, if the values of  $\Omega_{\text{mat}}$  and  $\Omega_\Lambda$  are chosen just right, is a “loitering” universe. Such a universe starts in a matter-dominated state, expanding outward with  $a \propto t^{2/3}$ . Then, however, it enters a stage (called the loitering stage) in which  $a$  is very nearly constant for a long period of time (Ryden 2002, pg. 113).

Figure 13 shows the general behaviour of  $a(t)$  as a function of  $\Omega_{\text{mat}}$  and  $\Omega_\Lambda$ . In the region labeled “Big Crunch”, the universe starts with  $a = 0$  at  $t = 0$ , reaches a maximum scale factor  $a_{\max}$ , then recollapses to  $a = 0$  at a finite time  $t = t_{\text{crunch}}$ . Note that Big Crunch universes can be positively curved, negatively curved, or flat. In the region labeled “Big Chill”, the universe starts with  $a = 0$  at  $t = 0$ , then expands outward forever, with  $a \rightarrow \infty$  as  $t \rightarrow \infty$ . Like Big Crunch universes, Big Chill universes can have any sign for their curvature. In the region labeled “Big Bounce”, the universe starts in a contracting state, reaches a minimum scale factor  $a = a_{\min} > 0$  at some time  $t_{\text{bounce}}$ , then expands outward forever, with  $a \rightarrow \infty$  as  $t \rightarrow \infty$ . Universes which fall just below the dividing line between Big Bounce universes and Big Chill universes are loitering universes. The closer such a universe lies to the Big Bounce – Big Chill dividing line in Figure 13, the longer its loitering stage lasts (Ryden 2002, pg. 113).

To illustrate the different types of expansion and contraction possible, Figure 13 shows  $a(t)$  for a set of four model universes. Each of these universes has the same current density parameter for matter:  $\Omega_{\text{mat}}$ , measured at  $t = t_0$  and  $a = 1$ . These universes cannot be distinguished from each other by measuring their current matter density and Hubble constant. Nevertheless, thanks to their different values for the cosmological constant, they have very different pasts and very different futures (Ryden 2002, pg. 116). In this case, we can see that at fixed  $H_0$  (they all have the same current slopes) and fixed  $\Omega_{\text{mat}}$ , varying  $\Omega_\Lambda$  shows that in order of descending age we have a closed, flat, and open universe. In this case both the closed and flat universes are “accelerating” in the sense that  $\Omega_\Lambda > \Omega_{\text{mat}} > 0$ . (**THIS MAKES ME CONFUSED ABOUT WHAT EXACTLY THIS QUESTION IS ASKING**).

There is strong observational evidence that we do not live in a loitering or Big Bounce universe. If we lived in a loitering universe, then as we looked out into space, we would see nearly the same redshift  $z_{\text{loiter}} = 1/a_{\text{loiter}} - 1$  for galaxies with a very large range of distances. No such excess of galaxies is seen at any redshift in our universe. If we lived in a Big Bounce universe, then the largest redshift we would see for any galaxy would be  $z_{\max} = 1/a_{\text{bounce}} - 1$ . As we looked further into space, we would see redshifts increase to  $z_{\max}$ , then see the redshifts decrease until they actually became blueshifts. In our universe, we do not see such distant blueshifted galaxies. Our own universe seems to be a Big Chill universe, fated to eternal expansion (Ryden 2002, pg. 116).

#### OBSERVATIONAL AGE MEASUREMENTS

Lower limits on the age of the Universe can be placed by dating the ages of the globular clusters in the MW. Age determinations of globular clusters can be calculated by observing the mass scale of the MS turnoff of stars in the cluster. Another method involves determining the temperatures of white dwarfs and using the fact that they cool as they age to place an estimate on their age. Both of these methods have been employed on the globular cluster NGC 6397 by Gratton et al. (2003) and Hansen et al. (2007) dating it as  $13.4 \pm 0.8$  and  $11.47 \pm 0.47$  Gyr using the former and latter methods respectively.

This can be taken a step further to place constraints on the age of the MW and hence better constraints on the age of the Universe. Pasquini et al. (2004) describe a process by which tracing the abundance of beryllium in the stars of a given globular cluster can constrain the amount of time that passed between the formation of the first stars in the Milky Way and the formation of the cluster. Beryllium is produced in the ISM through the spallation of heavy elements (namely C, N, and O) by interactions with galactic cosmic rays (GCRs). Moreover, beryllium (namely,  $^9\text{Be}$ ) was not produced during the Big Bang and is not synthesized in stars. Since GCRs travel throughout the Milky Way, guided by the cosmic magnetic field, the beryllium abundance throughout the galaxy should be spatially uniform. Moreover, its abundance increases with time allowing it to be used as a type of “cosmic clock”<sup>4</sup>. By measuring the beryllium abundance in turnoff stars, Pasquini et al. (2004) find that the globular cluster NGC 6397 formed some 0.2 Gyr after the first stars in the MW, which using the results of Gratton et al. (2003), implies an age of the Milky

<sup>4</sup> For an interesting article see the ESO press release at <http://www.eso.org/public/news/eso0425/>

Way of  $13.6 \pm 0.8$  Gyr. Turnoff stars are used since they represent the less massive, less evolved stars in the globular cluster. This is important because  $^9\text{Be}$  is destroyed at temperatures above  $\sim 10^6$  K, so convective motions in the red giant phase will mix the initial beryllium content in the upper atmosphere with hot gas beneath, potentially suppressing its signal.

**QUESTION 10**

**What are the currently accepted relative fractions of the various components of the matter-energy density of the universe?**

### QUESTION 10

**What are the currently accepted relative fractions of the various components of the matter-energy density of the universe?**

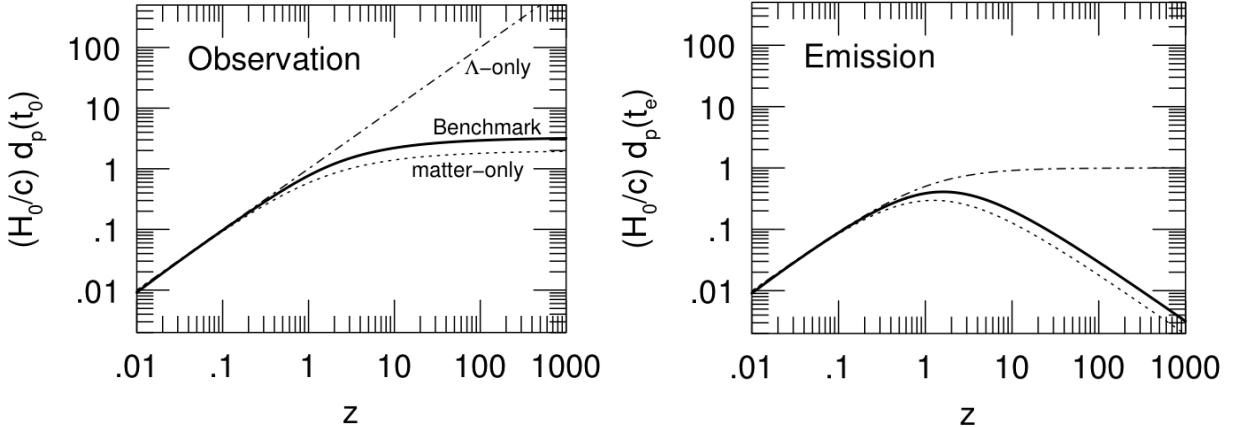


FIG. 14.— Proper distance to a light source with redshift  $z$  at the time of observation (left panel) and its distance at the time of emission (right panel). The bold solid line indicates the Benchmark Model, the dot-dash line a flat,  $\Lambda$ -only universe, and the dotted line a flat, matter-only universe. Image taken from Ryden (2002).

The seven-year WMAP data is well fit by a minimal six-parameter flat  $\Lambda$ CDM model. The parameters for this model, using the WMAP data in conjunction with baryon acoustic oscillation data from the SDSS and priors on  $H_0$  from the HST observations are:

$$\begin{aligned}\Omega_{\text{bary}} &= 0.046 \\ \Omega_{\text{CDM}} &= 0.227 \\ \Omega_{\text{mat}} &= 0.273 \\ \Omega_{\Lambda} &= 0.728 \\ \Omega_{\text{rad}} &= 8.5 \times 10^{-5} \\ \Omega_{\text{tot}} &= 1.002\end{aligned}\tag{102}$$

Most of these values are taken from Jarosik et al. (2011) while the radiation density can be inferred from the CMB temperature (as discussed below). Its value can also be obtained by taking the WMAP values for  $\Omega_{\text{mat}}$  and  $z_{\text{eq}}$ , and evolving them backwards in time to determine what the radiation density must have been. This method, however, is not as useful as the first since it acts to magnify the errors associated with the WMAP measurements.

#### PROPERTIES OF OUR UNIVERSE

We have seen that in the  $\Lambda$ CDM framework the Friedmann equation can be manipulated to read

$$H^2 = H_0^2 \left( \frac{\Omega_{\text{rad}}}{a^4} + \frac{\Omega_{\text{mat}}}{a^3} + \Omega_{\Lambda} + \frac{1 - \Omega_{\text{tot}}}{a^2} \right). \tag{103}$$

The values above suggest that our universe is *flat* with  $\Omega_{\text{tot}} \simeq 1$ ; for good measure, however, we will include the last term on the right-hand side of equation (103). Since  $H = \dot{a}/a$ , multiplying both sides of equation (103) by  $a^2$ , and taking the square root yields the relation

$$H_0 t = \int_0^a \frac{da}{\sqrt{\Omega_{\text{rad}}/a^2 + \Omega_{\text{mat}}/a + \Omega_{\Lambda} a^2 + (1 - \Omega_{\text{tot}})}}. \tag{104}$$

This integral relates the cosmic time  $t$  with scale factor  $a$ , and in general, must be done numerically for a given set of density parameters. However, at particular epochs in time equation (104) can be simplified using physical reasoning. For instance, we know at early times radiation dominated the energy density of the universe, and so computing only its contribution gives

$$H_0 t \approx \int_0^a \frac{ada}{\sqrt{\Omega_{\text{rad}}}} \approx \frac{1}{2\sqrt{\Omega_{\text{rad}}}} a^2 \Rightarrow a(t) \approx \left( 2\sqrt{\Omega_{\text{rad}}} H_0 t \right)^{1/2}. \tag{105}$$

This expression is roughly valid up to matter-radiation equality occurring at  $a_{\text{eq}} = 3 \times 10^{-4}$ . On the other hand, if the universe continues to expand forever, then in the limit  $a \rightarrow 0$ , the cosmological constant term will dominate and a similar procedure shows that  $a(t) \approx \exp(\sqrt{\Omega_{\Lambda}} H_0 t)$ ; that is, it expands exponentially in time. Within the intervening time, matter is assumed to dominate

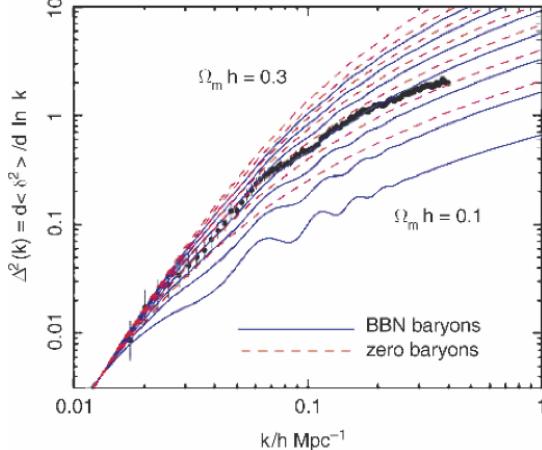


FIG. 15.— (left) Power spectrum of the galaxy distribution as measured in the 2dFGRS (points with error bars), here represented as  $\Delta^2(k) \propto k^3 P(k)$ . The curves show power spectra from CDM models with different shape parameter, and two values of  $\Omega_{\text{bary}}$ : one as obtained from primordial nucleosynthesis (BBN, solid curves), and the other for models without baryons (dashed curves). Image taken from Schneider (2002).

and we find that  $a(t) \propto t^{2/3}$ . In general, these approximations can be used to describe the three separate phases of the universe, though near the points of equality it is more accurate to assume two-component models (Ryden 2002, pg. 102).

Once  $a(t)$  has been computed for a given cosmology, a variety of other quantities can be computed. For instance, Figure 15 shows the current proper distance to a galaxy with redshift  $z$ . The heavy solid line is the result for the **Benchmark Model** (i.e. our universe); for purposes of comparison, the result for a flat  $\Lambda$ -only universe is shown as a dot-dash line and the result for a flat matter-only universe is shown as the dotted line. In the limit  $z \rightarrow \infty$ , the proper distance  $d_p(t_0)$  approaches a limiting value  $d_p \rightarrow 3.24c/H_0$ , in the case of the Benchmark Model. Thus, the Benchmark Model has a finite horizon distance,

$$d_{\text{hor}}(t_0) = \frac{3.24c}{H_0} = 14 \text{ Gpc} = 46 \text{ Gly.} \quad (106)$$

If the Benchmark Model is a good description of our own universe, then we can't see objects more than 14 Gpc away because light from them has not yet had time to reach us. Figure 15 also shows  $d_p(t_e)$ , the distance to a galaxy with observed redshift  $z$  at the time the observed photons were emitted. For the Benchmark Model,  $d_p(t_e)$  has a maximum for galaxies with redshift  $z = 1.6$ , where  $d_p(t_e) = 1.8 \text{ Gpc} = 5.9 \text{ Gyr}$  (Ryden 2002, pg. 121).

As the universe expands and ages, photons from increasingly distant objects have more time to complete their journey to Earth. This means that as time increases, we might expect that more of the universe will come into causal contact with us (i.e. the horizon distance above will increase in time). When we compute the horizon distance for the three phases of the universe we find that  $d_h = 2ct$  and  $3ct$  in the radiation and matter era, respectively. Note that  $d_h \propto t$  in these regimes while  $a$  is proportional to lower powers of  $t$ . Hence, during this time the size of the observable universe increased more rapidly than the universe expanded, so the universe becomes increasingly causally connected as it ages. However, when we do a similar calculation for the  $\Lambda$  era we find that  $d_h$  converges to basically the value above, with  $d_h \sim 14 \text{ Gyr}$ . This occurs because the exponential expansion is occurring too fast. Ultimately, an object located at the particle horizon will remain at the particle horizon as the universe expands, so its light will never reach us. Although photons from the object will continue to arrive, they will be deeply redshifted and their arrival rates will decline to zero due to cosmological time dilation (Carroll & Ostlie 2007, pg. 1205).

#### MEASURING COSMOLOGICAL PARAMETERS

Measuring  $\Omega_{\text{rad}}$  can be achieved through measurements of the CMB temperature. In particular, the number density of photons is given by the relation

$$\rho_\gamma = 2 \int \frac{d^3 p}{(2\pi)^3} \frac{1}{e^{p/T} - 1} p = \frac{\pi^2}{15} T^4, \quad (107)$$

where the factor of 2 accounts for the two spin states of photons, and the Bose-Einstein distribution is used to describe the number of photons in a given region of momentum space. Note that since we know that  $\rho_\gamma \propto a^{-4}$ , this relation implies  $T \propto a^{-1}$ . Dividing equation (107) by the critical density allows us to compute  $\Omega_{\text{rad}}$  based on knowledge of the present temperature  $T_0$  (Dodelson 2003, pg. 40)<sup>5</sup>.

The total matter density  $\Omega_{\text{mat}}$  can be measured through calculations of the power spectrum  $P_g(k)$  obtained through galaxy surveys. As we have discussed previously, the shape of  $P(k)$  is related to the shape parameter  $\Gamma = h\Omega_{\text{mat}}$ . Moreover, the shape of the BAO are useful in constraining how much of  $\Omega_{\text{mat}}$  is contributed by  $\Omega_{\text{bary}}$  (Schneider 2002, pg. 314). Constraints on  $\Omega_{\text{mat}}$  can also be obtained by counting the number of massive galaxy clusters we observe to the expected number of large halos (i.e.

<sup>5</sup> Note that we have been ignorant of the other forms of radiation existing in the universe (e.g., the starlight we see every night!). It turns out that by comparing the integrated luminosity for all galaxies, their total contribution to the radiation energy budget is about 10% that of the CMB. In addition, neutrinos, which are currently relativistic with energies  $E \gtrsim m_\nu c^2$ , must be included in the calculation of  $\Omega_{\text{rad}}$ . Calculations pertaining to the decoupling of neutrinos show that they should contribute about 66% to the energy budget of radiation (Ryden 2002, pg. 83).

Press-Schechter) as the latter depends on the total matter density. Another thing we can do is independently measure the mass  $M$  and luminosity  $L$  of galaxy clusters, and assuming that this represents the average  $M/L$  ratio of the universe, use this to determine  $\Omega_{\text{mat}}$ . This can then be decoupled to determine the value of  $\Omega_{\text{bary}}$  by measuring how much mass they contribute to the cluster (i.e. through X-ray emission), and assuming that this is representative of the universe (Schneider 2002, pg. 323).

In other questions, we talk about how quantities like  $\Omega_{\text{mat}}$  and  $\Omega_{\Lambda}$  can be computed from Type Ia SNe surveys. In addition, the CMB provides all kinds of methods for determining the different density components.

**QUESTION 11**

**Outline the history of the Universe. Include the following events: reionization, baryogenesis, formation of the Solar system, nucleosynthesis, star formation, galaxy formation, and recombination.**

## QUESTION 11

**Outline the history of the Universe. Include the following events: reionization, baryogenesis, formation of the Solar system, nucleosynthesis, star formation, galaxy formation, and recombination.**

Below we provide a brief chronological description of the evolution of the universe.

- **Planck Epoch** [ $0 - 10^{-44}$  s]: The Planck time is the time it takes light to travel one Planck length, and gives the length of time over which quantum gravity effects are significant (Charles). Cosmological models based on the Friedmann equation cannot be extrapolated back to this epoch since it marks the breakdown of general relativity. In cosmological contexts, general relativity assumes that the energy content of the universe is smooth down to arbitrarily small scales, instead of being parcelled into individual quanta. As long as a radiation-dominated universe has many, many quanta, or photons, within a horizon distance, then the approximation of a smooth, continuous energy density is justifiable, and we may safely use the results of general relativity. However, if there are only a few photons within the visible universe, then quantum mechanical effects *must* be taken into account, and the classical results of general relativity no longer apply. In order to accurately described the universe at its very earliest stages, prior to the Planck time, a theory of quantum gravity is needed (Ryden 2002, pg. 96).
- **Grand Unification Epoch** [ $10^{-44} - 10^{-36}$  s]: It is an article of faith for physicists that before the Planck time, the four fundamental forces of nature were merged into one all-encompassing “Theory of Everything” (TOE). When the universe reached the Planck time, the TOE force spontaneously separated into the gravitational force and a grand unified theory of the three remaining forces (Carroll & Ostlie 2007, pg. 1235).
- **Inflation** [ $10^{-36} - 10^{-34}$  s]: At the end of the GUTs epoch, the universe entered an extremely peculiar state called the false vacuum. This was not the true vacuum meaning the universe was not in the state with the lowest possible energy density. Inflation began when quantum fluctuations allowed a small region of space to enter a true vacuum state in a universe otherwise filled with false vacuum. The greater pressure inside the bubble caused the bubble to grow at an astounding rate. Thus, during this epoch, the constant energy density of the false vacuum became dominant in the acceleration equation and the universe grew by a factor of roughly  $e^{100} \sim 10^{43}$ . Today’s observable universe began in a bubble of true vacuum (Carroll & Ostlie 2007, pg. 1241).
- **Electroweak Epoch** [ $10^{-34} - 10^{-11}$  s]: The episode of inflationary growth came to a halt at the end of the GUTs epoch, when the strong nuclear force became distinct from the electroweak force. The elevated energy density of the false vacuum was then released, like the release of latent heat that occurs in freezing. This energy reheated the universe to its pre-inflation value, and generated a burst of particle-antiparticle creation (Carroll & Ostlie 2007, pg. 1243). At the end of the electroweak epoch, the weak nuclear and electromagnetic forces decouple from each other (Charles).
- **Quark Epoch** [ $10^{-11} - 10^{-5}$  s]: During the quark epoch the universe was filled with a dense, hot quark-gluon<sup>6</sup> plasma, containing quarks, leptons (e.g., electrons and neutrinos) and their antiparticles. Collisions between particles were too energetic to allow quarks to combine into mesons or baryons (Wikipedia).
- **Quark-Hadron transition** [ $10^{-5}$  s]: At this time the universe has cooled ( $T \sim 200$  MeV) to the point that quarks binding into hadrons (i.e., baryons and mesons) is energetically feasible, so that at lower temperatures, the constituents of the universe are hadrons, leptons, and photons (Olive 1991). As the universe cooled further it became too cold to create hadron/anti-hadron pairs, and all remaining pairs annihilate, leaving a small hadron excess due to **baryogenesis**. Obviously at some point before this the universe satisfied the Sakharov conditions<sup>7</sup> for baryogenesis, resulting in the universe being dominated by matter over antimatter. This likely occurred much earlier in the electroweak epoch (Charles).
- **Neutrino Decoupling** [0.1 s]: The rate for weak interactions keeping neutrinos coupled to the rest of the cosmic plasma drop beneath the expansion rate. As a result, neutrinos decouple from the plasma (Dodelson 2003, pg. 62). This also brings neutrons, protons, and electrons out of equilibrium, and neutrons begin to quickly decay (Olive 1991).
- **Electron-positron annihilation** [1.3 s]: At this time the temperature of the universe cools to the point that electron-positron pairs can no longer be spontaneously generated and all remaining pairs annihilate. **Leptogenesis?**
- **Big Bang Nucleosynthesis** [3–17 min]: At temperatures  $T \gg 1$  MeV nucleosynthesis cannot proceed even though the rate for forming the first isotope, deuterium, is sufficiently rapid. To begin with, at  $T > 1$  MeV, deuterium is photodissociated because the photon energy is greater than its binding energy of 2.2 MeV ( $\bar{E} = 2.7T$  for a blackbody). Furthermore, the density of photons is very high,  $n_\gamma/n_{\text{bary}} \sim 10^{10}$ , and the high-energy tail of the photon distribution will continue to dissociate deuterium at smaller temperatures. It isn’t until about  $T \sim 0.1$  MeV that this deuterium bottleneck can be overcome (Olive 1991). After about 10 minutes the temperature of the universe drops below that required to sustain fusion (Charles).

<sup>6</sup> There six types of **quarks**: up, down, strange, charm, bottom, and top, along with their antiquarks. Particles made up of quarks are **hadrons**. There are two types of hadrons: **baryons** (made of three quarks) and **mesons** (formed by a quark-antiquark pair). **Gluons** are force-carrying particles that mediate the strong interaction and bind quarks together (Carroll & Ostlie 2007, pg. 1230).

<sup>7</sup> A small baryon asymmetry  $\eta = (n_b - n_b^*)/n_\gamma$  may have been produced in the early universe if three necessary conditions are satisfied: (a) baryon-number (B) violation; (b) violation of C (charge conjugation symmetry) and CP (the composition of parity and C); and (c) departure from thermal equilibrium. (Riotto & Trodden 1999).

- **Radiation-Matter Equality** [0.07 Myr]: Matter and photon energy density equalize at this point.
- **Recombination** [0.3 Myr]: At this time the universe cooled to the point that electrons could recombine with protons and neutrons. This marked a phase transition where the universe went from being predominantly neutral to mostly ionized.
- **Dark Ages** [0.3 – 500 Myr]: The universe is mostly dark except for the fading glow of the CMB and the 21-cm transition of neutral hydrogen.
- **Star and Galaxy Formation** [ $\sim$  500 Myr]: The first stars and dwarf galaxies form in the universe and their collective luminosity puts an end to the cosmological dark ages.
- **Reionization** [0.5 – 1 Gyr]: The collective radiation from the first stars and galaxies progressively ionizes the surrounding IGM. This marks a phase transition in the universe from being predominantly neutral to mostly ionized.
- **Solar System Formation** [9 Gyr]: Formation of the solar system from the collapse of a molecular cloud some 4.5 Gyr ago.

**QUESTION 12**

**Explain how measurements of the angular power spectrum of the cosmic microwave background are used in the determination of cosmological parameters.**

## QUESTION 12

**Explain how measurements of the angular power spectrum of the cosmic microwave background are used in the determination of cosmological parameters.**

The CMB consists of photons that last interacted with matter at  $z \sim 1000$ . Since the Universe must have already been inhomogeneous at this time, in order for the structures present in the Universe today to be able to form, it is expected that these spatial inhomogeneities are reflected in a (small) anisotropy of the CMB: the angular distribution of the CMB temperature reflects the matter inhomogeneities at the redshift of decoupling of radiation and matter. Temperature fluctuations originating at this time are called **primary anisotropies**. Later, as the CMB photons propagate through the Universe, they may experience a number of distortions along their way which, again, may change their temperature distribution on the sky. These effects then lead to **secondary anisotropies** (Schneider 2002, pg. 336).

The most basic mechanisms causing primary anisotropies are the following:

- **Sachs-Wolfe Effect:** Inhomogeneities in the gravitational potential cause photons which originate in regions of higher density to climb out of a potential well, and thus experience a gravitational redshift. This effect is partly compensated for by the fact that, besides the gravitational redshift, a gravitational time delay also occurs: a photon that originates in an overdense region will be scattered at a slightly earlier time, and thus at a slightly higher temperature of the Universe, compared to a photon from a region of average density.
- **Doppler Shifts:** Photons that Thomson scatter off electrons for the last time do not follow the pure Hubble flow, but will have an additional peculiar velocity.
- **BAO:** Acoustic oscillations induced by the strong coupling between baryons and radiation on scales smaller than the horizon scale at recombination.
- **Silk Damping:** The coupling of baryons and photons is not perfect since, owing to the finite mean free path of photons, the two components are decoupled on small spatial scales. This implies that on small length-scales, the temperature fluctuations can be smeared out by the diffusion of photons.

Secondary anisotropies include things like the thermal and kinetic Sunyaev-Zel'dovich effect (described in another question) and the **integrated Sachs-Wolfe effect** (ISW). The latter has to do with the continued redshifting or blueshifting of CMB photons due to ongoing structure formation within the universe that is constantly changing the gravitational potential energy landscape (Schneider 2002, pg. 337).

Consider the temperature fluctuations  $\delta T/T$  observed in the CMB. Since these fluctuations are defined on the surface of the celestial sphere, it is useful to expand in spherical harmonics (basically the 2D analog of a Fourier transform):

$$\frac{\delta T}{T}(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^l a_{lm} Y_{lm}(\theta, \phi). \quad (108)$$

However, what concerns us is not the exact pattern on the sky, but rather its statistical properties; the most important statistical property is the **correlation function**  $C(\theta)$ . Consider two points on the last scattering surface. Relative to an observer, they are in the directions  $\hat{n}$  and  $\hat{n}'$ , and are separated by an angle  $\theta$  given by the relation  $\cos \theta = \hat{n} \cdot \hat{n}'$ . To find the correlation function  $C(\theta)$ , multiply together the values of  $\delta T/T$  at the two points, then average the product over all points separated by the angle  $\theta$ :

$$C(\theta) = \left\langle \frac{\delta T}{T}(\hat{n}) \frac{\delta T}{T}(\hat{n}') \right\rangle. \quad (109)$$

Using the expansion of  $\delta T/T$  in equation (108), we can reduce equation (109) to

$$C(\theta) = \frac{1}{4\pi} \sum_{l=0}^{\infty} (2l+1) C_l P_l(\cos \theta), \quad (110)$$

where  $P_l$  are the Legendre polynomials. In this way, a measured correlation function  $C(\theta)$  can be broken down into its multipole moments  $C_l$ . Generally speaking, a term  $C_l$  is a measure of temperature fluctuations on the angular scale

$$\theta \sim \frac{\pi}{l} \sim \frac{180^\circ}{l}. \quad (111)$$

Thus, the multipole  $l$  is interchangeable, for all practical purposes, with the angular scale  $\theta$ . The  $l = 0$  (monopole) term of the correlation function vanishes if you've defined the mean temperature correctly. The  $l = 1$  (dipole) term results primarily from the Doppler shift due to our motion through space. Thus, the moments  $l \geq 2$  are those for which we are interested in. Finally, we note that in presenting CMB results it is customary to plot the function

$$\Delta_T \equiv \left( \frac{l(l+1)}{2\pi} C_l \right)^{1/2} \langle T \rangle, \quad (112)$$

since this function tells us the contribution per logarithmic interval in  $l$  to the total temperature fluctuation  $\delta T$  of the CMB (Ryden 2002, pg. 197).

### THE HORIZON SCALE

To explain the basic features of CMB fluctuations, we first point out that a characteristic length-scale exists at  $z_{\text{rec}}$ , namely the horizon length. We have previously seen that the (comoving) horizon distance can be calculated as

$$d_{\text{hor}} = \int_0^a \frac{c da'}{a'^2 H(a')} \approx 2 \frac{c}{H_0} \frac{1}{\sqrt{(1+z)\Omega_{\text{mat}}}}, \quad (113)$$

where the approximation is valid during matter domination (e.g., during recombination). Transforming this to proper units,  $d_{\text{hor,prop}}$ , we see that the angular size on the sky that this length corresponds to is

$$\theta_H = \frac{d_{\text{hor,prop}}(z_{\text{rec}})}{d_A(z_{\text{rec}})}, \quad (114)$$

where  $d_A$  is the angular-diameter distance to the surface of last scattering (Schneider 2002, pg. 171). We have seen before that the angle subtended by a standard yardstick depends on the specific cosmology through the functional form of  $d_A$ . A common yardstick to use is the first peak in the CMB, which occurs at the angular scale subtended by the sound horizon at the surface of last scatter. Since  $\rho_\gamma > \rho_{\text{bary}}$  at that epoch, the sound speed of the universe is close to  $c/\sqrt{3}$ , so that the sound horizon corresponds to a nearly fixed physical scale (Jungman et al. 1996). In particular, the angular scale it subtends is  $\theta_1 \approx \theta_H/\sqrt{3}$ . Although degeneracy exists within  $d_A$ , we note that increasing the distance to the surface of last scattering also decreases the angular extent of the acoustic features in the CMB. This distance depends mainly on the expansion rate and hence the matter content of the universe. Because the same factor contributes to both the numerator and denominator of equation (114), this dependency nearly cancels out leaving the much stronger dependence on curvature (Hu et al. 1997).

On scales  $\theta \gg \theta_H$ , the Sachs-Wolfe effect dominates, since oscillations in the baryon-photon fluid can occur only on scales below the horizon length. For this reason, the CMB angular spectrum directly reflects the fluctuation spectrum  $P(k)$  of matter. In particular, if we have  $P(k) \propto k$ , we expect to find

$$l(l+1)C_l \approx \text{const. for } l \ll \frac{180^\circ}{\theta_H} \simeq 100, \quad (115)$$

and the amplitude of fluctuations immediately yields the amplitude of  $P(k)$ . Note that this flat behaviour of the fluctuation spectrum for  $n_s = 1$  is modified by the ISW effect. On angular scales  $\theta < \theta_H$ , fluctuations are observed that were inside the horizon prior to recombination, hence physical effects may act on these scales. We have already mentioned that these fluctuations arise from acoustic oscillations within the photon-baryon fluid, and we will explore their importance in great detail in another question. On even larger scales, Silk damping will dominate. Since recombination is not instantaneous but extends over a finite range in redshift, CMB photons are last scattered within a shell of finite thickness. Considering a length-scale that is much smaller than the thickness of this shell, several maxima and minima of  $T$  are located within this shell along a line-of-sight. For this reason, the temperature fluctuations on these small scales are averaged out in the integration along the line-of-sight. The thickness of the recombination shell is roughly equal to the diffusion length of the photons (Schneider 2002, pg. 339).

### MODEL DEPENDENCE OF THE FLUCTUATIONS

Figure 16 shows the power spectra of CMB fluctuations where, starting from some reference model, individual cosmological parameters are varied. First we note that the spectrum is basically characterized by three distinct regions in  $l$  (or in the angular scale). For  $l \lesssim 100$ ,  $l(l+1)C_l$  is a relatively flat function if – as in the figure – a Harrison-Zel'dovich spectrum is assumed. In the range  $l \gtrsim 100$ , local maxima and minima can be seen that originate from the acoustic oscillations. For  $l \gtrsim 2000$ , the amplitude of the power spectrum is strongly decreasing due to Silk damping (Schneider 2002, pg. 340).

Figure 16(a) shows the dependence of the power spectrum on the curvature of the universe. We see that the curvature has two fundamental effects on the spectrum: first, the locations of the minima and maxima of the acoustic peaks are shifted, and second, the spectral shape at  $l \lesssim 100$  depends strongly on  $\Omega_{\text{tot}}$ . The latter is a consequence of the ISW effect because the more the world model is curved, the stronger the time variations of the gravitational potential  $\phi$ . The shift in the acoustic peaks is essentially a consequence of the change in the geometry of the Universe: the size of the sound horizon depends only weakly on the curvature, but the angular diameter distance  $d_A(z_{\text{rec}})$  is a very sensitive function of this curvature, so that the angular scale that corresponds to the sound horizon changes accordingly (Schneider 2002, pg. 340).

The dependence on the cosmological constant for flat models is displayed in Figure 16(b). Here we see that the effect of  $\Omega_\Lambda$  on the locations of the acoustic peaks is comparatively small, so that these basically depend on the curvature of the universe. The most important influence of  $\Omega_\Lambda$  is seen for small  $l$ . For  $\Omega_\Lambda = 0$ , the ISW effect vanishes and the power spectrum is flat (for  $n_s = 1$ ), whereas larger  $\Omega_\Lambda$  always produce a strong ISW effect (Schneider 2002, pg. 340). We explain the slight movement in the acoustic peaks when changing  $\Omega_\Lambda$  by noting that the distance back to the surface of last scatter is smaller in a  $\Lambda$ -dominated flat universe than in a matter-dominated flat universe, thus shifting the peaks to larger angular scales, or smaller  $l$ 's (Jungman et al. 1996).

The influence of the baryon density is presented in Figure 16(c). An increase in the baryon density causes the amplitude of the first acoustic peak to rise, whereas that of the second peak decreases. In general, the amplitudes of the odd-numbered acoustic peaks increase, and those of the even-numbered peaks decrease with increasing  $\Omega_{\text{bary}}$ . Furthermore, the damping of fluctuations sets in at smaller  $l$  (hence, larger angular scales) if  $\Omega_{\text{bary}}$  is reduced, since in this case the mean free path of photons increases, and so the fluctuations are smeared out over larger scales. Finally, Figure 16(d) demonstrates the dependence of the temperature

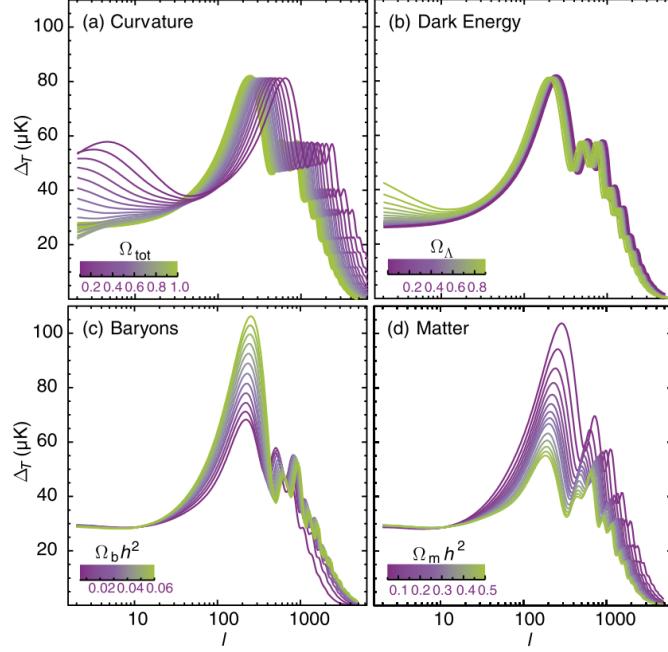


FIG. 16.— Dependence of the CMB fluctuation spectrum on cosmological parameters. Image taken from Schneider (2002).

fluctuations on the density parameter  $\Omega_{\text{mat}}$ . Changes in this parameter result in both a shift in the locations of the Doppler peaks and in changes of their amplitudes (Schneider 2002, pg. 341).

#### COSMIC VARIANCE

The angular fluctuation spectrum of CMB anisotropies is quantified by the multipole coefficients  $C_l$ . For instance,  $C_1$  describes the strength of the dipole. The dipole has three components; these can be described, for example, by an amplitude and two angles which specify a direction on the sphere. Accordingly, the quadrupole has five independent components, and in general,  $C_l$  is defined by  $(2l+1)$  independent components (i.e. higher  $l$  are sampled more by our observations). Cosmological models of the CMB anisotropies predict the *expectation* value of the amplitude of the individual components  $C_l$ . In order to compare measurements of the CMB with these models one needs to understand that we will never measure the expectation value, but instead we measure only the mean value of the components contributing to the  $C_l$  on our microwave sky. In general, the statistical deviation of the average of  $C_l$  from the expectation value is

$$\Delta C_l = \frac{C_l}{\sqrt{2l+1}}. \quad (116)$$

This equation represents a fundamental limit to the statistical accuracy of our measurements, which cannot be overcome by any improvements in instrumentation. This effect is called the **cosmic variance** and it arises because there is only one CMB sky that we can observe. The precision of the WMAP measurements is, for all  $l \lesssim 350$ , better than the cosmic variance (Schneider 2002, pg. 347).

#### POLARIZATION OF THE CMB

The cosmic background radiation is blackbody radiation and should therefore be unpolarized. Nevertheless, measurements have confirmed a finite polarization exists, which we will attempt to explain. The scattering of photons on free electrons not only changes the direction of the photons, but also produces a linear polarization of the scattered radiation. The direction of this polarization is perpendicular to the plane spanned by the incoming and the scattered photons. Consider now a region of space with free electrons. Photons from this direction have either propagated from the epoch of recombination to us without experiencing any scattering, or they have been scattered into our direction by the free electrons. Through this scattering, the radiation is, in principle, polarized. If the CMB, as seen from the scattering electrons, was isotropic, an equal number of photons would enter from all directions so that the net polarization would vanish. However, the scattering electrons see a slightly anisotropic CMB sky, in much the same way as we observe it; therefore, the net-polarization will not completely vanish. This picture implies that the CMB radiation may be polarized. The degree of polarization depends on the probability of a CMB photon having been scattered since recombination, thus on the optical depth with respect to Thomson scattering. Since the optical depth depends on the redshift at which the Universe was reionized, this redshift can be estimated from the degree of polarization (Schneider 2002, pg. 348).

**QUESTION 13**

**Explain how measurements of baryon-acoustic oscillations can be used in the determination of cosmological parameters.**

### QUESTION 13

Explain how measurements of baryon-acoustic oscillations can be used in the determination of cosmological parameters.

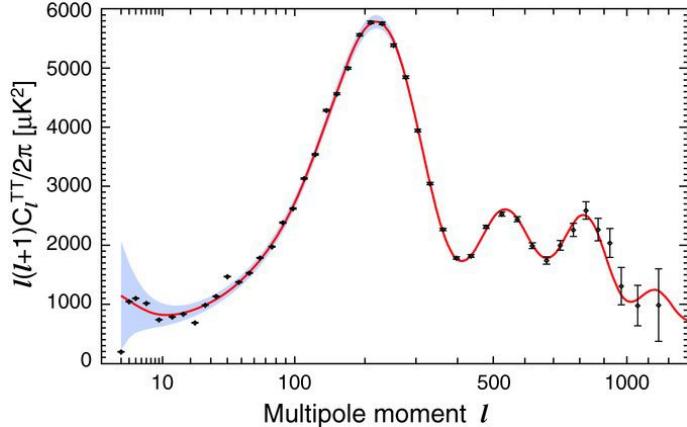


FIG. 17.— The CMB power spectrum taken from the seven-year WMAP data. The bottom axis represents multipole moment  $l$  which increases from left to right. This is related to the angular scale via  $\theta \sim \pi/l$  so that angular scales decrease from left to right. Although not shown here the limits of  $\theta$  go from  $90^\circ$  to  $0.2^\circ$  with the peak of the power spectrum occurring at  $\theta \sim 1^\circ$ .

When we observe the temperature fluctuations  $\delta T/T$  of the CMB it is customary to transform them into spherical harmonics and report the data in terms of the function  $C_l$  which describes the fluctuations in multipole moment  $l$ . In particular,  $C_l$  is a measure of the temperature fluctuations on the angular scale  $\theta \sim \pi/l$ , averaged over all points on the given CMB map that are separated by an angular scale  $\theta$ . For a given experiment we will then have that  $C_l$  is nonzero for all angular scales larger than the resolution of the experiment and for all angular scales smaller than the patch of the sky observed. The monopole term ( $l = 0$ ) is taken to be the mean temperature and the dipole term ( $l = 1$ ) arises from the Doppler shift of our peculiar velocity through space (Ryden 2002, pg. 197).

For plotting purposes the power spectrum of the CMB is usually represented as

$$\Delta_T^2 \equiv \frac{l(l+1)}{2\pi} C_l T^2, \quad (117)$$

which describes the power per logarithmic interval in multipole moment  $l$  (Hu & Dodelson 2002). Such a plot is displayed in Figure 17 which shows the seven-year WMAP power spectrum taken from Larson et al. (2011). We see that the power spectrum has a large peak at  $l_s \sim 200$  ( $\theta_s \sim 1^\circ$ ) which corresponds to the angular size of the sound horizon at the time of last scattering. The sound horizon is the distance through which perturbations in the photon-baryon fluid of the early universe could have traveled since the Big Bang. Since the sound speed in the photon-baryon fluid is close to the speed of light, the degree scale  $\theta_s$  also approximately marks the size of causally connected regions at the time of recombination (Hu & Dodelson 2002).

Angular scales  $\theta > \theta_s$  are therefore causally disconnected and so the temperature fluctuations in these regimes are not determined by the causal physics of interacting photons and baryons. Instead these fluctuations are determined by the gravitational effect of primordial density fluctuations in the distribution of dark matter. The reason for this is that recombination occurs during matter domination and dark matter is, of course, the greatest contributor to matter density. On the other hand, for angular scales  $\theta < \theta_s$  the temperature fluctuations result from the acoustic baryon oscillations that arise in the photon-baryon fluid at the time of recombination. Prior to recombination the photon-baryon fluid was a tightly coupled system with an energy density roughly a third that of dark matter. This fluid therefore moved primarily under the gravitational influences of dark matter with a sound speed  $c_s = \sqrt{w_{pb}c}$  where  $w_{pb}$  is the equation-of-state parameter, roughly intermediate the value of  $w = 1/3$  for photons and  $w = 0$  for baryons. As the fluid falls into the potential wells of dark matter it will be compressed until the radiation pressure of the photons begins to fight the gravitational pull of the dark matter. This eventually causes the fluid to expand until the radiation pressure again shrinks and the fluid falls in on itself again. Under this process the fluid oscillates between compressions and rarefactions which are known as the **baryonic acoustic oscillations (BAO)** (Ryden 2002, pg. 203).

If the photon-baryon fluid within a potential well is at maximum compression at the time of photon decoupling, its density will be higher than average, and the liberated photons, since  $T \propto \rho_\gamma^{1/4}$ , will be hotter than average. Conversely, if the photon-baryon fluid within a potential well is at maximum expansion at the time of decoupling, the liberated photons will be slightly cooler than average. Furthermore, if the photon-baryon fluid is in the process of expanding or contracting at the time of decoupling, the Doppler effect will cause the liberated photons to be cooler or hotter than average, depending on whether the photon-baryon fluid was moving away from our location or toward it at the time of photon decoupling (Ryden 2002, pg. 203).. The net result is a collection of hot and cold spots on CMB maps that can be transformed into a series of peaks and troughs in the angular power spectrum of CMB temperature fluctuations.

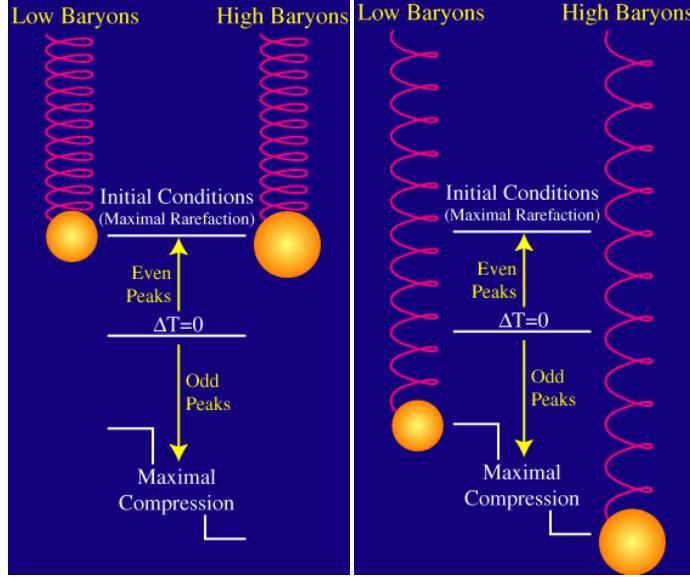


FIG. 18.— Baryons load down the photon-baryon plasma and add inertial (and gravitational) mass to the oscillating system. Their effect on the acoustic peaks is easy to understand. Remember what happens when you add mass to a spring and let it fall in the gravitational field of the Earth. With more mass loading the spring, it falls further before pulled back by the spring. On the other hand, it rebounds to the same position it started from. Since the odd numbered acoustic peaks are associated with how far the plasma “falls” into gravitational potential wells (how much the plasma compresses), they are enhanced by an increase in the amount of baryons in the universe. The even numbered peaks are associated with how far the plasma “rebounds” (how much the plasma rarefies). Images taken from <http://background.uchicago.edu/~whu/intermediate/baryons.html>

To get a feel for this we begin by examining an ideal photon-baryon fluid where we will neglect the effects of gravity and baryons. The following construction is taken from Hu & Dodelson (2002) unless cited otherwise. In this system the continuity and Euler equations for an ideal fluid reduce to

$$\ddot{\Theta} + c_s^2 k^2 \Theta = 0, \quad (118)$$

where  $\Theta$  is the photon temperature fluctuation  $\Theta = \Delta T / T$  and overdots are with respect to conformal time  $\eta$ . We write this as an equation in Fourier space since a general spatial fluctuation on the surface of last scattering (SLS) can be written as the superposition of normal modes  $k$  that evolve independently (Hu et al. 1997). We can solve this equation explicitly to determine the temperature distribution at recombination:

$$\Theta(\eta_*) = \Theta(0) \cos(k s_*), \quad (119)$$

where  $s$  is the sound horizon (the distance sound can travel by  $\eta$ ) and asterisks denote recombination at  $z_*$ . For some initial fluctuation  $\Theta(0)$  this represents the ensuing harmonic oscillations representing the continued heating and cooling of the fluid that arises from continued compression and rarefaction. For scales that are much larger than the horizon,  $ks \ll 1$ , a Taylor expansion of equation (119) shows that the perturbation is frozen into its initial condition (this is why small multipoles in Figure 17 tell us about the large-scale initial dark matter distribution). For small scales, equation (119) shows that the amplitude of the Fourier modes exhibit temporal oscillations. Modes that are caught at maxima or minima of their oscillation at recombination (when they stop oscillating) correspond to peaks in the power. Because sound takes half as long to travel half as far, modes corresponding to peaks follow a harmonic relationship  $k_n = n\pi/s_*$ , where  $n$  is an integer.

#### CURVATURE AND DARK ENERGY

We can interpret angular scales  $\theta$  in the power spectrum to represent physical inhomogeneities on a length scale  $\lambda$  via  $\theta \sim \lambda/d_A$ , where  $d_A(z)$  is the comoving angular diameter distance to the SLS (in a flat universe  $d_A^* = \eta_0 - \eta_* \approx \eta_0$ ). Equation (119) suggests a series of acoustic peaks located at  $l_n = nl_s$  where  $l_s = \pi d_A^*/s_*$  and  $n \geq 1$  is an integer. These peaks correspond to modes caught in maximal compression and rarefaction at the time of recombination, which results in the largest temperature fluctuations about the mean. In a flat matter-dominated universe  $\eta \propto \sqrt{1+z}$  so that  $\eta_*/\eta_0 \sim 2^\circ$ . From this simple picture the first thing we can glean from the power spectrum is the curvature of the universe. A given object at a fixed distance subtends a larger (smaller) angle in a closed (open) universe than in a flat universe. A closed (open) universe will therefore shift the peaks in the power spectrum to the left (right) relative to a flat universe. This effect is replicated by dark energy which has the effect of increasing the age of the universe for fixed conditions. Increasing (decreasing)  $\Omega_\Lambda$  causes the peaks in the power spectrum to shift to the left (right) since the age of the universe is increased (decreased) causing the sound horizon to also increase (decrease). Varying dark energy, however, has a less exaggerated effect in this case than varying the curvature parameter.

#### GRAVITATIONAL FORCING

Let's now consider the addition of gravity onto the oscillator. This leads to two additional contributions to the oscillator equation. Firstly, gravity causes the oscillations to become a competition between pressure gradients  $\Theta$  and gravity gradients  $\Psi$

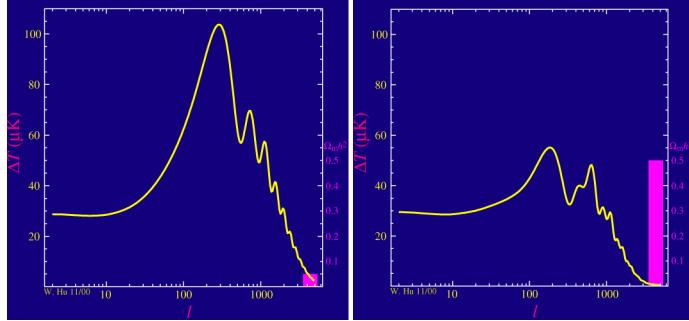


FIG. 19.— Effect of chaining the dark matter density  $\Omega_{\text{dm}}$  with a small (left panel) value and large (right panel value). As we raise the physical density of the dark matter the radiation driving effect goes away such that its amplitude decreases. Note that decreasing the matter density also affects the baryon loading since the dark matter potential wells go away leaving nothing for the baryons to fall into. Having a third peak that is boosted to a height comparable to or exceeding the second peak is an indication that dark matter dominated the matter density in the plasma before recombination. Notice also that the location of the peaks, and that of the first peak in particular, changes as we change the dark matter density. The matter to radiation ratio also controls the age of the universe at recombination and hence how far sound can travel relative to how far light travels after recombination. This is the leading order ambiguity in the measurement of the spatial curvature of the universe. We see here that that ambiguity will be resolved when at least three peaks are precisely measured. Images taken from <http://background.uchicago.edu/~whu/intermediate/driving2.html>

of the Newtonian potential  $\Phi$ , with an equilibrium point at  $\Theta = -\Psi$ . Secondly, gravity generates temperature perturbations by analogy to cosmological redshift,  $\delta\Theta = -\delta\Phi$ . Equation (118) is now modified as

$$\ddot{\Theta} + c_s^2 k^2 \Theta = -\frac{k^2}{3} \Psi - \ddot{\Phi}. \quad (120)$$

In a flat universe in the absence of pressure,  $\Phi$  and  $\Psi$  are constant, and with the absence of baryons  $c_s^2 = 1/3$ , so that equation (120) reduces to equation (118) with  $\Theta$  replaced by  $\Theta + \Psi$ . We therefore see that with the inclusion of gravity, the quantity that fluctuates is actually  $\Theta + \Psi$ , and these fluctuations arise even if there is no initial temperature fluctuation  $\Theta(0)$ ; the oscillations arise from the infall of the fluid into potential wells<sup>8</sup>. The quantity  $\Theta + \Psi$  is also the quantity we measure when we look at CMB maps since photons travelling out of an overdense (underdense) region at recombination are redshifted (blueshifted) by an amount  $\delta T/T \propto \Psi$ .

#### BARYON LOADING

Including baryons alters the oscillator equation to become

$$c_s^2 \frac{d}{d\eta} (c_s^{-2} \dot{\Theta}) + c_s^2 k^2 \Theta = -\frac{k^2}{3} \Psi - c_s^2 \frac{d}{d\eta} (c_s^{-2} \dot{\Phi}). \quad (121)$$

The addition of baryons decreases the sound speed by an amount  $c_s = 1/\sqrt{3(1+R)}$  where  $R$  is the baryon momentum density ratio  $R = (p_b + \rho_b)/(p_\gamma + \rho_\gamma)$ . To get a feel for the implications of baryons we take the limit where  $R$ ,  $\Theta$ , and  $\Psi$  are all constant. Then it can be shown that equation (121) can be written in the form of equation (118) with  $\Theta$  replaced by  $\Theta + (1+R)\Psi$ . This causes the equilibrium point to shift to  $\Theta = -(1+R)\Psi$  and since it is the quantity  $\Theta + \Psi$  that we measure for the temperature fluctuations, this breaks the symmetry of the oscillations. The baryons act to cause the fluid to fall further into the potential wells during compression, thereby increasing the amplitude of the odd numbered peaks in the power spectrum. We can think of this physically by analogy of a mass  $m = 1+R$  on a spring in a constant gravitational field. Increasing the mass on the spring causes the system to fall deeper into the potential well (increased compression) while continuing to rebound to the same position; see Figure 18.

#### RADIATION DRIVING

Up to now we have been considering the interactions of the photon-baryon fluid with the gravitational landscape of dark matter perturbations. However, high multipole modes that had oscillated multiple times up to recombination had actually started oscillating in the epoch of radiation domination. In this era the radiation density is creating the potential well in which the fluid is oscillating. As the fluid compresses into the well and the radiation pressure stabilizes the fluctuation, the gravitational potential begins to decay since the expansion of the universe dilutes the density of photons (i.e., the potential decays because the photons are what is actually causing this field in the first place). Consequently, when the fluid begins to rebound from the compression it no longer has as strong a gravitational potential to fight against and the amplitudes of the oscillations increase. The net effect of this is that across the horizon scale of matter-radiation equality,  $l_{\text{eq}}$ , the acoustic amplitudes increase by a factor of  $\sim 5$ . Determining this transition thus allows us to determine the ratio  $\Omega_{\text{mat}}/\Omega_{\text{rad}}$ , and since the latter is constrained by  $T_{\text{CMB}}$  itself, this transition allows us to determine the total matter density; see Figure 19. In addition, it should be noted that since radiation driving eliminates the gravitational potential it also eliminates the alternating peak heights that result from baryon loading. The observed high third peak in the power spectrum is a good indication that cold dark matter both exists and dominates the energy density at recombination.

<sup>8</sup> We could neglect this when considering large-scale perturbations, since at scales above the horizon length, the photon-baryon fluid, which travels at a speed  $< c$ , will not have had time to fall into the centre of these perturbations (Ryden 2002, pg. 203).

#### DIFFUSION DAMPING

Damping processes cut off the power spectrum of acoustic oscillations at small angular scales explaining the rapid decay at high multipoles seen in Figure 17. This is caused by the random walk of photons in the fluid which allows hot and cold regions to mix, thereby cancelling out fluctuations (Hu et al. 1997). The random walk is the result of Compton scattering off of free electrons and thus depends on the baryon density at the time of recombination. Higher free electron abundances translate into a smaller photon mean free path which causes damping to occur later, at higher multipoles. Hence, observing where this damping tail occurs allows a measurement of  $\Omega_{\text{bary}}$ . In general, however, the damping tail is used as a consistency check to the data since  $\Omega_{\text{bary}}$  can already be determined by measuring the positions and heights of the peak themselves. Knowing the abundance of free electrons, we can determine the physical mean free path of photons, and use this as a standard ruler to test the curvature of the universe.

#### BAO IN GALAXY CLUSTERING

As we will learn in another question, the BAO leave their signature imprinted on the clustering of low-redshift galaxies. The complementary probes of the CMB and galaxy clustering observations can be combined to break each others' parameter degeneracies in order to better constrain cosmological parameters. For instance, for a fixed primordial spectrum, increasing DM density shifts the galaxy power spectrum up to the right whilst shifting the CMB peaks down to the left. On the other hand, the addition of baryons boosts the amplitude of odd-numbered peaks in the CMB spectrum, but suppresses the galaxy spectrum rightward of its peak as well as making it wigglier. Finally, increasing the abundance of hot dark matter (i.e. neutrinos) suppresses galaxy clustering on small scales while having essentially no effect on the CMB (Tegmark 2002).

**QUESTION 14**

**Explain how weak lensing measurements can be used in the determination of cosmological parameters.**

#### QUESTION 14

**Explain how weak lensing measurements can be used in the determination of cosmological parameters.**

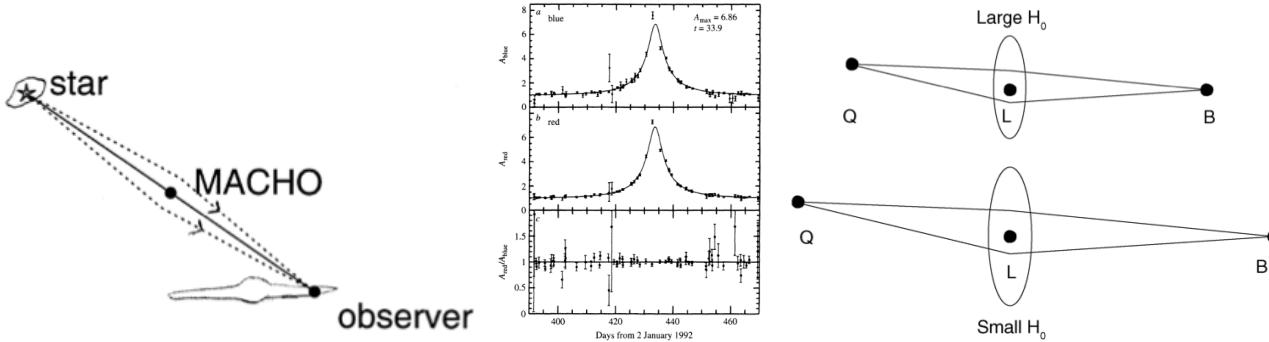


FIG. 20.— (left) Light from a star in the LMC is deflected by a MACHO on its way to an observer in the disk of the MW. Image taken from Ryden (2002). (middle) The light curve of a star in the LMC brightened over a period of 33 days, apparently because a MACHO passed through the line of sight. The data are shown for (a) blue light, (b) red light, and (c) the ratio of blue light to red light. Image taken from Carroll & Ostlie (2007). (right) Lens geometry in two universes with different Hubble constant. All observables are dimensionless – angular separations, flux ratios, redshifts – except for the difference in the light travel time. This is larger in the universe at the bottom than in the one at the top; hence,  $\Delta t \propto H_0^{-1}$ . If the time delay  $\Delta t$  can be measured, and if one has a good model for the mass distribution of the lens, then  $H_0$  can be derived. Image taken from Schneider (2002).

Many classical tracers of DM involve trying to detect the gravitational influence it has on luminous matter. For instance, we can detect DM around spiral galaxies because it affects the motions of stars and interstellar gas. We can also detect DM in clusters of galaxies because it affects the motions of galaxies and intracluster gas (Ryden 2002, pg. 170). Other methods involve tracing the luminous matter directly and assuming some sort of conversion factor to determine the DM abundance. In this case, assumptions must be made about things like mass-to-light ratios and ionized fractions (e.g., X-ray emission from galaxy clusters), and thus trace the DM distribution inasmuch as these assumptions are correct (Yevgeni).

#### BARYON-TO-MATTER FRACTION

After extra matter was identified through the rotation curves of spiral galaxies, it became a common curiosity to determine the nature of the DM halo. It cannot be in the form of interstellar dust, because dust betrays its presence through the extinction of starlight. Furthermore, the DM halo cannot be composed of gas, because absorption lines would be apparent when observing halo stars. One hypothesized scenario suggest that DM may be made up of **massive compact halo objects (MACHOs)**. MACHOs that could supply the unseen mass may be in the form of WDs, NSs, BHs, or less exotic brown dwarfs (Carroll & Ostlie 2007, pg. 897).

One way to detect the abundance of MACHO objets is to realize that matter not only gravitationally affects the trajectory of other matter, but also the trajectory of photons. In another question, we find that a photon passing a compact massive object at an impact parameter  $b$  is deflected through an angle

$$\alpha = \frac{4GM}{c^2 b}, \quad (122)$$

where  $M$  is the mass of the compact object. Equation (122) is valid as long as  $\alpha \ll 1$ , which is the case for weak gravitational fields (see justification for our use of this equation below) (Schneider 2002, pg. 64).

Suppose then that a MACHO in the halo of the MW passes between us and a star in the LMC, as demonstrated in Figure 20. As the MACHO deflects the light from the distant star, it produces an image of the star which is both distorted and amplified. If the MACHO is *exactly* along the line of sight between us and the LMC, the image produced is a perfect ring, with angular radius

$$\theta_E = \left( \frac{4GM}{c^2 d} \frac{1-x}{x} \right)^{1/2}, \quad (123)$$

where  $d$  is the distance from the observer to the lensed star, and  $xd$  (where  $0 < x < 1$ ) is the distance from the observer to the lensing MACHO. The angle  $\theta_E$  is called the **Einstein radius**. If  $x \approx 0.5$  (the MACHO is at the halfway point between us and the LMC), then

$$\theta_E = 4 \times 10^{-4} \left( \frac{M}{M_\odot} \right)^{1/2} \left( \frac{d}{50 \text{ kpc}} \right)^{-1/2} \text{ arcsec.} \quad (124)$$

If the MACHO does not lie perfectly along the line of sight to the star, then the image of the star will be distorted into two or more arcs instead of a single unbroken ring. Although the Einstein radius for an LMC star being lensed by a MACHO is too small to be resolved, it is possible, in some cases, to detect the amplification of the flux from the star. For the amplification to be significant, the angular distance between the MACHO and the lensed star, as seen from Earth, must be comparable to, or smaller than, the Einstein radius. Given the small size of the Einstein radius, the probability of any particular star in the LMC being

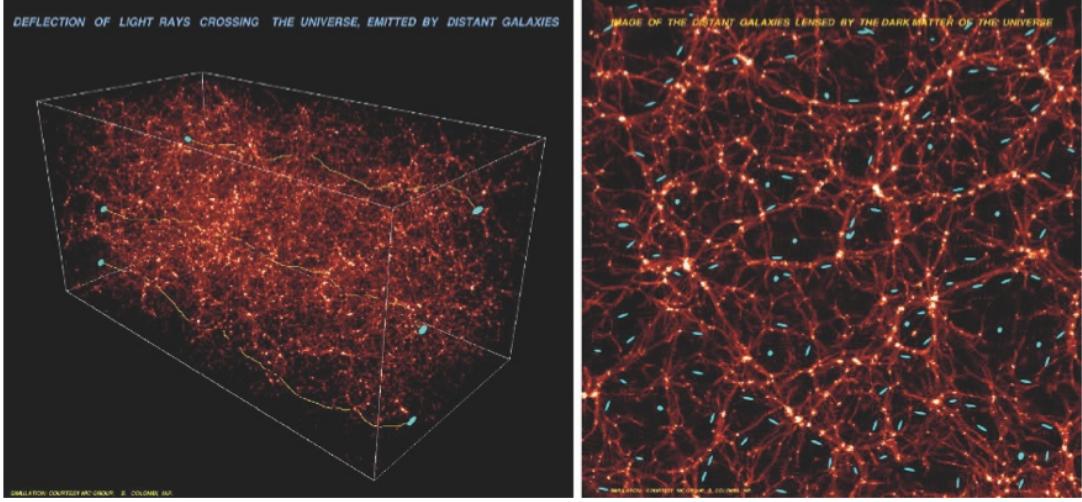


FIG. 21.— As light beams propagate through the universe they are affected by the inhomogeneous matter distribution; they are deflected, and the shape and size of their cross-section changes. This effect is displayed schematically here – light beams from sources at the far side of the cube are propagating through the large-scale distribution of matter in the universe, and we observe the distorted images of the sources. In particular, the image of a circular source is elliptical to a first approximation. Since the distribution of matter is highly structured on large scales, the image distortion caused by light deflection is coherent: the distortion of two neighbouring light beams is very similar, so that the observed ellipticities of neighbouring galaxies are correlated. From a statistical analysis of the shapes of galaxy images, conclusions about the statistical properties of the matter distribution in the universe can be drawn. Hence, the ellipticities of images of distant sources are closely related to the (projected) matter distribution, as displayed schematically in the right panel. Image taken from Schneider (2002).

lensed at any moment is tiny. It has been calculated that if the dark halo of our galaxy were entirely composed of MACHOs, then the probability of any given star in the LMC being lensed at any given time would still only be  $P \sim 10^{-7}$  (Ryden 2002, pg. 173).

To detect lensing by MACHOs, various research groups took up the daunting task of monitoring millions of stars in the LMC to watch for changes in their flux. Since the MACHOs in our dark halo and the stars in the LMC are in constant relative motion, the typical signature of a “lensing event” is a star which becomes brighter as the angular distance between star and MACHO decreases, then becomes dimmer as the angular distance increases again; see Figure 20. The typical time scale for a lensing event is the time it takes a MACHO to travel through an angular distance equal to  $\theta_E$  as seen from Earth; for a MACHO halfway between here and the LMC, this is

$$t_E = \frac{d\theta_E}{2v} \approx 90 \left( \frac{M}{M_\odot} \right)^{1/2} \left( \frac{v}{200 \text{ km s}^{-1}} \right)^{-1} \text{ days}, \quad (125)$$

where  $v$  is the relative transverse velocity of the MACHO and the lensed star as seen by the observer on Earth. Generally speaking, equation (125) shows that more massive MACHOs produce larger Einstein rings and thus will amplify the lensed star for a longer time (Ryden 2002, pg. 173).

The research groups which searched for MACHOs found a scarcity of short duration lensing events, suggesting that there is not a significant population of brown dwarfs (with  $M < 0.08 M_\odot$ ) in the dark halo of our Galaxy. The total number of lensing events which they detected suggest that as much as 20% of the halo mass could be in the form of MACHOs. The long time scales of the observed lensing events, which have  $\Delta t > 35$  days, suggest typical MACHO masses of  $M > 0.15 M_\odot$ <sup>9</sup>. Alternatively, the observed lensing events could be due, at least in part, to lensing objects within the LMC itself. In any case, the search for MACHOs suggests that most of the matter in the dark halo of our galaxy is due to a smoothly distributed component, instead of being congealed into MACHOs of roughly stellar mass (Ryden 2002, pg. 173). We can thus use these results to constrain the ratio  $\Omega_{\text{bary}}/\Omega_{\text{mat}}$  within the MW.

Gravitational lensing occurs at all mass scales. Suppose, for instance, that a cluster of galaxies, with  $M \sim 10^{14} M_\odot$ , at a distance  $d \sim 500$  Mpc from the MW, lenses a background galaxy at  $d \sim 1000$  Mpc. The Einstein radius for this configuration

$$\theta_E \approx 0.5 \left( \frac{M}{10^{14} M_\odot} \right)^{1/2} \left( \frac{d}{1000 \text{ Mpc}} \right)^{-1/2} \text{ arcmin}. \quad (126)$$

The arc-shaped images into which the background galaxy is distorted by the lensing cluster can thus be resolved. The mass of clusters can be estimated by the degree to which they lens background galaxies. The masses calculated in this way are in general agreement with the masses found by applying the virial theorem to the motions of galaxies in the cluster or by applying the equation of hydrostatic equilibrium to the hot intracluster gas (Ryden 2002, pg. 174). By independently measuring the baryonic mass of the cluster (e.g., through X-ray emission and integrated luminosity), we can again use these results to determine the fraction  $\Omega_{\text{bary}}/\Omega_{\text{mat}}$  for extragalactic sources.

#### COSMIC SHEAR

<sup>9</sup> This mass may be comparable to WDs, though the implied abundance seems to contradict the current model for the IMF and the observed metallicity content of the universe (Schneider 2002, pg. 74).

On traversing the inhomogeneous matter distribution in the Universe, light beams are deflected and distorted, where the distortion is caused by the tidal gravitational field of the inhomogeneously distributed matter. This effect, called **cosmic shear**, is sketched in Figure 21. By measuring the shapes of images of distant galaxies this tidal field can be mapped. From probing the tidal field, conclusions can be drawn about the matter distribution. For instance, the two-point correlation function of the image ellipticities can be measured. This is linked to the power spectrum  $P(k)$  of the matter distribution. Thus, by comparing measurements of cosmic shear with cosmological models we obtain constraints on the cosmological parameters, without the need to make any assumptions about the need to make any assumptions about the relation between luminous and DM. The most significant result that has been obtained from cosmic shear so far is a derivation of a combination of the matter density  $\Omega_{\text{mat}}$  and the normalization  $\sigma_8$  of the power spectrum of density fluctuations (Schneider 2002, pg. 330).

#### HUBBLE CONSTANT

The light travel times along the different paths (according to the multiple images) are not the same. On the one hand the paths have different geometrical lengths, and on the other hand the light rays traverse different depths of the gravitational potential of the lens, resulting in a (general relativistic) time dilation effect. The difference in the light travel times  $\Delta t$  is measurable because luminosity variations of the source are observed at different times in the individual images.  $\Delta t$  can be measured from this difference in arrival time, called the time delay. It is easy to see that  $\Delta t$  depends on the Hubble constant, or in other words, on the size of the Universe. If a universe is twice the size of our own,  $\Delta t$  would be twice as large as well; see Figure 20. Thus if the mass distribution of the lens can be modelled sufficiently well, by modelling the geometry of the image configuration, then the Hubble constant can be derived from measuring the difference in the light travel time. However, this is not considered a precision measurement of  $H_0$  since there are larger errors associated with the mass determination (Schneider 2002, pg. 131).

#### WEAK GRAVITATIONAL FIELDS

To characterize the strength of a gravitational field, we refer to the gravitational potential  $\Phi$ . The ratio  $\Phi/c^2$  is dimensionless and therefore well suited to distinguishing between strong and weak gravitational fields. For weak fields,  $\Phi/c^2 \ll 1$ . Another possible way to quantify the field strength is to apply the virial theorem: if a mass distribution is in virial equilibrium, then  $v^2 \sim \Phi$ , and weak fields are therefore characterized by  $v^2/c^2 \ll 1$ . Because the typical velocities in galaxies are  $\sim 200 \text{ km s}^{-1}$ , for galaxies  $\Phi/c^2 \sim 10^{-6}$ . The typical velocities of galaxies in a cluster of galaxies are  $\sim 1000 \text{ km s}^{-1}$ , so that in clusters  $\Phi/c^2 \sim 10^{-5}$ . Thus the gravitational fields occurring are weak in both cases (Schneider 2002, pg. 122).

**QUESTION 15**

**Describe cosmological inflation. List at least three important observations which it is intended to explain.**

## QUESTION 15

**Describe cosmological inflation. List at least three important observations which it is intended to explain.**

### FLATNESS PROBLEM

The **flatness problem** can be summarized by the statement, “The universe is nearly flat today, and was even flatter in the past.” We get into the heart of the flatness problem by writing the Friedmann equation in the form

$$1 - \Omega(t) = \frac{H_0^2(1 - \Omega_0)}{H(t)^2 a(t)^2} \quad (127)$$

Note that since the right-hand side of equation (127) can't change sign as the universe expands, neither can the left hand side. If  $\Omega < 1$  at any time, it remains less than one for all time; similarly, if  $\Omega > 1$  at any time, it remains greater than one for all times, and if  $\Omega = 1$  at any time,  $\Omega = 1$  at all times (Ryden 2002, pg. 65). The results of the Type Ia supernova observations and the measurements of the CMB anisotropy are consistent with the value  $|1 - \Omega_0| < 0.2$ , and so we can use equation (127) to determine how close to flatness the universe was in the past. If we extrapolate back to radiation-domination when  $|1 - \Omega| \propto a^2 \propto t$ , then at, say the Planck time  $t_P = 10^{-44}$  s, we find that  $|1 - \Omega_P| = 10^{-60}$  (Ryden 2002, pg. 235). In essence, the flatness problem arises since  $\Omega = 1$  is an unstable equilibrium. The slightest deviations from  $\Omega$  being unity shortly after the Big Bang would have been greatly amplified over the course of 13 Gyr to the present day. The fact that we observe the universe to be so close to flatness today implies that it must have been *incredibly* close to flatness. We explore this “fine-tuning” problem in more detail in the next question.

### HORIZON PROBLEM

The **horizon problem** is simply the statement that the universe is nearly homogeneous and isotropic on very large scales. To see why this is a problem, consider two antipodal points on the last scattering surface. The current proper distance to the last scattering surface is

$$d_p(t_0) = c \int_{t_{ls}}^{t_0} \frac{dt}{a(t)}. \quad (128)$$

Since the last scattering of the CMB photons occurred a long time ago ( $t_{ls} \ll t_0$ ), the current proper distance to the last scattering surface is only slightly smaller than the current horizon distance. Thus, two antipodal points on the last scattering surface, separated by  $180^\circ$  as seen by an observer on Earth, are currently separated by a proper distance of  $\approx 2d_{\text{hor}}(t_0)$ ; hence, they are causally *disconnected*. That is, they haven't had time to send messages to each other, and in particular, haven't had time to come into thermal equilibrium with each other. Nevertheless, from CMB measurements, we know that the two points have the same temperature to within one part in  $10^5$  (Ryden 2002, pg. 237).

The near-isotropy of the CMB is still more remarkable when it is recalled that the temperature fluctuations in the CMB result from the density and velocity fluctuations that existed at the time of last scattering. During this time, the universe was matter-dominated, so the horizon distance at that time is  $d_{\text{hor}}(t_{ls}) \approx 2c/H(t_{ls}) = 0.4$  Mpc. Thus, points more than 0.4 Mpc apart at the time of last scattering were not in causal contact at the time their signature was left imprinted on the CMB. Since the angular-diameter distance to the last scattering surface is  $d_A \approx 13$  Mpc, points on the last scattering surface that were separated by a horizon distance will have an angular separation equal to  $\theta_H = 0.4 \text{ Mpc}/13 \text{ Mpc} \approx 2^\circ$ , as seen from the Earth today. Therefore, points on the last scattering surface separated by an angle as small as  $\sim 2^\circ$  were out of contact with each other at the time the temperature fluctuations were stamped upon the CMB. Nevertheless, we find that  $\delta T/T$  is as small as  $10^{-5}$  on scales  $\theta > \theta_H$  (Ryden 2002, pg. 238).

### THE MONOPOLE PROBLEM

The **monopole problem** is defined by the apparent lack of magnetic monopoles in the universe. This is not a purely cosmological problem, but one that results when combining the Hot Big Bang scenario with the particle physics concept of a Grand Unified Theory (GUT). In particle physics, a GUT is a theory which attempts to unify the electromagnetic force, the weak nuclear force, and the strong nuclear force. In fact, it has been demonstrated that at particle energies greater than  $E_{ew} \sim 1$  TeV, the electromagnetic force and the weak force unite to form a single “electroweak” force. This corresponds to a thermal temperature of  $T_{ew} \sim 10^{16}$  K that would have occurred when the universe was  $t_{ew} \sim 10^{-12}$  s old. Similarly, by extrapolating the known properties of the strong and electroweak forces to higher particle energies, physicists estimate that at an energy  $E_{\text{GUT}} \sim 10^{12}$  TeV, the strong and electroweak forces should be unified as a single GUT. The corresponding universe age is  $t_{\text{GUT}} \sim 10^{-36}$  s and temperature  $T_{\text{GUT}} \sim 10^{28}$  K; the GUT energy is about four orders of magnitude smaller than the Planck energy,  $E_P \sim 10^{16}$  TeV. One of the predictions of GUTs is that the universe underwent a *phase transition* as the temperature dropped below the GUT temperature. Generally speaking, phase transitions are associated with a spontaneous loss of symmetry as the temperature of a system is lowered (e.g., when we freeze water, the water goes from having an isotropic, rotationally symmetric distribution of orientations, to an anisotropic crystalline structure). In general, phase transitions associated with a loss of symmetry give rise to flaws known as **topological defects** (e.g., domain walls in aluminum railings). It is predicted that the GUT phase transition creates point-like topological defects which act as magnetic monopoles (i.e., an isolated north or south pole) (Ryden 2002, pg. 239).

The rest energy of the magnetic monopoles created in the GUT phase transition is predicted to be  $m_{MC}^2 \sim E_{\text{GUT}} \sim 10^{12}$  TeV. This corresponds to a mass of over a nanogram (comparable to that of a bacterium), which is a lot of mass for a single particle to

be carrying around. At the time of the GUT phase transition, points further apart than the horizon size will be out of causal contact with each other. Thus, we expect roughly one topological defect per horizon volume, due to the mismatch of fields which are not causally linked. The number density of magnetic monopoles, at the time of their creation, would be  $n_M \sim 1/(2ct_{\text{GUT}}) \sim 10^{82} \text{ m}^{-3}$ , and with a density  $\rho_M \propto a^{-3}$ , they should have dominated the energy content of the universe by a time  $t \sim 10^{-16} \text{ s}$ . Obviously, the universe is *not* dominated by magnetic monopoles today. In fact, there is no strong evidence that they exist at all. Since there is not yet a single, definite GUT, and in some variants of the GUT theme, magnetic monopoles are not produced, one solution to this problem is that magnetic monopoles never existed in the first place. Nevertheless, even if this is the case, the flatness and horizon problems still present problems to the standard Big Bang scenario (Ryden 2002, pg. 241).

#### THE INFLATION SOLUTION

Inflation is posited to solve each of the flatness, horizon, and monopole problems. In a cosmological context, inflation can most generally be defined as the hypothesis that there was a period, early in the history of our universe, when the expansion was exponentially accelerating outward. The usual implementation of inflation states that the universe was temporarily dominated by a positive cosmological constant  $\Lambda_i$  (with  $w = -1$ ) so that  $a(t) \propto e^{H_i t}$ , where  $H_i = (\Lambda_i/3)^{1/2}$  (Ryden 2002, pg. 242).

To see how a period of exponential growth can resolve the flatness, horizon, and monopole problems, suppose that the universe had a period of exponential expansion sometime in the midst of its early, radiation-dominated phase. For simplicity, suppose the exponential growth was switched on instantaneously at a time  $t_i$ , and lasted until some later time  $t_f$ , when the exponential growth was switched off instantaneously, and the universe reverted to its former state of radiation-dominated expansion. In this simple case, we can write the scale factor as

$$a(t) = \begin{cases} a_i(t/t_i)^{1/2} & t < t_i \\ a_i e^{H_i(t-t_i)} & t_i < t < t_f \\ a_i e^{H_i(t-t_i)}(t/t_f)^{1/2} & t > t_f \end{cases} \quad (129)$$

Thus, between the time  $t_i$  when the exponential inflation began and the time  $t_f$  when the inflation stopped, the scale factor increased by a factor

$$\frac{a(t_f)}{a(t_i)} = e^N, \quad (130)$$

where  $N$ , the number of e-foldings of inflation, was  $N \equiv H_i(t_f - t_i)$ . If the duration of inflation,  $t_f - t_i$ , was long compared to the Hubble time during inflation,  $H_i^{-1}$ , then  $N$  was large, and the growth in scale factor during inflation was enormous. For concreteness, we can take one model of inflation where exponential expansion starts around the GUT time and lasts for  $N = 100$  Hubble times, so that  $a$  grows by a factor of  $10^{43}$  (Ryden 2002, pg. 243).

To solve the flatness problem, we note that during inflation equation (127) would proceed as

$$|1 - \Omega(t)| \propto e^{-2H_i t}, \quad (131)$$

so that the difference between  $\Omega$  and unity decreases exponentially with time. If we compare the density parameter at the beginning of exponential inflation with the density parameter at the end of inflation, we find

$$|1 - \Omega(t_f)| = e^{-2N} |1 - \Omega(t_i)| \sim 10^{-87} |1 - \Omega(t_i)|, \quad (132)$$

for our fiducial inflation scenario. Thus, even if the universe was actually fairly strongly curved, with  $|1 - \Omega(t_i)| \sim 1$ , equation (132) shows that 100 e-foldings of inflation would have flattened it like a proverbial pancake. With the current limits on the density parameter, we would require a minimum of  $N > 60$  e-foldings of inflation, if it occurred around the GUT time (Ryden 2002, pg. 244).

Next we show how inflation solves the horizon problem. We can use equation (129) to show that the horizon size at the end of inflation is

$$d_{\text{hor}}(t_f) = e^N c(2t_i + H_i^{-1}), \quad (133)$$

so that the horizon size grows exponentially with time during inflation. If inflation started at  $t_i \sim 10^{-36} \text{ s}$ , with a Hubble parameter  $H_i \approx t_i^{-1} \sim 10^{36} \text{ s}^{-1}$ , and lasted for  $N = 100$  e-foldings, then the horizon size immediately before and after inflation was

$$\begin{aligned} d_{\text{hor}}(t_i) &\approx 2ct_i \sim 10^{-28} \text{ m} \\ d_{\text{hor}}(t_f) &\approx e^N 3ct_i \sim 0.8 \text{ pc}. \end{aligned} \quad (134)$$

During the brief period of  $\sim 10^{-34} \text{ s}$  that inflation lasts in this model, the horizon size is boosted exponentially from submicroscopic scales to nearly a parsec. At the end of the inflationary epoch, the horizon size reverts to growing at a sedate linear rate. The net result of inflation is to increase the horizon length in the post-inflationary universe by a factor  $\sim e^N$  over what it would have been without inflation. For instance, we found that, in the absence of inflation, the horizon size at the time of last scattering was  $d_{\text{hor}}(t_{ls}) \approx 0.4 \text{ Mpc}$ . Given a hundred e-foldings of inflation in the early universe, however, the horizon size at last scattering would have been  $\sim 10^{43} \text{ Mpc}$ , obviously gargantuan enough for the entire last scattering surface to be in causal contact (Ryden 2002, pg. 246).

If magnetic monopoles were created before or during inflation, then the number density of monopoles was diluted to any undetectably low level. During a period when the universe was expanding exponentially ( $a \propto e^{H_i t}$ ), the number density of monopoles, if they were neither created nor destroyed, was decreasing exponentially ( $n_M \propto e^{-3H_i t}$ ). For instance, if inflation

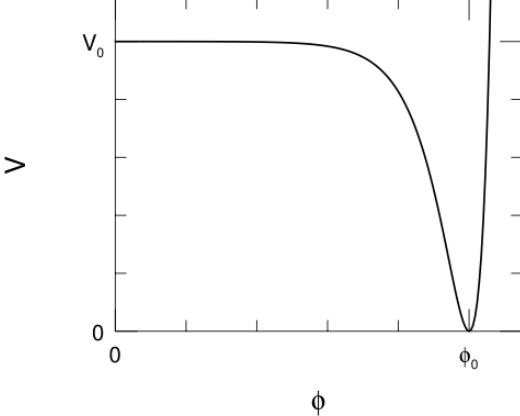


FIG. 22.— A potential which can give rise to an inflationary epoch. The global minimum in  $V$  (or “true vacuum”) is at  $\phi = \phi_0$ . If the scalar field starts at  $\phi = 0$ , it is in a “false vacuum” state. Image taken from Ryden (2002).

started around the GUT time with  $N = 100$ , then the end result is that we should expect to find  $n_M \sim 10^{-61} \text{ Mpc}^{-3}$  today. The probability of finding even a single monopole within the last scattering surface would be astronomically small (Ryden 2002, pg. 247).

#### THE PHYSICS OF INFLATION

Suppose the universe contains a scalar field  $\phi(\vec{r}, t)$ , called the **inflaton field**, whose value can vary as a function of position and time. Generally speaking, a scalar field can have an associated potential energy  $V(\phi)$ . If  $\phi$  has units of energy and  $V$  has units of energy density, then the energy density and pressure of the inflaton field are

$$\begin{aligned}\rho_\phi &= \frac{1}{2} \frac{1}{\hbar c^3} \dot{\phi}^2 + V(\phi) \\ P_\phi &= \frac{1}{2} \frac{1}{\hbar c^3} \dot{\phi}^2 - V(\phi).\end{aligned}\quad (135)$$

If the inflaton field changes only very slowly as a function of time, with

$$\dot{\phi}^2 \ll \hbar c^3 V(\phi), \quad (136)$$

then the inflaton field acts like a cosmological constant, with

$$\rho_\phi = -P_\phi = V(\phi). \quad (137)$$

Thus, an inflaton field can drive exponential inflation if there is a temporary period when its rate of change of  $\dot{\phi}$  is small, and its potential energy  $V(\phi)$  is large enough to dominate the energy density of the universe (Ryden 2002, pg. 248).

By plugging  $\rho_\phi$  and  $P_\phi$  into the fluid equation, with the simplification in equation (137), it can be shown that the requirement of exponential inflation is

$$\left( \frac{dV}{d\phi} \right)^2 \ll \frac{V_0^2}{E_P^2}, \quad (138)$$

where  $E_P$  is the Planck energy and  $V_0$  is the initial potential at  $\phi = \phi_{\text{init}}$ . Hence, if the slope of the inflaton’s potential is sufficiently shallow, satisfying equation (138), and if the amplitude of the potential is sufficiently large to dominate the energy density of the universe, then the inflaton field is capable of giving rise to exponential expansion (Ryden 2002, pg. 249).

As a concrete example of a potential  $V(\phi)$  which can give rise to inflation, consider the potential shown in Figure 22. If the plateau where  $V \approx V_0$  satisfies equation (138), then while  $\phi$  is slowly rolling toward the global minimum at  $\phi_0$ , the inflaton field contributes an energy density  $\rho_\phi \approx V_0 = \text{const.}$  to the universe. When an inflaton field has a potential similar to that of Figure 22, it is referred to as being in a **metastable false vacuum state** when it is near the maximum at  $\phi = 0$ . Such a state is not truly stable; if the inflaton field is nudged from slightly, it will continue to slowly roll toward the **true vacuum** state at  $\phi = \phi_0$  and  $V = 0$ . However, if the plateau is sufficiently broad as well as sufficiently shallow, it can take many Hubble times for the inflaton field to roll down to the true vacuum state. The ensuing exponential expansion ends when the inflaton field reaches the true vacuum at  $\phi = \phi_0$  (Ryden 2002, pg. 251).

After rolling off the plateau in Figure 22, the inflaton field  $\phi$  oscillates about the minimum at  $\phi_0$ . If the inflaton field is coupled to any of the other fields in the universe, however, the oscillations in  $\phi$  are damped more rapidly, with the energy of the inflaton field being carried away by photons or other relativistic particles. These photons **reheat** the universe after the precipitous drop in temperature (falls off like  $e^{-N}$ ) caused by inflation. The energy lost by the inflaton field after its phase transition from the false vacuum to the true vacuum can be thought of as the latent heat of that transition. That is, the transition from false to true vacuum

releases an energy  $V_0$  which goes to reheat the universe. Note that this reheating process is required in order to compensate for the incredible drop in temperature during the inflationary epoch. As the energy density associated with the inflaton field was converted to relativistic particles such as photons, the temperature of the universe was restored to its pre-inflationary value (Ryden 2002, pg. 252).

Inflation successfully explains the flatness, homogeneity, and isotropy of the universe. It ensures that we live in a universe with a negligibly low density of magnetic monopoles, while the inclusion of reheating ensures that we don't live in a universe with a negligibly low density of photons. In some ways, though, inflation seems to be too successful. It makes the universe homogeneous and isotropic all right, but it makes it *too* homogeneous and isotropic, to the point that we would expect the CMB to be much smoother than it is today. Remember, however, the saga of how a submicroscopic patch of the universe was inflated to macroscopic size, before growing to the size of the currently visible universe. Inflation excels in taking submicroscopic scales and blowing them up to macroscopic scales. On submicroscopic scales, the vacuum, whether true or false, is full of constantly changing quantum fluctuations, with virtual particles popping into and out of existence. On quantum scales, the universe is intrinsically inhomogeneous. Inflation takes the submicroscopic quantum fluctuations in the inflaton field and expands them to macroscopic scales. The energy fluctuations that result are the origin, in the inflationary scenario, of the inhomogeneities in the current universe (Ryden 2002, pg. 253). Therefore, it is inflation coupled with quantum fluctuations that produces the large scale structures we see today.

**QUESTION 16**

**Define and describe the ‘fine tuning problem’. How do anthropic arguments attempt to resolve it?**

### QUESTION 16

Define and describe the ‘fine tuning problem’. How do anthropic arguments attempt to resolve it?

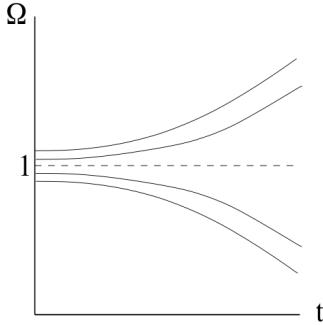


FIG. 23.— The flatness problem arises since  $\Omega = 1$  is an unstable equilibrium in that  $\Omega(z)$  tends to evolve away from unity as the universe evolves. Image taken from Albrecht (2001).

The flatness problem is most easily investigated when we consider writing the Friedmann equation in the form

$$H^2 = \frac{8\pi G}{3} \rho(t) - \frac{\kappa c^2}{a^2} \Rightarrow \frac{1 - \Omega(t)}{\Omega(t)} = -\frac{3\kappa c^2}{8\pi G \rho a^2}. \quad (139)$$

If we extrapolate back to radiation-dominated times where  $\rho \propto a^{-4}$  and  $a \propto t^{1/2}$ , this relation becomes

$$\left| \frac{1 - \Omega(t)}{\Omega(t)} \right| \propto t \Rightarrow \left| \frac{1 - \Omega(t)}{\Omega(t)} \right| = \left( \frac{t}{t_0} \right) \left| \frac{1 - \Omega_0}{\Omega_0} \right|. \quad (140)$$

The most recent seven-year measurements from WMAP place  $\Omega_0 = 1.002$  (Jarosik et al. 2011). Plugging this value into equation (140) and extrapolating back to  $t = 1$  s after the Big Bang yields

$$\left| \frac{1 - \Omega(1 \text{ s})}{\Omega(1 \text{ s})} \right| \approx 10^{-20}. \quad (141)$$

We thus see that in order for the universe to be so close to flatness today, it must have been even closer to flatness at earlier times, with the requirement becoming more stringent the further back we probe. Physically, this occurs because a flat universe for which  $\Omega = 1$  represents an unstable point in the evolution of the universe. Any deviations from flatness at early times become amplified by the expansion of the universe. We can see this from looking at the expression in equation (139). Because  $\rho \propto a^{-3}$  or  $a^{-4}$  throughout the early history of the universe, the  $\rho$  term in the Friedmann equation falls away much more quickly than the  $\kappa/a^2$  term as the universe expands, and the  $\kappa/a^2$  term comes to dominate (Albrecht 2001). This behaviour is illustrated in Figure 23.

The fact that  $\Omega$  must have been so close to unity at early times is often referred to as a **fine-tuning problem**. Generally speaking, fine-tuning refers to circumstances when the parameters of a model must be adjusted very precisely in order to agree with observations (Etsuko). However, it may be seem obvious to question the extent to which this issue can be regarded as a scientific problem. In fact, it is often acknowledged that the legitimacy of the flatness problem cannot be deduced logically, but rather holds on intuitive grounds. Possibly the most convincing intuitive argument in favour of the flatness problem is the analogy drawn by Albrecht (2001). In this analogy, you imagine walking into your boss’ office and finding a pencil balanced on its point. You continue to find this pencil balanced every day for the next year until you finally inquire about its origins. Your boss acknowledges that it is quite remarkable to find the pencil this way, but claims that there is no need to explain its origins since he found it this way when he moved into his office. You would probably not find this answer satisfactory, but, on the other hand, if the pencil had simply been lying there, you probably wouldn’t have asked about it in the first place. The pencil remaining balanced for so long is analogous to finding  $\Omega_0$  equal to unity. As in the pencil analogy, it is unsatisfactory to ignore the extraordinary origins that must have brought  $\Omega_0$  to 1. Albrecht (2001) argues that this is what constitutes the flatness problem as a true scientific problem.

We have already seen how an exponential inflation scenario solves the flatness problem. Here we will focus on a solution based on the **anthropic principle**. Hawking (1974) considers the consequences that different initial conditions of the Big Bang would have had on the evolution of the universe. For instance, suppose that the universe is open ( $\Omega_0 < 1$ ). In this case, the density is insufficient to halt the expansion of the universe resulting from the Big Bang. After only a short period of time the universe would be expanding too quickly for gravitationally bound systems to form. On the other hand, consider that the universe is closed ( $\Omega_0 > 1$ ). In this case, the density of the universe will halt the expansion of the universe and cause it to collapse at some finite time. In the event that the universe is flat ( $\Omega_0 = 1$ ), the universe will expand at just the correct rate to allow both gravitationally bound systems to form and to avoid collapse. The idea presented here is that only the flat universe presents conditions that are

favourable to the development of intelligent life. In order for intelligent life to develop, gravitationally bound systems such as galaxies and stars must exist. This will not occur in an open universe since it will expand too quickly for bound systems to form. In addition, not only do gravitationally bound systems need to exist, but they need to exist long enough for biological evolution to occur. This condition cannot be satisfied in a closed universe since it will collapse too quickly. In particular, if the rate of expansion of the universe differed by one part in  $10^{12}$  at the time the universe was  $T \sim 10^{10}$  K, the universe would have started recollapsing when it was only  $10^{-4}$  of its present size and the temperature was still  $10^4$  K. Hence, there is not enough time for biological evolution to sustain intelligent life in a closed universe. However, in a flat universe the expansion will be just right for both gravitationally bound systems to form and ample time for intelligent life to evolve. Therefore, the only universe in which intelligent life can exist is in one which is flat. Since the human population clearly exists, then it must be that the universe is flat. In this sense, the flatness problem is no concern since it has been shown that we are living in a universe where  $\Omega$  is exactly equal to unity (Hawking 1974).

This consideration can be interpreted as follows: we live in a universe which had, at a very early time, a very precisely tuned density parameter, because only in such a universe can life evolve and astronomers exist to examine the flatness of the universe. In all other conceivable universes this would not be possible. This approach is meaningful only if a large number of universes existed – in this case we should not be too surprised about living in one of those where this initial fine-tuning took place – in the other ones, we, and the question about the cosmological parameters, would just not exist. This anthropic reasoning may either be seen as an “explanation” for the flatness of *our* universe, or as a capitulation – where we give up attempting to solve the question of the origin of the flatness of the Universe (Schneider 2002, 173).

The anthropic principle has had a controversial history regarding the validity of its use as a scientific proof. One argument against its use has been presented by Smolin (2004). Here it is argued that most anthropic arguments within inflationary cosmology are based on two assumptions. The first, which we will call **A**, is that there exists a large ensemble of universes,  $\mathcal{M}$ , with differing physical parameters, and are causally disjoint in the sense that we have no ability to make observations in other universe than our own. The second, which we will call **B**, is that the parameters within  $\mathcal{M}$  are randomly distributed, and that the parameters we find in our universe are in some sense rare. In the context of the flatness problem, **A** would represent the collection of universes with differing values for  $\Omega_0$  and **B** is the assumption that our universe is rare in that  $\Omega_0$  is so close to unity. The argument presented by Smolin (2004) is that any theory based on **A** and **B** cannot be falsifiable in the sense that there is no possible experiment that could contradict both **A** and **B**. Here is the basic argument: if such an anthropic theory applies to nature, then it follows that our universe is a member of  $\mathcal{M}$ . Thus we can assume that whatever properties our universe is known to have, or is discovered to have in the future, it remains true that there is at least one member of  $\mathcal{M}$  that has those properties. Therefore, no experiment, present or future, could contradict **A** and **B** (since there is no way for us to experiment in any other universe). Moreover, since, by **B**, we already assume that there are properties of our universe that are improbable in  $\mathcal{M}$ , it is impossible to make even a statistical prediction that, were it not borne out, would contradict **A** and **B**.

**QUESTION 17**

**Define the two-point correlation function. How is it related to the power spectrum? How is the  $C_l$  spectrum of the CMB related to low redshift galaxy clustering?**

### QUESTION 17

**Define the two-point correlation function. How is it related to the power spectrum? How is the  $C_l$  spectrum of the CMB related to low redshift galaxy clustering?**

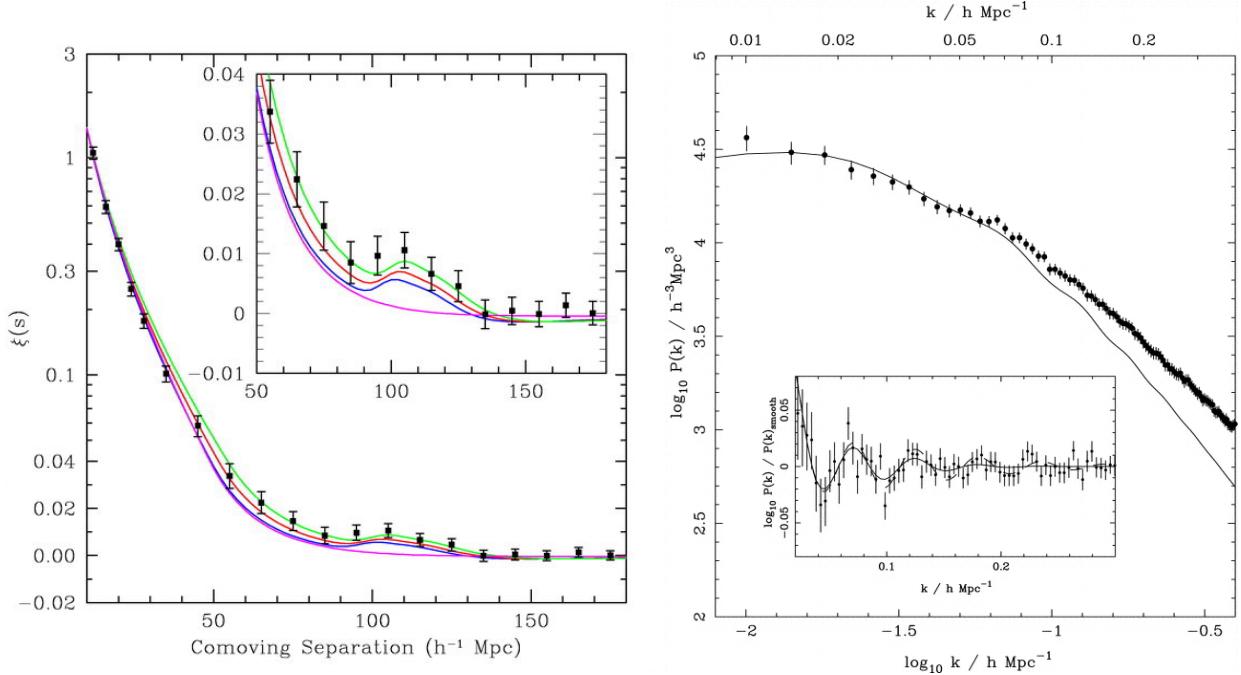


FIG. 24.— (left) The correlation function computed from SDSS galaxies. The peak at  $100/h$  Mpc separation denotes the horizon scale at recombination. The solid lines show models with  $\Omega_{\text{mat}}h^2 = 0.12, 0.13, 0.14$  (green, red, blue respectively), all with  $\Omega_{\text{bary}}h^2 = 0.024$ . The magenta lines shows a pure CDM model ( $\Omega_{\text{mat}}h^2 = 0.105$ ) which lacks the acoustic peak. Image taken from Eisenstein et al. (2005). (right) The matter power spectrum as computed from analysis of galaxies identified by the SDSS. The inset shows the baryon wiggles caused by the baryonic acoustic oscillations (BAO). Image taken from Percival et al. (2007).

We begin by investigating the probability of finding a galaxy within the volume element  $dV$  around some location  $\vec{x}$ . If  $\bar{n}$  is the mean number density of galaxies then this probability is simply  $P_1 = \bar{n}dV$ , assuming that the universe is spatially homogeneous. Now if the distribution of galaxies were uncorrelated, then the probability of finding a galaxy within  $dV$  at  $\vec{x}$  at the same time as finding one within  $dV$  at  $\vec{y}$  would simply be  $P_2 = P_1^2$ . However, we know that the distribution of galaxies is in fact correlated and so we modify this to be

$$P_2 = P_1^2(1 + \xi(\vec{x}, \vec{y})), \quad (142)$$

where  $\xi$  is the **two-point correlation function**. Since the universe is considered homogeneous and isotropic, the correlation function is only a function of the magnitude of the separation,  $r = |\vec{x} - \vec{y}|$ , so that  $\xi = \xi(r)$ . The correlation function  $\xi(r)$  takes into account the fact that galaxies are not randomly distributed in space, but are rather found in organized groups and clusters. As a consequence, the probability of finding a galaxy at some location  $\vec{x}$  is not independent of whether there is a galaxy in the vicinity of  $\vec{x}$ . In other words, because of their clustering, it is more likely to find a galaxy in the vicinity of another (Schneider 2002, pg. 283).

The correlation function is obtained observationally by averaging over the density products for a large number of pairs of galaxies with given separation  $r$ . This is achieved through galaxy redshift surveys such as the Two-degree-Field Galaxy Redshift Survey (2dFGRS) and the Sloan Digital Sky Survey (SDSS). For example, Eisenstein et al. (2005) compute  $\xi(r)$  from the SDSS galaxies with the result being displayed in Figure 24. Because the universe has a significant fraction of baryons, the baryon acoustic oscillations (BAO) in the early-universe plasma will be imprinted onto the late-time matter power spectrum. This is observed in Figure 24 where a peak at a separation of  $100/h$  Mpc reflects the scale of the sound horizon at recombination. This plot shows how the location of this peak changes with changing values of the matter density for a fixed baryon density and also contrasts this to the case of pure dark matter. Evidently, when  $\Omega_{\text{mat}}$  is increased, the peak shifts to smaller scales (more on this below).

An equivalent description of the statistical properties of the structure in the universe is to look at the matter power spectrum  $P(k)$ , which describes matter density fluctuations as a function of length scale  $L \simeq 2\pi/k$ ; large values of  $P(k)$  denote large fluctuations in wave number  $k$ . The power spectrum and correlation function are related via a Fourier transform<sup>10</sup>:

$$P(k) = 2\pi \int_0^\infty r^2 \frac{\sin kr}{kr} \xi(r) dr. \quad (143)$$

<sup>10</sup> This looks different than the standard three-dimensional Fourier transform because the angular integrals have already been evaluated (this is done since  $\xi$  is isotropic, depending only on separation).

The correlation function can similarly be derived from the power spectrum via an inverse Fourier transformation (Schneider 2002, pg. 284). From this equation we should expect that the peaks in  $\xi(r)$  should be transformed into wiggles in  $P(k)$ ; see the right panel of Figure 24. Note that the power spectrum shown here can be extended to smaller and larger scales though other methods. In particular,  $P(k)$  is largely constrained at scales  $k < 0.01 h/\text{Mpc}$  from the CMB measurements which provide the primordial potential fluctuations via the Sachs-Wolfe effect. On the other hand, it can be expanded to smaller scales by measuring the density of neutral hydrogen from Ly $\alpha$  absorption along the line of sight to quasars. Again the difficulty with this method is determining how to relate the density of hydrogen to the density of matter. Finally, the large-scale power spectrum can be observed through gravitational lensing (Rich 2010, pg. 257).

We can therefore determine the matter power spectrum by first calculating the correlation function of galaxies through redshift surveys and applying a Fourier transform prescribed by equation (143). Note that the obvious caveat in constructing the power spectrum in this way is that the fluctuations in the number of galaxies do not necessarily faithfully trace fluctuations in matter density. As a result, it is customary to assume that the matter density contrast is proportional to the galaxy number density contrast via

$$\frac{\Delta n_{\text{gal}}}{n_{\text{gal}}} = b \frac{\Delta \rho}{\rho}, \quad (144)$$

where  $b$  is called the bias parameter. This parameter can indeed be a complicated function of scale, environment, and galaxy properties and, as such, the difficulty in estimating  $b$  dominates the systematic uncertainties in determining  $P(k)$  (Rich 2010, pg. 257).

The thin solid line in the right panel of Figure 24 shows the calculated  $P(k)$  using the 3-year WMAP parameters. The excess power observed for  $k > 0.1$  may be due to a scale-dependent bias parameter  $b$  in the galaxy sample that was not included in the construction of the power spectrum. Another type of bias that can arise is from galaxy type bias. The power spectra of two different types of galaxies should be proportional to each other, up to an amplitude difference arising from different bias parameters  $b$ . Indeed, comparing the power spectra of red and blue galaxies from the 2dFGRS shows that the bias factor for red galaxies is  $\sim 1.4$  larger than that for blue galaxies. This result is consistent with red galaxies being preferentially located in large clusters, with the opposite being true for blue galaxies. Hence, red galaxies appear to follow the dark matter density profile more closely than blue galaxies (Schneider 2002, pg. 314).

#### PHYSICAL INTERPRETATION

Many of the physical processes left imprinted on the  $C_l$  spectrum of the CMB also arise in the clustering of low-redshift galaxies; most notable are the BAO. To understand how this can be the case, we consider what happens after recombination. Imagine that recombination happened instantaneously; then right at that moment there are density fluctuations in the DM component as well as the acoustic oscillations in the baryons. The photons can stream freely, due to the absence of free electrons, and the sudden drop of pressure in the baryon component reduces the sound speed from  $\sim c$  essentially to zero. At this point, the baryons can now fall into the DM potential wells. However, since the cosmic baryon density is only about six times smaller than that of the dark matter, the baryonic density fluctuations at recombination are not completely negligible compared to those of the DM. Therefore, they form their own potential wells, and part of the dark matter will fall into them. After some time, baryons and DM have about the same spatial distribution, which is described by the linear evolution of the density field, where the initial condition is a superposition of the DM fluctuations at recombination plus that of the baryonic oscillations. Whereas the corresponding density contrast of the latter is small compared to that of the dark matter, it has the unique feature that it carries a well-defined length-scale, namely the sound horizon at recombination. The matter correlation function in the local universe should therefore contain a feature at just this length-scale. If galaxies trace the underlying matter distribution, this length-scale should then be visible in the galaxy correlation function (Schneider 2002, pg. 334). The peak in  $\xi(r)$  at a separation of  $\sim 100$  Mpc reflects the scale of the sound horizon at recombination. Note that the subsequent acoustic peaks observed in  $P(k)$  transform to a single peak in the correlation function. The decreasing envelope of the higher harmonics, due to Silk damping, as well as non-linear gravity, corresponds to broadening of the single peak (Eisenstein et al. 2005).

The form of  $P(k)$  can be understood in the context of the  $\Lambda$ CDM model. First, the clustering of DM is suppressed on scales small enough to have been traversed by the neutrinos and/or photons during the radiation-dominated period of the universe. This introduces the characteristic turnover in the DM power spectrum at the scale of the horizon at matter-radiation equality. This length scales as  $\Omega_{\text{mat}}^{-1}$ . Second, the acoustic oscillations have a characteristic scale known as the sound horizon, which depends both on the expansion history of the early universe and on the sound speed in the plasma; this scales with  $\Omega_{\text{mat}}^{-0.25}$  and  $\Omega_{\text{bary}}^{-0.08}$ . We can also use our standard cosmological model to understand the shape of  $\xi(r)$ . We saw above that increasing  $\Omega_{\text{mat}}$  caused the peak in  $\xi(r)$  to shift to smaller separations. We know that larger values of  $\Omega_{\text{mat}}$  correspond to earlier epochs of matter-radiation equality, which increase the amount of small-scale power compared to large, which in turn decreases the correlations on large scales when holding the small-scale amplitude fixed (Eisenstein et al. 2005).

**QUESTION 18**

Consider a cosmological model including a positive cosmological constant. Show that, in such a model, the expansion factor eventually expands at an exponential rate. Sketch the time dependence of the expansion factor in the currently favoured cosmological model.

### QUESTION 18

Consider a cosmological model including a positive cosmological constant. Show that, in such a model, the expansion factor eventually expands at an exponential rate. Sketch the time dependence of the expansion factor in the currently favoured cosmological model.

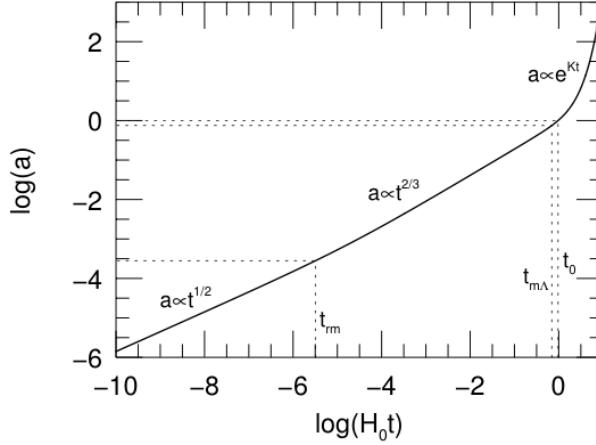


FIG. 25.— The scale factor  $a$  as a function of time  $t$  (measured in units of the Hubble time), computed for the Benchmark Model. The dotted lines indicate the time of radiation-matter equality,  $a_{\text{rm}} \approx 3 \times 10^{-4}$ , the time of matter- $\Lambda$  equality,  $a_{\text{m}\Lambda} \approx 0.7$ , and the present moment,  $a_0 = 1$ . Image taken from (Ryden 2002).

We have previously seen that the Friedmann equation can be written in the form

$$H(a) = H_0 \sqrt{\frac{\Omega_{\text{rad}}}{a^4} + \frac{\Omega_{\text{mat}}}{a^3} + \Omega_\Lambda + \frac{\Omega_\kappa}{a^2}}, \quad (145)$$

and since  $H(a) \equiv \dot{a}/a$ , this can be rewritten as

$$\dot{a} = H_0 \sqrt{\frac{\Omega_{\text{rad}}}{a^2} + \frac{\Omega_{\text{mat}}}{a} + \Omega_\Lambda a^2 + \Omega_\kappa}. \quad (146)$$

In the limit that  $a \gg 1$  we see that only one term survives:

$$\frac{da}{dt} \approx H_0 \sqrt{\Omega_\Lambda} a \Rightarrow a(t) \propto \exp[H_0 \sqrt{\Omega_\Lambda} t]. \quad (147)$$

Hence, when the cosmological constant dominates the Friedmann equation, the scale factor grows exponentially in time. We can similarly evaluate equation (146) in the limits that radiation and matter dominate the energy density of the universe:

$$a_{\text{RD}}(t) \propto t^{1/2}, \quad a_{\text{MD}}(t) \propto t^{2/3}. \quad (148)$$

Given that there are three components in the universe, each of which evolves differently with the scale factor, we can calculate the scale factor when the densities of two components are equal. This is called an epoch of equality. For instance, radiation-matter equality occurs when

$$\frac{\rho_{\text{rad}}(a)}{\rho_{\text{mat}}(a)} = 1 \rightarrow \frac{\Omega_{\text{rad}}(a)}{\Omega_{\text{mat}}(a)} = \frac{\Omega_{\text{rad},0}/a^4}{\Omega_{\text{mat},0}/a^3} = 1 \Rightarrow a_{\text{rm}} = \frac{\Omega_{\text{rad},0}}{\Omega_{\text{mat},0}} = \frac{8.5 \times 10^{-5}}{0.273} \approx 3 \times 10^{-4}. \quad (149)$$

Similarly, matter- $\Lambda$  equality occurs when

$$\frac{\rho_{\text{mat}}(a)}{\rho_\Lambda(a)} = 1 \rightarrow \frac{\Omega_{\text{mat}}(a)}{\Omega_\Lambda(a)} = \frac{\Omega_{\text{mat},0}/a^3}{\Omega_\Lambda} = 1 \Rightarrow a_{\text{m}\Lambda} = \left(\frac{\Omega_{\text{mat},0}}{\Omega_\Lambda}\right)^{1/3} = \left(\frac{0.273}{0.728}\right)^{1/3} \approx 0.7 \quad (150)$$

Thus,  $z_{\text{rm}} \approx 3333$  and  $z_{\text{m}\Lambda} \approx 0.4$ . For  $z \ll z_{\text{rm}}$  the deceleration parameter is  $q(a) = \Omega_{\text{mat}}(a)/2 - \Omega_\Lambda$ . We can use this to compute the epoch at which the acceleration of the universe changed sign:

$$\frac{\rho_{\text{mat}}(a)}{2\rho_\Lambda(a)} = 1 \rightarrow \frac{\Omega_{\text{mat}}(a)}{2\Omega_\Lambda(a)} = \frac{\Omega_{\text{mat},0}/a^3}{2\Omega_\Lambda} = 1 \Rightarrow a_{\text{accel}} = \left(\frac{\Omega_{\text{mat},0}}{2\Omega_\Lambda}\right)^{1/3} = \left(\frac{0.273}{1.456}\right)^{1/3} \approx 0.6, \quad (151)$$

so that  $z_{\text{accel}} \approx 0.7$ ; the acceleration became positive *before* the cosmological constant dominated the energy density of the universe (Carroll & Ostlie 2007, pg. 1195). A sketch of the time dependence of the scale factor for the currently accepted cosmological model is shown in Figure 25.

### DARK ENERGY

Cosmological models are based on the Friedmann equation,

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2}\rho - \frac{\kappa c^2}{R_0^2 a^2}, \quad (152)$$

and the fluid equation

$$\dot{\rho} + 3\frac{\dot{a}}{a}(\rho + P) = 0. \quad (153)$$

If we take the time derivative of equation (152), divide by  $2\dot{a}a$ , and then substitute in equation (153), we can derive the **acceleration equation**:

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(\rho + 3P). \quad (154)$$

This set of equations is then complemented with an equation of state of the form  $P = w\rho$ , where  $w$  is a dimensionless number. From equation (154) we see that  $w < -1/3$  will produce a positive acceleration  $\ddot{a} > 0$ . A component of the universe satisfying  $w < -1/3$  is generically referred to as **dark energy**. One form of dark energy is of special interest; some observational evidence indicates that our universe may contain a **cosmological constant**, defined such that  $w = -1$ . Note that using  $w = -1$  in equation (153) implies that  $\dot{\rho} = 0$ ; that is, the energy density of such a component remains constant in time (Ryden 2002, pg. 71).

The cosmological constant was initially introduced by Einstein based on erroneous interpretations of the universe which were the result of insufficient observational data at time time. In particular, it was thought that the universe was static and dominated by non-relativistic matter. However, such notions are mutually incompatible. If we create a matter-filled universe which is initially static, then gravity will cause it to contract. On the other hand, if we create a matter-filled universe which is initially expanding, then it will either expand forever or reach a maximum radius and then collapse; the exact scenario depends on the energy density of matter. Trying to make a matter-filled universe which neither expands nor collapses is like throwing a ball into the air and expecting it to hover there. To solve this problem, Einstein added a new term  $\Lambda$  to his field equations which allowed the universe to be both matter-filled and static. This comes at the price of making the universe open and teetering at an unstable equilibrium point. Expanding (contracting) the universe slightly from its current position would result in runaway expansion (contraction) as the  $\Lambda$  ( $\rho_{\text{mat}}$ ) takes over (Ryden 2002, pg. 74).

Einstein subsequently dropped the use of  $\Lambda$  when Hubble's Law was realized. However, soon after,  $\Lambda$  was brought back into action in order to reconcile the fact that the originally measured  $H_0$  implied an age of the universe much smaller than the age of the Earth. Since then  $\Lambda$  has regained popularity due to Type Ia SNe measurements indicating that the universe has a positive acceleration. In order to give  $\Lambda$  a real physical meaning, we need to identify some component of the universe whose energy density  $\rho_\Lambda$  remains constant as the universe expands or contracts. Currently, the leading candidate for this component is the **vacuum energy**. This energy is associated with virtual particle-antiparticle pairs that are quantum-mechanically allowed to spontaneously appear and then annihilate out of the vacuum. Unfortunately, computing the value of  $\rho_{\text{vac}}$  is an exercise in QFT which has not yet been successfully completed. It has been suggested that the natural value for the vacuum energy density is the Planck energy density,

$$\rho_{\text{vac}} \sim \frac{E_P}{l_P^3}, \quad (155)$$

where the Planck energy is  $E_P \sim 10^{28}$  eV and the Planck length is  $l_P \sim 10^{-35}$  m. Unfortunately, this gives a value of  $\rho_{\text{vac}} \sim 10^{133}$  eV m<sup>-3</sup>; 124 orders of magnitude larger than the current critical density for our universe (Ryden 2002, pg. 76).

### PLANCK SYSTEM

The Planck System is based on the universal constants  $G$ ,  $c$ , and  $\hbar$ . In this system, the natural unit of length, called the **Planck length** is

$$l_P = \left(\frac{G\hbar}{c^3}\right)^{1/2} = 1.6 \times 10^{-35} \text{ m}. \quad (156)$$

Similarly, the **Planck mass**,

$$m_P = \left(\frac{\hbar c}{G}\right)^{1/2} = 2.2 \times 10^{-8} \text{ kg}, \quad (157)$$

and **Planck time**,

$$t_P = \left(\frac{G\hbar}{c^5}\right)^{1/2} = 5.4 \times 10^{-44} \text{ s}, \quad (158)$$

can be derived. The corresponding **Planck energy** is  $E_P = m_P c^2 = 1.2 \times 10^{28}$  eV and **Planck temperature** is  $T_P = E_P/k = 1.4 \times 10^{32}$  K. When distance, mass, time, and temperature are measured in the appropriate Planck units, then  $c = k = \hbar = G = 1$  (Ryden 2002, pg. 4).

**QUESTION 19**

**Define and describe the epoch of reionization. What are the observational constraints on it?**

## QUESTION 19

**Define and describe the epoch of reionization. What are the observational constraints on it?**

### OVERVIEW OF REIONIZATION

The fact that the main sources of reionization were likely to be much less luminous than present-day galaxies, combined with their large large luminosity distance, means that the details of the reionization process are shrouded in mystery, beyond most current observational probes. The notable exceptions are observations of the polarization of the cosmic microwave background (CMB), which imply an optical depth to Thomson scattering of  $\tau \sim 0.09$  and the absence of the Gunn-Peterson trough in the spectra of distant quasars, indicating that reionization was largely complete by  $z \sim 6$ . Reionization is therefore thought to have mainly taken place over the redshift range  $z \sim 6 - 15$ . Due to the lack of more specific constraints, much of our current understanding about the epoch of reionization comes from theoretical studies in the context of the  $\Lambda$ CDM cosmology.

The picture which emerges is of small scale gaseous structures forming at  $z > 20$ , due to the collapse of DM halos at the Jeans scale, roughly  $10^4 M_\odot$ . The gas was just cool enough to fall into halos at this mass, leading to strong inhomogeneities on a scale of tens of comoving parsecs. At the same time, slightly more massive halos, with masses on the order of  $\sim 10^6 M_\odot$ , formed enough H<sub>2</sub> molecules in their cores to cool efficiently, leading to the formation of the first stars in the universe. The ionizing radiation from these stars is thought to have created substantial, yet short-lived H II regions, which were shaped by the surrounding inhomogeneous gas distribution.

Eventually, sufficiently large halos formed to trigger the formation of the first galaxies, containing tens to thousands of stars. These nascent dwarf galaxies would have created longer lived, but still isolated, H II regions. It is not yet understood how these galaxies would have grown into the much more luminous ones that have been observed at redshifts as high as  $z \sim 8$ . Nevertheless, it is widely believed that as the first galaxies grew and merged, their collective radiative output would have created a large and complex patch-work of ionized bubbles, with characteristic sizes on the order of tens to hundreds of comoving Mpc. During this time, dense systems in the IGM likely impeded the growth of the ionized bubbles. At the end of reionization, so-called “Lyman-limit” systems, dense clouds of gas optically-thick to ionizing radiation observed in the spectra of quasars at  $z < 6$ , dominated the overall opacity of the IGM to ionizing radiation, with important implications for the percolation phase of reionization and the evolution and structure of the ionizing background afterwards.

### MORE DETAILED PHYSICS

Understanding reionization is directly linked to studying the first generation of stars<sup>11</sup>. The cosmological Jeans mass can be written as

$$M_J = \frac{\pi^{5/2}}{6} \left( \frac{c_s^2}{G} \right)^{3/2} \frac{1}{\sqrt{\rho}}, \quad (159)$$

and interpreted as the minimum mass of a DM halo required for gravitational infall of gas. The Jeans mass depends on the temperature of the gas, expressed through the sound speed  $c_s$ , and on the mean cosmic matter density  $\bar{\rho}(z) = \bar{\rho}_0(1+z)^3$ . The baryon temperature has a somewhat complicated dependency on redshift. For sufficiently high redshifts, the small fraction of free electrons that remain after recombination provide a thermal coupling of the baryons to the cosmic background radiation, by means of Compton scattering. Hence, up to the **decoupling redshift**  $z_{\text{dec}} \sim 140$ , the baryon temperature simply follows that of the CMB; afterward, it adiabatically cools (Schneider 2002, pg. 383).

The Jeans criterion is a necessary condition for the formation of proto-galaxies. In order to form stars, the gas in the halos needs to be able to cool further. Here, we are dealing with the particular situation of the first galaxies, whose gas is metal-free, so metal lines cannot contribute to the cooling. This means that cooling can only happen via hydrogen and helium. Since the first excited state of hydrogen has a high energy (that of the Ly $\alpha$  transition, thus  $E \sim 10.2$  eV), this cooling is efficient only above  $T \gtrsim 10^4$  K. However, the halos which form at high redshift have low mass, so that their virial temperature is considerably below this energy scale. Therefore, atomic hydrogen is a very inefficient coolant for these first halos, insufficient to initiate the formation of stars. Furthermore, helium is of no help in this context, since its excitation temperature is even higher than that of hydrogen. Only in recent years has it been discovered that molecular hydrogen represents an extremely important component in cooling processes. Despite its very small transition probability, H<sub>2</sub> dominates the cooling rate of primordial gas at temperatures below  $T \sim 10^4$  K. By means of H<sub>2</sub>, the gas can cool in halos with a temperature exceeding about  $T_{\text{vir}} \sim 3000$  K, corresponding to a mass of  $M \gtrsim 10^4 M_\odot$ ; the exact value depending on redshift. In these halos, stars may then be able to form. However, these stars will certainly be different from those known to us, because they do not contain any metals. Therefore, the opacity of the stellar plasma is much lower. Such stars, which at the same mass presumably have a much higher temperature and luminosity (and thus a shorter lifetime), are called Pop III stars. Due to their high temperature they are much more efficient sources of ionizing photons than stars with “normal” metallicity (Schneider 2002, pg. 384).

The binding energy of H<sub>2</sub> is only 11.26 eV meaning that subsequent star formation through H<sub>2</sub> cooling will be suppressed once the most luminous stars form. Since the universe is transparent to photons with  $E_\gamma < 13.6$  eV, these photons will be able to travel relatively far distances and destroy molecular hydrogen in their near vicinities. SNe explosions from the first massive stars can also suppress subsequent star formation as they dispel gas outside of their parent halos. For gas to cool in halos without H<sub>2</sub>, their virial temperature needs to exceed  $10^4$  K; halos of this mass form with appreciable abundance at  $z \sim 10$ . In these halos, efficient star formation can then take place; the first proto-galaxies will form. These will then ionize the surrounding IGM through a

<sup>11</sup> Collisional ionization as a source of reionization can be ruled out since the IGM would not have been hot enough, as evidenced by the nearly perfect Planck spectrum of the CMB

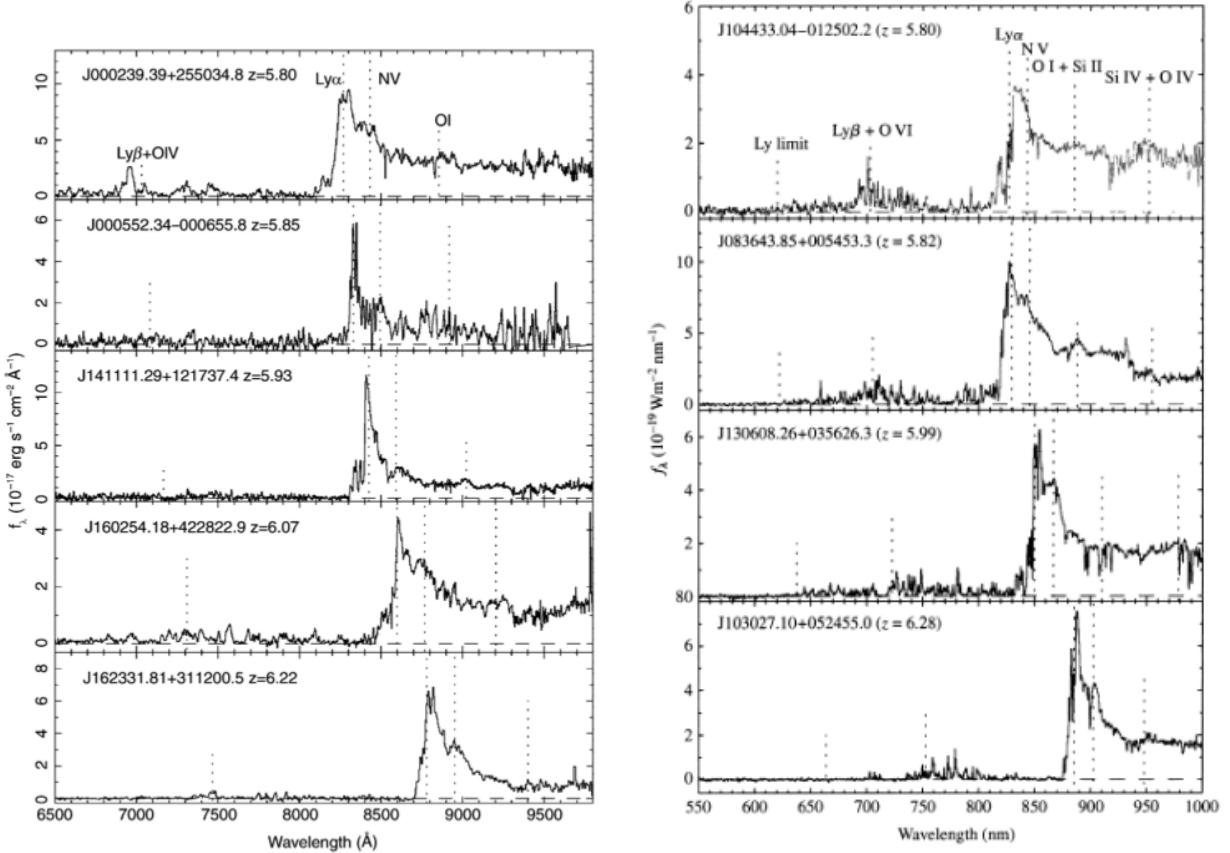


FIG. 26.— (left) Spectra of five QSOs at redshifts  $z > 5.7$ , discovered in multicolour data from the SDSS. The positions of the most important emission lines are marked. Particularly remarkable is the complete lack of flux bluewards of the Ly $\alpha$  emission line in some of the QSOs, indicating a strong Gunn-Peterson effect. However, this absorption is not complete in all QSOs, which points at strong variations in the density of neutral hydrogen in the IGM at these high redshifts. Either the hydrogen density varies strongly for different lines-of-sight, or the degree of ionization is very inhomogeneous. Image taken from (Schneider 2002). (right) The Gunn-Peterson trough observed in four high- $z$  quasars illustrates the rapid suppression of the Ly $\alpha$  forest with increasing  $z$ . This indicates the abundance of ionized hydrogen has declined significantly from  $z \sim 5$  to  $z \sim 6$  and that the universe is approaching the reionization epoch at  $z \sim 6$ . Image taken from Carroll & Ostlie (2007).

network of H II regions. We therefore conclude that reionization is a two- stage process. In a first phase, Pop III stars form through cooling of gas by molecular hydrogen, which is then destroyed by these very stars. Only in a later epoch and in more massive halos is cooling provided by atomic hydrogen, then leading to reionization (Schneider 2002, pg. 385).

#### LYMAN ALPHA FOREST

In the spectrum of any QSO a large number of absorption lines at wavelengths shorter than the Ly $\alpha$  emission line of the QSO are found. The major fraction of these absorption lines originate from the Ly $\alpha$  transition of neutral hydrogen located along the line-of-sight to the source. Since the absorption is found in the form of a line spectrum, the absorbing hydrogen cannot be distributed homogeneously. A homogeneous IGM containing neutral hydrogen would be visible in continuum absorption (Schneider 2002, pg. 331).

We first ask whether part of the baryons in the universe may be contained in a homogeneous IGM. This question can be answered by means of the **Gunn-Peterson test**. Neutral hydrogen preferentially absorbs (due to its large absorption cross-section) photons at a rest wavelength of  $\lambda = \lambda_{\text{Ly}\alpha} = 1216 \text{ Å}$  (10.2 eV). Photons from a QSO at redshift  $z_{\text{QSO}}$  attain this wavelength  $\lambda_{\text{Ly}\alpha}$  somewhere along the line-of-sight between us and the QSO if they are emitted by the quasar at  $\lambda_{\text{Ly}\alpha}(1+z_{\text{QSO}})^{-1} < \lambda < \lambda_{\text{Ly}\alpha}$ . However, if the wavelength at emission is  $\lambda > \lambda_{\text{Ly}\alpha}$ , the radiation can nowhere on its way to us be absorbed by neutral hydrogen. Hence, a jump in the observed continuum radiation should occur between the red and the blue side of the Ly $\alpha$  emission line of the QSO: this is the Gunn-Peterson effect. Such a jump in the continuum radiation of QSOs across their Ly $\alpha$  emission line has not been observed for QSOs at  $z \lesssim 5$ . At even higher redshift observations become increasingly difficult, because the Ly $\alpha$  forest then becomes so dense that hardly any continuum radiation is visible between the individual absorption lines. From the upper limit for the optical depth of absorption, one obtains bounds for the density of neutral hydrogen,  $\Omega_{\text{HI}} \sim 10^{-8}$ . From this we conclude that hardly any homogeneously distributed baryonic matter exists in the IGM or that hydrogen in the IGM is virtually fully ionized (Schneider 2002, pg. 331).

The statistical properties of these absorption lines are essentially the same for all QSOs and seem to depend only on the redshift of the Ly $\alpha$  lines, but not on  $z_{\text{em}}$ . This interpretation is confirmed by the fact that for nearly any line in the Ly $\alpha$  forest, the corresponding Ly $\beta$  line is found if the quality and the wavelength range of the observed spectra permit this. The Ly $\alpha$  forest is

further subdivided, according to the strength of the absorption, into narrow lines, Lyman-limit systems, and damped Ly $\alpha$  systems. Narrow Ly $\alpha$  lines are caused by absorbing gas of neutral hydrogen column densities of  $N_H \lesssim 10^{17} \text{ cm}^{-2}$ . Lyman-limit systems derive their name from the fact that at column densities of  $N_H \gtrsim 10^{17} \text{ cm}^{-2}$ , neutral hydrogen almost totally absorbs all radiation at  $\lambda \lesssim 912 \text{ \AA}$  (in the hydrogen rest-frame), where photons ionize hydrogen. If such a system is located at  $z_{\text{limit}}$  in the spectrum of a QSO, the spectrum at  $\lambda < (1 + z_{\text{limit}})912 \text{ \AA}$  is almost completely suppressed. Damped Ly? systems occur if the column density of neutral hydrogen is  $N_H \gtrsim 10^{20} \text{ cm}^{-2}$ . In this case, the absorption line becomes very broad due to the extended damping wings of the Voigt profile (Schneider 2002, pg. 221).

Unfortunately, Ly $\alpha$  emission cannot penetrate past regions with neutral fractions greater than about  $10^{-3}$ , meaning that it cannot be used to probe the era of reionization itself. In fact, reionization being essentially complete for a long time by  $z \sim 6$  is consistent with QSO spectra. One thing that QSO spectra can help us constrain is the evolution in the ionizing background. In general, we find that the number density of Ly $\alpha$  absorption lines at those redshifts which are only slightly smaller than the emission line redshift of the QSO itself, is lower than the mean absorption line density at this redshift (averaged over many different QSO lines-of-sight). This effect indicates that the QSO has some effect on the absorption lines, if only in its immediate vicinity; for this reason, it is named the **proximity effect**. For ionized gas in ionization equilibrium we have that

$$n_{\text{HI}} = \frac{\alpha_B}{\Gamma} n_P^2, \quad (160)$$

where  $\alpha_B$  is the case B recombination coefficient and  $\Gamma$  is the photoionization rate. This results shows that  $n_{\text{HI}}$  is inversely proportional to the number density of ionizing photons. However, the IGM in the vicinity of the QSO does not only experience the ionizing background radiation field but, in addition, the energetic radiation from the QSO itself. Therefore, the degree of ionization of hydrogen in the immediate vicinity of the QSO is higher, and consequently less Ly $\alpha$  absorption can take place there. Since the contribution of the QSO to the ionizing radiation depends on the distance of the gas from the QSO, and since the spectrum and ionizing flux of the QSO is observable, examining the proximity effect provides an estimate of the intensity of the ionizing background radiation as a function of redshift (Schneider 2002, pg. 333).

#### CMB POLARIZATION

CMB polarization measurements can be used to constrain the start of reionization. H II region electrons have a tendency to Thomson scatter incoming CMB photons. Thomson scattering also has a natural polarization axis (defined in the rest frame of the electron by the incoming and outgoing light). If in the electron rest frame the CMB were isotropic, no net polarization would be created, but if the incoming radiation were anisotropic, which is the case (as we have observed it to be), CMB radiation can become polarized. Since the CMB is naturally unpolarized, the degree of CMB polarization is directly related to the Thomson scattering optical depth and therefore the abundance of free electrons (Charles). WMAP observations of CMB polarization find a relatively large optical depth, indicating an early start to reionization at  $z \sim 15$ , since many free electrons must have been around. Specifically, this result is found by plotting the cross power spectrum between the temperature anisotropies and the degree of polarization. A surprisingly large value of the cross-power is observed for small  $l$ , meaning a large degree of polarization on large angular scales (Schneider 2002, pg. 348). Because the WMAP measurements are calculating a quantity integrated along the line of sight, it cannot easily constrain the period of time over which reionization occurred; if reionization were instant, WMAP gives  $z \sim 11$ , but if it occurred over an extended period of time from  $z \sim 7$  onward, reionization could have begun closer to  $z \sim 15$  (Charles).

#### 21-CM OBSERVATIONS

21-cm line emission and absorption from H I could prove an invaluable tool to studying reionization. In the late ( $z \lesssim 9$ ) reionization epoch, the 21-cm line intensity is simply proportional to the ionized fraction, and therefore it is straightforward to translate fluctuations in 21-cm emission to the progression of reionization. While QSO observations are restricted to  $z \lesssim 6$  and CMB observations are integrated, 21-cm line emission and absorption could probe to much higher redshifts while maintaining fidelity in both time and space (Charles).

#### HELIUM REIONIZATION

He II has an ionizing potential energy of 54.4 eV, and the soft emission from Pop III stars is unable to ionize it; in contrast, Pop III stars can easily ionize He I, with an ionizing potential of 24.6 eV. As a result He only fully ionizes when the quasar population is large enough for large numbers of hard quasar-emitted photons to percolate the universe, which occurs at  $z \sim 3$ . The most significant observation evidence for reionization completion at  $z \sim 3$  comes from far-UV studies of the He Ly $\alpha$  forest along lines of sight to bright quasars, which show a significant increase in He II optical depth near  $z \sim 3$  (Charles).

**QUESTION 1**

**Sketch out the Hubble sequence. What physical trends are captured by the classification system.**

## QUESTION 1

**Sketch out the Hubble sequence. What physical trends are captured by the classification system?**

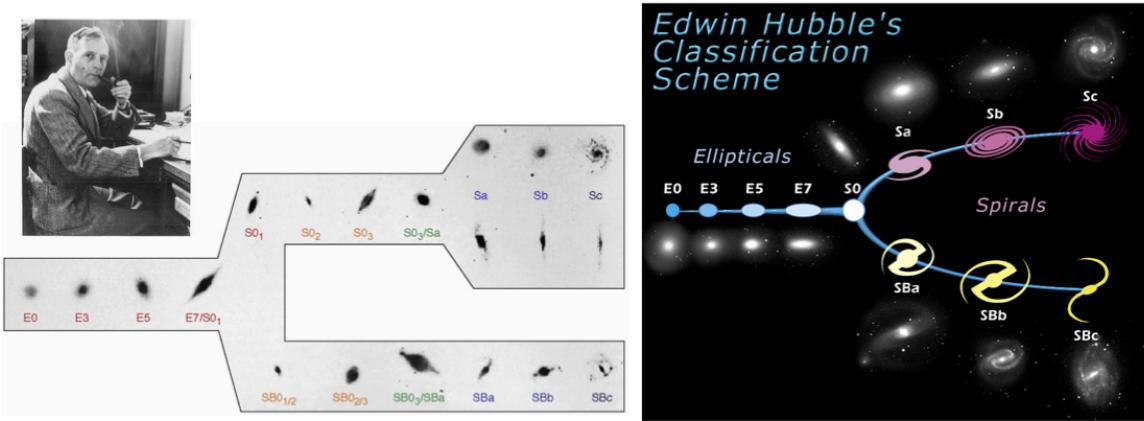


FIG. 27.— The Hubble sequence of galaxy classification. Image taken from Etsuko’s notes.

Hubble produced a morphological classification scheme of galaxies based on optical observations; this is known as the **Hubble sequence** and is shown in Figure 27. According to this scheme, three main types of galaxies exist (Schneider 2002, pg. 88):

- **Elliptical Galaxies:** These have nearly elliptical isophotes<sup>12</sup> without any clearly defined structure. They are subdivided according to their ellipticity  $\epsilon = 1 - b/a$ , where  $a$  and  $b$  denote the semimajor and the semiminor axes, respectively. Ellipticals are found over a relatively broad range in ellipticity,  $0 \leq \epsilon \lesssim 0.7$ . The notation  $E_n$  is commonly used to classify the ellipticals with respect to  $\epsilon$ , with  $n = 10\epsilon$  (e.g., an  $E4$  galaxy has an axis ratio of  $b/a = 0.6$ , and  $E0$ 's have circular isophotes).
- **Spiral Galaxies:** These consist of a disk with spiral arm structure and a central bulge. They are divided into two subclasses: normal spirals (S's) and barred spirals (SB's). In each of these subclasses, a sequence is defined that is ordered according to the brightness ratio of bulge and disk, and that is denoted by a, ab, b, bc, c, cd, d. Objects along this sequence are often referred to as being either an early-type or a late-type; hence, an  $Sa$  galaxy is an early-type spiral, and an  $SBc$  galaxy is a late-type barred spiral. We stress explicitly that this nomenclature is not a statement of the evolutionary stage of the objects but is merely a nomenclature of purely historical origin.
- **Irregular Galaxies:** These are galaxies with only weak (Irr I) or no (Irr II) regular structure. The classification of Irr's is often refined. In particular, the sequence of spirals is extended to the classes Sdm, Sm, Im, and Ir (m stands for Magellanic; the Large Magellanic Cloud is of type SBm).
- **S0 (Lenticular) Galaxies:** These are a transition between ellipticals and spirals. They are also called lenticulars as they are lens-shaped galaxies which are likewise subdivided into S0 and SB0, depending on whether or not they show a bar. They contain a bulge and a large enveloping region of relatively unstructured brightness which often appears like a disk without spiral arms. Ellipticals and S0 galaxies are referred to as early-type galaxies, spirals as late-type galaxies. As before, these names are only historical and are not meant to describe an evolutionary track.

Besides morphological criteria, colour indices, spectroscopic parameters (based on emission or absorption lines), the broadband spectral distribution (galaxies with/without radio- and/or X-ray emission), as well as other features may also be used for galaxy classification (Schneider 2002, pg. 88). For instance, along the the Hubble sequence galaxies are correlated by colour, with early-type galaxies appearing redder and late-type galaxies appearing bluer; see Figure 28. This is likely related to their present star formation rates and star formation histories (Bob AST2040). Table 2 summarizes the most notable differences between the three main galaxies in the Hubble sequence using information form Schneider (2002).

TABLE 2  
COMPARISON OF HUBBLE TYPES

Class	$M_B$	Mass [ $M_{\odot}$ ]	Diameter [kpc]
Ellipticals	-8 to -25	$10^7$ to $10^{14}$	10 to 1000
Spirals	-16 to -23	$10^9$ to $10^{12}$	5 to 100
Irregulars	-13 to -18	$10^8$ to $10^{10}$	0.5 to 50

<sup>12</sup> Isophotes are contours along which the surface brightness of a sources is constant.

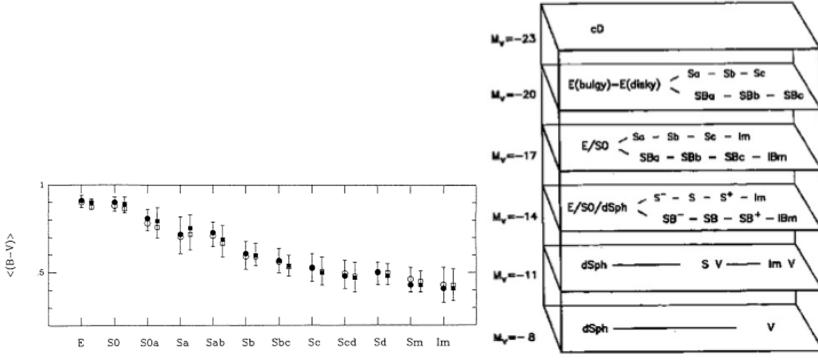


FIG. 28.— (left) Colour versus morphological galaxy type displaying a clear trend along the Hubble sequence. (right) The Hubble sequence with B-band luminosity being used as a third dimension. Images taken from Bob's AST2040 Lecture Notes.

### ELLIPTICAL GALAXIES

Ellipticals span a broad class of galaxies which differ in their luminosities and sizes:

- **Normal Ellipticals:** This class includes giant ellipticals (gE's), those of intermediate luminosity (E's), and compact ellipticals (cE's), covering a range in absolute magnitudes from  $M_B \sim -23$  to  $M_B \sim -15$ .
- **Dwarf Ellipticals:** These differ from the cE's in that they have a significantly smaller surface brightness and a lower metallicity.
- **cD Galaxies:** These are extremely luminous (up to  $M_B \sim -25$ ) and large (up to  $R \lesssim 1$  Mpc) galaxies that are only found near the centres of dense clusters of galaxies. Their surface brightness is very high close to the centre, they have an extended diffuse envelope, and they have a very high M/L ratio.
- **Blue compact dwarf galaxies:** These “blue compact dwarfs” (BCD's) are clearly bluer than the other ellipticals, and contain an appreciable amount of gas in comparison.
- **Dwarf spheroidals:** These exhibit a very low luminosity and surface brightness. They have been observed down to  $M_B \sim -8$ . Due to these properties, they have thus far only been observed in the Local Group.

Normal ellipticals and cD's follow a de Vaucouleurs profile which describes the surface brightness  $I$  as a function of the distance  $R$  from the centre of the galaxy:

$$\log \left( \frac{I(R)}{I_e} \right) = -3.33 \left[ \left( \frac{R}{R_e} \right)^{1/4} - 1 \right]. \quad (161)$$

The effective radius  $R_e$  is strongly correlated with the absolute magnitude  $M_B$  (larger are brighter), with relatively little scatter (Schneider 2002, pg. 90).

Except for the BCD's, elliptical galaxies appear red when observed in the optical, which suggests an old stellar population. It was once believed that ellipticals contain neither gas nor dust, but these components have now been found, though at a much lower mass-fraction than in spirals. The metallicity of ellipticals and S0 galaxies increases towards the galaxy centre, as derived from colour gradients. Also in S0 galaxies the bulge appears redder than the disk (Schneider 2002, pg. 92).

Analyzing the morphology of elliptical galaxies raises the question as to why they are not round in shape. A simple explanation would be rotational flattening (i.e., as in a rotating self-gravitating gas ball, the stellar distribution bulges outwards at the equator due to centrifugal forces, as is also the case for the Earth). If this explanation were correct, the rotational velocity  $v_{\text{rot}}$ , which is measurable in the relative Doppler shift of absorption lines, would have to be of about the same magnitude as the velocity dispersion of the stars  $\sigma_v$  that is measurable through the Doppler broadening of lines. However, for luminous ellipticals one finds that, in general,  $v_{\text{rot}} \ll \sigma_v$ , so that rotation cannot be the major cause of their ellipticity. In addition, many ellipticals are presumably triaxial, so that no unambiguous rotation axis is defined. Thus, luminous ellipticals are in general *not* rotationally flattened. For less luminous ellipticals and for the bulges of disk galaxies, however, rotational flattening can play an important role (Schneider 2002, pg. 93). Instead, their shapes are caused by the anisotropic velocity dispersion (i.e., their motion is not random, but rather follows some preferential direction) of stars in the galaxies; in this sense, they are said to be **pressure-supported** (Carroll & Ostlie 2007, pg. 988).

The isophotes of many of the normal elliptical galaxies are well approximated by ellipses. These elliptical isophotes with different surface brightnesses are concentric to high accuracy. However, in many cases the ellipticity varies with radius, so that the value for  $\epsilon$  is not a constant. In addition, many ellipticals show a so-called isophote twist: the orientation of the semi-major axis of the isophotes changes with the radius. This indicates that elliptical galaxies are not spheroidal, but triaxial systems (Schneider 2002, pg. 95). Often the isophotes are described in terms of the degree of **boxiness** or **diskiness** that their isophotal surfaces exhibit. Disky galaxies tend to be rotationally supported while boxy galaxies are largely pressure-supported (Carroll & Ostlie 2007, pg. 990).

### THE FUNDAMENTAL PLANE

A relation for elliptical galaxies, analogous to the Tully-Fisher relation (discussed in another question), is the **Faber-Jackson Relation**. This is based on the observational discovery that the velocity dispersion in the centre of elliptical galaxies,  $\sigma_0$ , scales with the luminosity:

$$L \propto \sigma_0^4. \quad (162)$$

This scaling relation can be derived under the same assumptions as the Tully-Fisher relation. However, the dispersion of ellipticals about this relation is larger than that of spirals about the Tully-Fisher relation (Schneider 2002, pg. 107).

It is thus common to speak in terms of the **fundamental plane** instead, which has a smaller dispersion amongst ellipticals. This relation can be derived by noting that a relation between the surface brightness and the effective radius exists for ellipticals:

$$R_e \propto \langle I \rangle_e^{-0.83}, \quad (163)$$

where  $\langle I \rangle_e$  is the average surface brightness within the effective radius, so that

$$L = 2\pi R_e^2 \langle I \rangle_e. \quad (164)$$

The form of equation (164) and the relation in equation (163) restate the observation we made before that larger ellipticals are brighter. We can use these two equations to show that

$$\langle I \rangle_e \propto L^{-1.5}. \quad (165)$$

Hence, more luminous ellipticals have smaller surface brightnesses. By means of the Faber-Jackson relation in equation (162),  $L$  is related to  $\sigma_0$ , and therefore  $\sigma_0$ ,  $\langle I \rangle_e$ , and  $R_e$  are all related to each other. The distribution of elliptical galaxies in such a three-dimensional parameter space is located close to a plane defined by

$$R_e \propto \sigma_0^{1.4} \langle I \rangle_e^{-0.85}. \quad (166)$$

This is what is referred to as the fundamental plane (Schneider 2002, pg. 107).

We can explain this result by noting that the mass within  $R_e$  can be derived from the virial theorem,  $M \propto \sigma_0^2 R_e$ . Combining this with equation (164) yields

$$R_e \propto \frac{L}{M} \frac{\sigma_0^2}{\langle I \rangle_e}, \quad (167)$$

which agrees with the fundamental plane in the form of equation (166) provided that

$$\left( \frac{M}{L} \right) \propto M^{0.2} \Leftrightarrow \left( \frac{M}{L} \right) \propto L^{0.25}. \quad (168)$$

Hence, the fundamental plane follows from the virial theorem if the mass-to-light ratio of ellipticals increases slightly with mass (Schneider 2002, pg. 108).

Another scaling relation for ellipticals which is of substantial importance in practical applications is the  $D_n - \sigma$  relation.  $D_n$  is defined as that diameter of an ellipse within which the average surface brightness  $I_n$  corresponds to a value of 20.75 mag arcsec $^{-2}$  in the B-band. After assuming a self-similar brightness profile and integrating over the de Vaucouleurs profile we find that

$$D_n \propto \sigma_0^{1.4} I_e^{0.05}. \quad (169)$$

This implies that  $D_n$  is nearly independent of  $I_e$  and only depends on  $\sigma_0$ . The  $D_n - \sigma$  relation in equation (169) describes the properties of ellipticals considerably better than the Faber-Jackson relation and, in contrast to the fundamental plane, it is a relation between only two observables (Schneider 2002, pg. 109).

### SPIRAL GALAXIES

Looking at the sequence of early-type spirals (i.e., Sa's or SBA's) to late-type spirals, we find a number of differences that can be used for classification purposes:

- **Bulge-to-Disk Luminosity:** A decreasing luminosity ratio of bulge and disk is observed with  $L_{\text{bulge}}/L_{\text{disk}} \sim 0.3$  for Sa's and  $\sim 0.05$  for Sc's (i.e., the bulge becomes *dimmer* as we move from early-type to late-type spirals).
- **Spiral Arm Winding:** An increasing opening angle of spiral arms is observed with  $\sim 6^\circ$  for Sa's and  $\sim 18^\circ$  for Sc's (i.e., the spiral arms become *less* tightly wound as we move from early-type to late-type spirals).
- **Spiral Arm Smoothness:** An increasing brightness structure along the spirals: Sa's have a *smooth* distribution of stars along the spiral arms, whereas the light distribution in the spiral arms of Sc's is resolved into bright knots of stars and H II regions (i.e., the spiral arms become more *clustered* as we move from early-type to late-type spirals).

Compared to ellipticals, the spirals cover a distinctly smaller range in absolute magnitude and mass; see Table 2. Bars are common in spiral galaxies, with  $\sim 70\%$  of all disk galaxies containing a large-scale stellar bar. Such a bar perturbs the axial symmetry of the gravitational potential in a galaxy, which may have a number of consequences. One of them is that this perturbation can lead to a redistribution of angular momentum of the stars, gas, and dark matter. In addition, by perturbing the orbits, gas can be

driven towards the centre of the galaxy which may have important consequences for triggering nuclear activity. The light profile of the bulge of spirals is described by a de Vaucouleurs profile to a good approximation while the disk follows an exponential brightness profile. Observationally, it has been determined that the central surface brightness of spirals has a very low spread (Schneider 2002, pg. 98).

The colour of spiral galaxies depends on their Hubble type, with later types being bluer. This means that the fraction of massive young stars increases along the Hubble sequence towards later spiral types. This conclusion is also in agreement with the findings for the light distribution along spiral arms where we clearly observe active star-formation regions in the bright knots in the spiral arms of Sc's. Furthermore, this colour sequence is also in agreement with the decreasing bulge fraction towards later types. The formation of stars requires gas, and the mass fraction of gas is larger for later types, as can be measured, for instance, from the 21-cm emission of H I from H $\alpha$  within H II regions, and from CO emission from cool molecular clouds. Dust, in combination with hot stars, is the main source of (FIR) emission from galaxies. Sc galaxies emit a larger fraction of FIR radiation than Sa's, and barred spirals have stronger FIR emission than normal spirals. The FIR emission arises due to dust being heated by the UV radiation of hot stars and then reradiating this energy in the form of thermal emission. A prominent colour gradient is observed in spirals: they are red in the centre and bluer in the outer regions. We can identify at least two reasons for this trend. The first is a metallicity effect, as the metallicity is increasing inwards and metal-rich stars are redder than metal-poor ones, due to their higher opacity. Second, the colour gradient can be explained by star formation. Since the gas fraction in the bulge is lower than in the disk, less star formation takes place in the bulge, resulting in a stellar population that is older and redder in general. Furthermore, it is found that the metallicity of spirals increases with luminosity (Schneider 2002, pg. 102).

The spiral arms are the bluest regions in spirals and they contain young stars and H II regions. For this reason, the brightness contrast of spiral arms increases as the wavelength of the (optical) observation decreases. In particular, the spiral structure is very prominent in a blue filter. It is suspected that spiral arms are a wave structure, the velocity of which does not coincide with the physical velocity of the stars. Spiral arms are quasi-stationary density waves, regions of higher density (possibly 10–20% higher than the local disk environment). If the gas, on its orbit around the centre of the galaxy, enters a region of higher density, it is compressed, and this compression of molecular clouds results in an enhanced star-formation rate. This accounts for the blue colour of spiral arms (Schneider 2002, pg. 103).

Hot gas resulting from the evolution of supernova remnants may expand out of the disk and thereby be ejected to form a gaseous halo of a spiral galaxy. Such a halo of hot ( $T \sim 10^6$  K) coronal gas has been identified outside the disk of spiral galaxies through their X-ray emission (Schneider 2002, pg. 104).

#### OTHER TYPES OF GALAXIES

The light from “normal” galaxies is emitted mainly by stars. Therefore, the spectral distribution of the radiation from such galaxies is in principle a superposition of the spectra of their stellar population. The spectrum of stars is, to a first approximation, described by a Planck function that depends only on the star's surface temperature. A typical stellar population covers a temperature range from a few thousand Kelvin up to a few tens of thousand Kelvin. Since the Planck function has a well-localized maximum and from there steeply declines to both sides, most of the energy of such “normal” galaxies is emitted in a relatively narrow frequency interval that is located in the optical and NIR sections of the spectrum (Schneider 2002, pg. 89).

In addition to these, other galaxies exist whose spectral distribution cannot be described by a superposition of stellar spectra. One example is the class of **active galaxies** which generate a significant fraction of their luminosity from gravitational energy that is released in the infall of matter onto a supermassive black hole. The activity of such objects can be recognized in various ways. For example, some of them are very luminous in the radio and/or in the X-ray portion of the spectrum, or they show strong emission lines with a width of several thousand km/s if the line width is interpreted as due to Doppler broadening. In many cases, by far the largest fraction of luminosity is produced in a very small central region: the **active galactic nucleus** (AGN) that gave this class of galaxies its name. In quasars, the central luminosity can be of the order of  $\sim 10^{13} L_\odot$ , about a thousand times as luminous as the total luminosity of our MW (Schneider 2002, pg. 89).

Another type of galaxy also has spectral properties that differ significantly from those of “normal” galaxies, namely the **starburst galaxies**. Normal spiral galaxies like our MW form new stars at a star-formation rate of  $\sim 3 M_\odot \text{ yr}^{-1}$  which can be derived, for instance, from the Balmer lines of hydrogen generated in the H II regions around young, hot stars. By contrast, elliptical galaxies show only marginal star formation or none at all. However, there are galaxies which have a much higher star-formation rate, reaching values of  $\sim 100 M_\odot \text{ yr}^{-1}$  and more. If many young stars are formed we would expect these starburst galaxies to radiate strongly in the blue or in the UV part of the spectrum, corresponding to the maximum of the Planck function for the most massive and most luminous stars. This expectation is not fully met though: star formation takes place in the interior of dense molecular clouds which often also contain large amounts of dust. If the major part of star formation is hidden from our direct view by layers of absorbing dust, these galaxies will not be very prominent in blue light. However, the strong radiation from the young, luminous stars heats the dust; the absorbed stellar light is then emitted in the form of thermal dust emission in the infrared and submillimeter regions of the electromagnetic spectrum – these galaxies can thus be extremely luminous in the IR. They are called **ultra-luminous infrared galaxies (ULIRGs)** (Schneider 2002, pg. 90).

**QUESTION 2**

**What is the total mass (in both dark matter and in stars) of the Milky Way galaxy? How does this compare to M31 and to the LMC? How is this mass determined?**

## QUESTION 2

**What is the total mass (in both dark matter and in stars) of the Milky Way galaxy? How does this compare to M31 and to the LMC? How is this mass determined?**

The masses of the MW and Andromeda galaxy (M31) have been measured by a variety of methods, but often with conflicting results that have led to a debate around which galaxy is more massive. Mass judging criteria based on observations of the surface brightness of the stellar halo, the number of globular clusters (which correlates with total mass albeit with scatter), and the amplitude of the inner gas rotation curve suggest that M31 is more massive. On the other hand, if the mass estimate is based on criteria such as the velocities of satellite galaxies, distant globular clusters, or tidal radii of nearby dwarf spheroidals, then the MW appears more massive. The current consensus, however, is that the two galaxies are roughly of the same mass ( $\sim 10^{12} M_\odot$ ), with M31 probably the slightly more massive of the two, though this is based on the rather indirect mass estimates described above for M31 (Watkins et al. 2010).

On the other hand, the masses of the two galaxies are reasonably well constrained within the first few kpc from knowledge of their gas rotation curves via 21-cm radio observations (Carroll & Ostlie 2007, pg. 914). Of course, this only samples the inner regions of the galaxies, and in order to probe further out into the vast dark matter halos it is necessary to resort to satellite kinematics. Unfortunately, the uncertainties in such techniques are plagued by low sample sizes as well as the fact that there is seldom knowledge of the proper motion of the satellites to complement their observed radial velocity and distance data. Because of the latter, assumptions must be made on the eccentricities of the satellite orbits thereby affecting the mass determination; see equation (170) below.

In order to convert satellite kinematics into mass estimates we begin by analyzing the virial theorem for a spherically symmetric collection of  $N$  test particles (e.g. planetary nebulae, stars, globular clusters, satellite galaxies) orbiting a point mass  $M$ . For this situation the virial theorem dictates that

$$GM = \frac{\langle v^2 \rangle}{\langle 1/r \rangle}, \quad (170)$$

where angular brackets denote average values. If the distribution of test particles is spherically symmetric then  $\langle v^2 \rangle = 3\langle v_r^2 \rangle$  and  $\langle 1/r \rangle = 2/\pi\langle 1/R \rangle$ , where  $v_r$  is the observed radial velocity and  $R$  the projected separation. Substituting these identities into equation (170) for a collection of  $N$  test particles yields a mass estimate of the form

$$M = \frac{3\pi}{2G} \frac{\sum_i v_r^2}{\sum_i 1/R_i}. \quad (171)$$

Despite its easy appearance, the virial theorem does not provide accurate mass estimates. There are many problems associated with its use including its failure to converge as  $N \rightarrow \infty$  (Bahcall & Tremaine 1981).

Instead of using the virial theorem, a more reliable mass estimate is based on the projected mass  $q \equiv v_r^2 R/G$ . The variable  $q$  has dimensions of mass and with a suitable multiplicative factor can be used as an estimator to the mass  $M$ . For a general distribution of test particles it turns out that the expectation value of  $q$  is

$$\langle q \rangle = \frac{\pi M}{32} (3 - 2\langle e^2 \rangle), \quad (172)$$

where  $\langle e^2 \rangle$  is the expectation value of the square of the eccentricities of the particles orbits. Using this relation and the definition of  $q$  it is straightforward to arrive at a mass estimate of the form

$$M = \frac{C}{G N} \sum_i v_r^2 R_i, \quad (173)$$

where  $C$  is a constant of order unity depending on the test particles' eccentricities. Unlike the virial theorem method, the central value theorem applies to this method and guarantees that the sum will converge to the true mass  $M$  with an error proportional to  $1/\sqrt{N}$  (Bahcall & Tremaine 1981).

Applying a modified form of equation (173) to 26 satellite galaxies of the Milky Way and 23 satellite galaxies of M31, Watkins et al. (2010) determine the masses of the two galaxies within 300 kpc from their centres to be  $M_{\text{MW}} \sim 3 \times 10^{12} M_\odot$  and  $M_{\text{M31}} \sim 1 \times 10^{12} M_\odot$ . These values are rather volatile inasmuch as the exclusion of the satellite galaxies with ambiguous velocity and distance measures changes the mass estimates by a factor of roughly 2.

The value for  $M_{\text{MW}}$  is in good agreement with the study by Xue et al. (2008) in which 2401 blue horizontal branch (BHB) stars (which have high luminosities and nearly constant absolute magnitudes within a restricted colour range) from the SDSS are used to constrain the MW's circular velocity curve up to 60 kpc. From this the total mass within 60 kpc is determined and subsequently used to estimate the mass of the entire halo to be  $M_{\text{MW}} \sim 1 \times 10^{12} M_\odot$ . Of course, Newton's theorem asserts that any mass outside of the limiting radius of 60 kpc will have no observational effect in a spherical or elliptical system and so estimating the halo mass in this way requires an initial assumption on the structure of the dark matter halo. Indeed, the estimate by Xue et al. (2008) is based on the assumption of an NFW halo profile.

Schommer et al. (1992) measure the velocities of individual stars in 83 star clusters in the LMC to arrive at a mass estimate for the galaxy. Using equation (173) they find the mass of the LMC to be  $M_{\text{LMC}} \sim 2 \times 10^{10} M_\odot$ , roughly 1/100 that of the Milky Way. They compare this to an estimate based on a rotation curve constructed from their cluster rotation data in addition to earlier

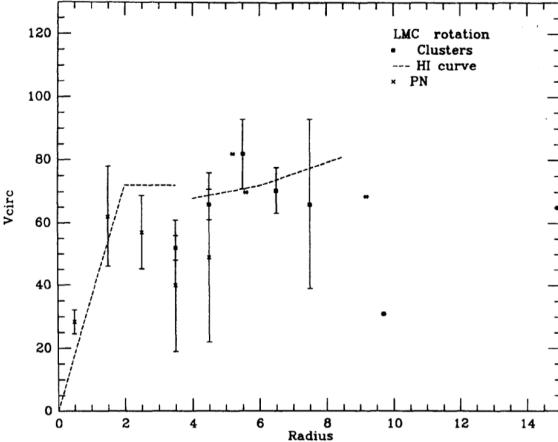


FIG. 29.— The rotation curve of the LMC built from rotational data of star clusters, H I, and PN. The dashed line shows the H I rotation curve whereas the points represent data from the clusters and PN. The error bars are the standard deviations of the mean of the velocities in the given bin while points with no error bars denote single objects. The bottom axis is in units of degrees and can be converted to physical lengths by noting that the LMC is 48 kpc from the MW. Image taken from Schommer et al. (1992).

data on the circular velocities of H I and planetary nebulae (PN) in the LMC. This is shown in Figure 29. To estimate the mass of the LMC from the rotation curve, one needs to compare the the circular velocity of the MW and the LMC at a distance of 8.5 kpc, the distance from the Sun to the centre of the MW. For the MW this value is roughly  $200 \text{ km s}^{-1}$  while for the LMC it is roughly  $20 \text{ km s}^{-1}$  since the circular velocity is roughly 20% larger than the rotation velocity (Weinberg 2000). Then since  $M \propto Rv^2$  we have that  $M_{\text{LMC}} \sim 1 \times 10^{10} M_{\odot}$ , in good agreement with the value above. This estimate would obviously be improved with more measurements of mass tracers at large radii. In principle, this can be achieved by H I though at such large radii its kinematics may be disturbed by hydrodynamical processes from tidal interactions (i.e. the Magellanic Stream).

Another quick check on the mass of the LMC is to investigate its tidal interactions with the MW. From the point of view of the MW, the LMC is an oversize globular cluster. Its tidal radius is measurable and depends both on the MW rotation curve and the LMC mass (and also weakly on the LMC mass profile). The tidal radius of the LMC can be estimated by observing the extent of its stellar halo. From this its mass is estimated via

$$M_{\text{LMC}} = 2 \left( \frac{r_t}{R_{\text{LMC}}} \right)^3 M_{\text{MW}}, \quad (174)$$

where  $r_t$  is the tidal radius and  $R_{\text{LMC}}$  is the distance to the LMC. This analysis is considered by Weinberg (2000) in which a tidal radius of 10.8 kpc is used to constrain the mass of the LMC at  $M_{\text{LMC}} \sim 2 \times 10^{10} M_{\odot}$ .

**QUESTION 3**

**How do we know that the intergalactic medium is ionized?**

### QUESTION 3

How do we know that the intergalactic medium is ionized?

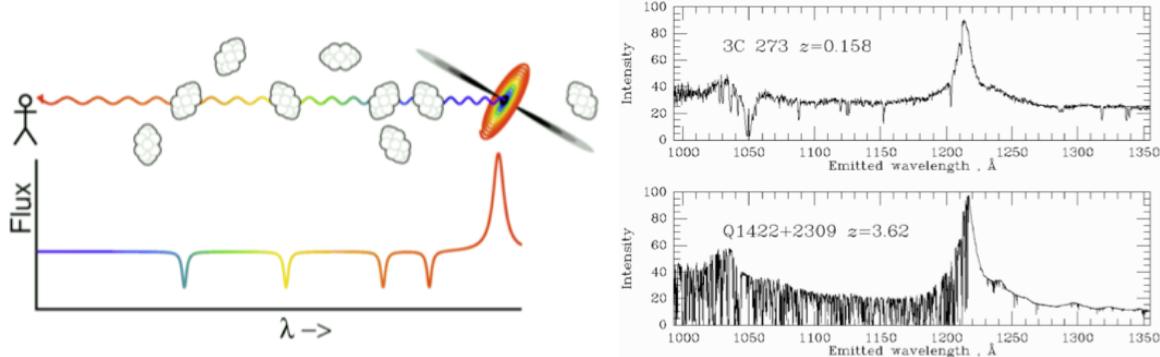


FIG. 30.— (left) Cartoon depicting the absorption of redshifted Ly $\alpha$  photons by intervening neutral hydrogen pockets. Image taken from Laura Parker's AST3X03 Lecture Notes from McMaster University. (right) A comparison of the Ly $\alpha$  forest for a nearby ( $z = 0.2$ ) and distant ( $z = 3.6$ ) quasar, showing that the more distant quasar has a thicker set of absorption lines. Image from Yevgeni's notes.

The spectra of high-redshift quasars always display a large number of narrow absorption lines superimposed on the quasar's continuous spectrum; these lines are in addition to any broad absorption lines that are associated with the quasar itself. These narrow lines are formed when the light from a quasar passes through material (e.g., an interstellar cloud or a galactic halo) that happens to lie along the line of sight. If the absorbing material is far from Earth then the expansion of the universe will cause these absorption lines to be strongly redshifted. Furthermore, if the light passes through more than one cloud or galactic halo during its trip to Earth, different sets of absorption lines will be seen. Each set of lines corresponds to the redshift of a particular cloud or halo. There are two classes of narrow absorption lines in quasar spectra:

- The **Ly $\alpha$  forest** is a dense thicket of hydrogen absorption lines. These lines are believed to be formed in intergalactic clouds and display a variety of redshifts. Absorption by primordial ionized helium (He II) has also been detected.
- Lines are also formed by **ionized metals**, primarily C and Mg, together with Si, Fe, Al, N, and O. The mix of elements is similar to that found in the interstellar medium of the MW, indicating that the material has been processed through stars and enriched in heavy elements. These lines are thought to be formed in the extended halos or disks of galaxies found along the line of sight to the quasar.

Most of these lines normally occur at UV wavelengths in their rest frame (Carroll & Ostlie 2007, pg. 1137).

Observations of QSO at high redshifts, show what is known as the Ly $\alpha$  forest; see Figure 30. Neutral hydrogen absorbs photons of energy roughly 10.2 eV, which corresponds to the electron transition from the ground states into the first energy level (the Ly $\alpha$  line at 1216 Å). If the universe was filled with a neutral IGM, then as photon of energies higher than Ly $\alpha$  in the rest frame of the QSO were redshifted into this line, they would be absorbed, as the photon propagated toward us. The resulting spectrum would show continuous absorption. However, the presence of the Ly $\alpha$  forest, which is a collection of absorption lines at wavelengths shorter than the rest Ly $\alpha$  line of the QSO, tells us that the light from the QSO encountered clumpy patches of neutral hydrogen (Yevgeni). The optical depth for Ly $\alpha$  absorption in models with  $\Omega_\Lambda$  is

$$\tau = 4 \times 10^{10} h^{-1} \frac{n_{\text{HI}}(z)/\text{cm}^{-3}}{(1+z)\sqrt{1+\Omega_{\text{mat}}z}}, \quad (175)$$

where  $n_{\text{HI}}(z)$  is the density of neutral hydrogen at the absorption redshift  $z$ . Computing the effective optical depth from the flux decrement in the absorption line allows us to solve equation (175) for the neutral hydrogen number density. Since the Ly $\alpha$  absorption cross-section is so large, only a small amount of neutral hydrogen is required to create a substantial  $\tau$ , and the result for actual quasars suggests

$$n_{\text{HI}} \lesssim 10^{-13} \text{ cm}^{-3} \Leftrightarrow \Omega_{\text{HI}} \lesssim 10^{-8}. \quad (176)$$

From this we conclude that hardly any homogeneously distributed baryonic matter exists in the IGM, or that hydrogen in the intergalactic medium is virtually fully ionized. However, from primordial nucleosynthesis we know the average density of hydrogen – it is *much* higher than the above limits – so that hydrogen must be present in essentially fully ionized form (Schneider 2002, pg. 331).

We deduce the size of the intergalactic clouds by comparing the Ly $\alpha$  forest in the spectra of pairs of lensed quasars. Many of the absorption lines are seen in both spectra, but some are not. This indicates that the clouds are, on average, about the size of the lensing galaxy. From the total calculated column density of hydrogen (ionized plus neutral), the mass of atypical cloud is somewhere around  $10^7 M_\odot$ . At the temperature estimated for a typical cloud ( $T \sim 3 \times 10^4$  K), its self-gravity would be too weak

to keep it from dispersing. It may be held together by the pressure of ales dense (but hotter) external IGM or by the presence of DM within the cloud (Carroll & Ostlie 2007, pg. 1139).

The comoving space density of intergalactic clouds appears to have been greater in the past than it is today, so the number of clouds has been decreasing as the universe ages. A statistical analysis of the clouds' redshifts reveals little evidence that the clouds tend to be grouped in clusters. Instead, they appear to be distributed randomly throughout space. In particular, there do not appear to be large voids in the distribution of these intergalactic clouds similar to those we observe for the large-scale clustering of galaxies; the significance of this is not yet clear (Carroll & Ostlie 2007, pg. 1139).

#### ABSORPTION SYSTEMS

The hydrogen between galaxies was reionized more than 12.5 Gyr ago. After this event, a largely uniform H I-ionizing background pervaded the IGM and kept the hydrogen highly ionized everywhere except within rare, overdense pockets. The amplitude of this background appears to have declined quickly above  $z \sim 6$  and to have stayed relatively constant over  $2 < z < 4$  (McQuinn et al. 2011). The observed high transmission of  $z \sim 3$  quasars at rest wavelengths blueward of H I Ly $\alpha$  reveals that the IGM is highly ionized. The presence of the Ly $\alpha$  forest demands an intense, extragalactic ultraviolet background (EUVB) radiation field. The quasars themselves provide a significant fraction of the required ionizing flux, buoyed by the emission from more numerous yet fainter star-forming galaxies. Several recent studies have argued that the latter population dominates the EUVB at  $z \gtrsim 3$ , where the quasar population likely declines (Prochaska et al. 2009).

Often the H I absorption system seen the spectra of QSOs are placed in three general categories dependent on the H I column density ( $N_{\text{HI}}$ ) of the absorber. The low column density Ly $\alpha$  forest absorbers ( $N_{\text{HI}} < 10^{16} \text{ cm}^{-2}$ ) are associated with the diffuse IGM. These systems probe low-density, highly ionized gas and are thought to trace the dark matter distribution throughout the IGM as well as contain the bulk of the baryons at high redshift and a significant amount of the baryons even today. At the other end, the high column density damped Ly $\alpha$  absorbers (DLAs,  $N_{\text{HI}} > 10^{20} \text{ cm}^{-2}$ ) appear associated with the main bodies of galaxies. These high-density, predominantly neutral systems serve as neutral gas reservoirs for high redshift star formation. The intermediate column density systems – known as Lyman Limit Systems – mark the transition from the optically thin Ly $\alpha$  forest to the optically thick absorbers found in and around the extended regions of galaxies. Typically these absorbers are easy to identify in QSO spectra due to the characteristic attenuation of QSO flux by the Lyman limit at  $\sim 912 \text{ \AA}$  in the rest frame (Ribaudo et al. 2011). In addition, they are optically thick enough to be harbouring neutral hydrogen cores.

**QUESTION 4**

**Describe as many steps of the distance ladder and the involved techniques as you can. What are the rough distances to the Magellanic Clouds, Andromeda, and the Virgo Cluster?**

#### QUESTION 4

**Describe as many steps of the distance ladder and the involved techniques as you can. What are the rough distances to the Magellanic Clouds, Andromeda, and the Virgo Cluster?**

##### *DISTANCE LADDER*

First we will provide an overview of methods used to determine distances within the MW. Unless otherwise cited, information from the following points is taken from Charles' notes and from Schneider (2002, pgs. 36-44).

- **Trigonometric Parallax:** This method is based on a purely geometric effect and is therefore independent of any physical assumptions. This involves a measurement of the position of a distant object from two different observation points separated by some physical distance called a “baseline”. The baseline divided by the angle by which the objects shifts when moving between the two points of observation gives a distance estimate. The technique is usually only applicable to nearby stars, but can be used to infinity depending on the photometric/pointing accuracy of a telescope and length of the baseline. Commonly the baseline used is the diameter of the Earth’s orbit around the Sun in which case we can feasibly determine the distance to all stars within the MW. The proper motion of the Sun can be used to further extend this baseline; a method known as secular parallax (Carroll & Ostlie 2007, pg. 1039).
- **Moving Cluster Parallax:** The stars in an (open) star cluster all have a very similar spatial velocity. This implies that their proper motion<sup>13</sup> vectors should be similar. To what extent the proper motions are aligned depends on the angular extent of the star cluster on the sphere. The moving cluster parallax is a projection effect, similar to that known from viewing railway tracks. The directions of velocity vectors pointing away from us seem to converge and intersect at the convergence point. The connecting line from the observer to the convergence point is parallel to the velocity vector of the star cluster. By measuring its radial velocity (Doppler measurements) we can use this geometric argument to determine its true three-dimensional velocity, then from that its true tangential velocity. Knowing its tangential velocity and measuring its proper motion thus allow us to determine its distance.
- **Dynamical Parallax:** The distance to a visual binary star may be estimated from the masses of its two components, the size of their orbit, and the period of their revolution around one another. In this technique the masses of the binaries must be estimated either through a mass-luminosity relation or from stellar spectra. Knowing the masses and period we can use Kepler’s Third Law to determine how far apart the binaries are physically separated. Then by measuring the angular extent of the semi-major axis, we can determine their distance from us.
- **Main Sequence Fitting\***: Most stars in the colour-magnitude diagram are located along the MS. This enables us to compile a calibrated MS of those stars whose trigonometric parallaxes are measured, thus with known distances. The stars of a star cluster define their own MS; since they are all located at the same distance, their MS is already defined in a colour-magnitude diagram in which only apparent magnitudes are plotted. This cluster main sequence can then be fitted to a calibrated main sequence<sup>3</sup> by a suitable choice of the distance.
- **Spectroscopic Parallax:** From the spectrum of a star, the spectral type as well as the luminosity class can be determined. The former is determined from the strength of various absorption lines in the spectrum, while the latter is obtained from the width of the lines. Combined with the apparent magnitude, this can be used to determine a distance modulus to the star. Technically spectroscopic parallax is useful up to 10 Mpc, but in practice it is only used up to 0.1 Mpc.
- **Expansion Velocities of SNe Ejecta:** By measuring the proper motion of the ejecta across the sky, combined with radial velocity measurements, and the assumption of an isotropic flow, this can be used as a distance estimate.

Now we will describe methods used in determining the distance to extragalactic sources (though some are also used and calibrated within the MW); a schematic of these methods is shown in Figure 31. Unless otherwise cited, information from the following points is taken from Charles' notes and from Carroll & Ostlie (2007, pgs. 1039-1051).

- **Wilson-Bappu Effect [0.1 Mpc]:** Some stars have specific features in their spectra that allow their absolute magnitudes, and hence their distances, to be calculated. For example, the K absorption line of calcium can be quite broad, reaching maximum strength at special type K0. In late-type stars with chromospheres (types G, K, and M), a narrow emission line is seen, centred on the wide K absorption line. The width of this emission line is strongly correlated with a star’s absolute visual magnitude; known as the Wilson-Bappu Effect.
- **Bright Red Supergiants [7 Mpc]:** The brightest blue and red supergiants have been used as standard candles since they have roughly constant V band and bolometric luminosities. This requires distinguishing individual stars, giving it the same range as spectroscopic parallax.
- **Tip of the RGB [7 Mpc]:** Uses the tip of the RGB as a standard candle. Stars travelling up the RGB will eventually experience an He flash and transition off the RGB to the zero-age HB. An He flash occurs when the He core of an RGB star reaches  $\sim 0.5 M_{\odot}$ , and the luminosity prior is dependent on properties of the H-burning shell, which in turn is dependent on the He core; this means that the most luminous a red giant can get is an almost constant value. A distance, can then be estimated from the brightest RGB stars in a galaxy, since they will be very near the RGB tip.

<sup>13</sup> The **proper motion** is the motion observed along the celestial sphere which results from peculiar velocity tangential to the line of sight.

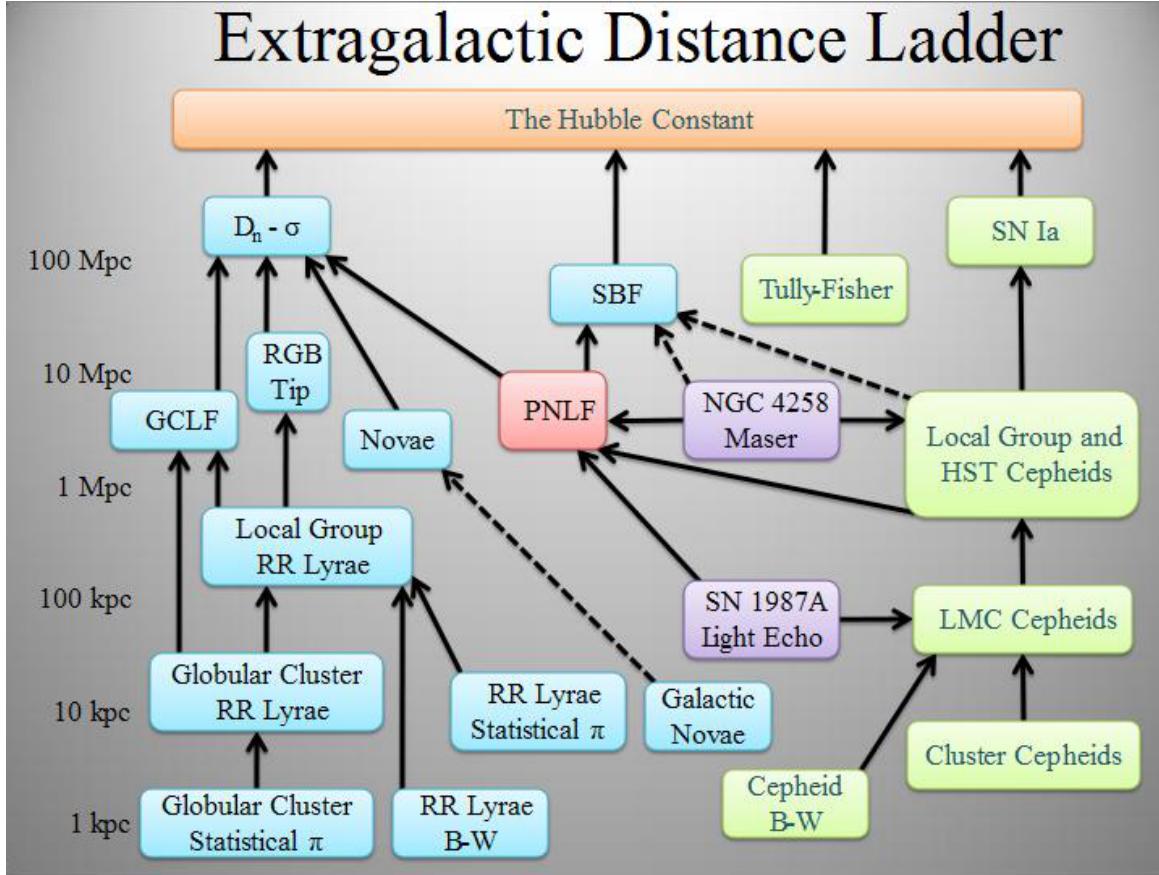


FIG. 31.— Plot of various extragalactic distance-determining techniques, with their maximum effective luminosity distance measures located to the left. The various colours represent applicability to different classes of object (but can effectively be ignored). Solid black lines indicate good calibration between two steps on the ladder; dashed lines indicate poor calibration. “GCLF” is the globular cluster luminosity function, “PNLF” is planetary nebula luminosity function, and “SBF” is surface brightness function. Image taken from Wikipedia.

- **Novae**[20 Mpc]: Novae, like their SNe Ia cousins, possess a light curve with a characteristic decay rate allowing them to be calibrated as standard candles. The relationship comes from the fact that more massive WDs tend to accrete smaller amounts of matter before a nuclear runaway occurs (i.e. producing a dimmer explosion), and this thinner layer is ejected more easily. Following this event, both novae and SNe ejected matter will begin to cool; a very coarse relationship between the temperature of the ejecta and its luminosity can be used to determine distance, but derived values are only good order of magnitude estimates.
- **Variables** [29 Mpc]: Several types of pulsating stars show periodic changes in their brightnesses, where the period of a star is related to its mass, and thus to its luminosity. This period-luminosity (PL) relation is ideally suited for distance measurements: since the determination of the period is independent of distance, one can obtain the luminosity directly from the period. A well-defined PL relation exists for three types of pulsating stars: classical Cepheids, W Virginis stars, and RR Lyrae stars (Schneider 2002, pg. 44).
- **Globular Clusters\*** [50 Mpc]: Appear to follow a luminosity function for many galaxies studied, though it is not obvious if this function is universal. Because the function has a turnover, fitting a sampling of globular clusters around a distant galaxy with the luminosity function, and comparing the turnover with that from the fit to a set of globular clusters with known distance provides a measure of relative distance.
- **Planetary Nebulae\*** [50 Mpc]: The brightness distribution of PN in a galaxy seems to have an upper limit which is nearly the same for each galaxy. If a sufficient number of PN are observed and their brightnesses measured, it enables us to determine their luminosity function from which the maximum apparent magnitude is then derived. By calibration on galaxies of known Cepheid distance, the corresponding maximum absolute magnitude can be determined, which then allows the determination of the distance modulus for other galaxies, thus their distances (Schneider 2002, pg. 116).
- **Surface Brightness Fluctuations\*** [50 Mpc]: This is based on the fact that the number of bright stars per area element in a galaxy fluctuates – purely by Poisson noise: If  $N$  stars are expected in an area element, relative fluctuations of  $\sqrt{N}/N = 1/\sqrt{N}$  of the number of stars will occur. These are observed in fluctuations of the local surface brightness. To demonstrate that this effect can be used to estimate distances, we consider a solid angle  $d\omega$ . The corresponding area element  $dA = D^2 d\omega$ .

depends quadratically on the distance  $D$  of the galaxy; the larger the distance, the larger the number of stars  $N$  in this solid angle, and the smaller the relative fluctuations of the surface brightness (Schneider 2002, pg. 116).

- **Tully-Fisher Relation\*** [ $> 100$  Mpc]: A relation exists between a spiral galaxy's luminosity and its maximum rotational velocity. Cosmological redshift-corrected Doppler shift analysis can be used to determine rotational velocities. The analogous relation for ellipticals is the **Faber-Jackson Relation**, which shows considerable spread and is much more difficult to use as a standard ruler. Physically, the Tully-Fisher relation means that the mass-to-light ratio and mean surface brightness of spirals is fairly constant. In fact, due to the changing fraction of baryons in gas instead of stars for lower mass spirals, the mass-to-light ratio does change – adding a correction term to the Tully-Fisher relation results in a much tighter relationship.
- **$D_n - \sigma$  Relation\*** [ $> 100$  Mpc]: A relation exists between an elliptical galaxy's angular diameter  $D_n$  out to a constant surface brightness ( $M_B = 20.75$  mag arcsec $^{-1}$ ) and its velocity dispersion. Since surface brightness is independent of distance,  $D_n$  is inversely proportional to the distance to the elliptical galaxy. Physically, this relation is a natural outcome of the fundamental plane and the (fairly good) assumption that all ellipticals are self-similar.
- **Masers** [ $> 1$  Gpc]: Masers are amplified microwave emissions coming from regions of interstellar media where populations (due to very low densities) can become inverted. Cosmic masers exist around some galactic nuclear regions with line luminosities up to  $10^4 L_\odot$ ; these maser sources orbit the SMBH, and observations of both the radial and proper motions of these sources can be made. If we assume circular orbits, these two values can be combined to give a distance. Source radial velocities can also be combined with an independent measure of the BH mass to determine distances.
- **Supernovae Ia** [ $> 1$  Gpc]: These SN have a strong relation between light curve decay time and peak luminosity (i.e., the Phillips relation). While most optical and infrared band observation can be used, this relation is has the smallest spread in the infrared. Other supernovae actually have similar relationships that could be used to determine luminosity (e.g., SNe IIP plateau longevity is correlated with maximum luminosity), though these relationships are generally not as well defined and/or studied, and SNe Ia are brighter in the optical and IR than other SNe.
- **Brightest Galaxies in Clusters\*** [ $> 1$  Gpc]: Fits a luminosity function to the galaxy cluster, and determines a distance modulus from the fit. This works on the same assumption that the brightest red supergiants and H II regions can be used as distance indicators to individual galaxies (i.e., it assumes that the brightest galaxies can be used as standard candles). Obviously, the danger in using this method to look out to great distances is that galaxies and galaxy clusters evolve in time.
- **Cosmological Redshift** [ $> 1$  Gpc]: Requires finding characteristic spectral features and applying the linearized form of Hubble's Law which is valid for low redshift. Use of this distance measure must take into account peculiar velocities of objects.

All techniques listed above that are denoted by an asterisk are **secondary distance indicators**, which only give distance scaling (i.e., the ratio of distance between two objects), and therefore require a calibration galaxy with a distance known by other means. **Primary distance indicators**, on the other hand, can be used on their own because they give absolute luminosity distances (Schneider 2002, pg. 116).

#### *DISTANCES TO MAGELLANIC CLOUDS, ANDROMEDA, AND VIRGO CLUSTER*

The distance of the LMC can be estimated using various methods. Since we can resolve and observe individual stars in the LMC, methods including the Wilson-Bappu, RGB tip/supergiants, and variable star methods (e.g., Cepheids), can all be used. The accepted value is that the LMC is  $\sim 50$  kpc away. Similar analysis can be performed on the SMC, yielding a distance of  $\sim 60$  kpc. Perhaps the most precise method of determining the distance to the LMC is a purely geometrical one. After SN 1987A exploded it illuminated a nearly perfectly elliptical ring. This ring consisted of material that was once ejected by the stellar winds of the progenitor star of the supernova and that is now radiatively excited by energetic photons from the supernova explosion. The corresponding recombination radiation is thus emitted only when photons from the SN hit this gas. Because the observed ring is almost certainly intrinsically circular and the observed ellipticity is caused only by its inclination with respect to the line-of-sight, the distance to SN 1987A can be derived from observations of the ring. First, the inclination angle is determined from its observed ellipticity. The gas in the ring is excited by photons from the SN a time  $R/c$  after the original explosion, where  $R$  is the radius of the ring. We do not observe the illumination of the ring instantaneously because light from the section of the ring closer to us reaches us earlier than light from the more distant part. Thus, its illumination was seen sequentially along the ring. Combining the time delay in the illumination between the nearest and farthest part of the ring with its inclination angle, we then obtain the physical diameter of the ring. Comparing this to its measured angular diameter, we are able to derive a distance estimate to the LMC (Schneider 2002, pg. 115).

The distance to Andromeda has been measured using a variety of methods including Cepheid variables, surface brightness fluctuations, red supergiant and RGB tip stars, the Wilson-Bappu effect, and measurements of eclipsing binaries (basically the dynamical parallax method). The accepted value is that Andromeda is roughly 0.8 Mpc away. The distance to the Virgo Cluster has been determined through the use of many of the extragalactic distance indicators applied to its component galaxies. These include Cepheid variables, novae, globular cluster and PN luminosity functions, surface brightness fluctuations, Tully-Fisher relation,  $D_n - \sigma$  relation, and Type Ia SNe. The accepted value is that the Virgo Cluster resides some 15 Mpc away (Carroll & Ostlie 2007, pg. 1051).

**QUESTION 5**

**What evidence is there that most galaxies contain nuclear black holes? How do those black holes interact with their host galaxies?**

## QUESTION 5

**What evidence is there that most galaxies contain nuclear black holes? How do those black holes interact with their host galaxies?**

In order to answer this question we must first define the meaning of a BH in an astronomical context. For most observational purposes, a BH is defined as a mass-concentration whose radius is smaller than the Schwarzschild radius  $r_S$  of its corresponding mass. Of course,  $r_S$  is very small: about 3 km for the Sun, and  $r_S \sim 10^7$  km  $\sim 15 R_\odot$  for the SMBH in the Galactic centre (GC). At a distance of  $D = R_0 \approx 8$  kpc, this corresponds to an angular radius of  $\sim 10^{-5}$  arcsec. Current observing capabilities are still far from resolving scales of order  $r_S$ , but in the near future Very Long Baseline Interferometry (VLBI) observations at very short radio wavelengths may achieve sufficient angular resolution to resolve the Schwarzschild radius for the Galactic black hole. If even for the closest SMBH, the one in the GC, the Schwarzschild radius is significantly smaller than the achievable angular resolution, *how can we hope to prove that SMBHs exist in other galaxies?* Like in the GC, this proof has to be found indirectly by detecting a compact mass concentration incompatible with the mass concentration of the stars observed (Schneider 2002, pg. 110).

### KINEMATICS

We first discuss the concept of the **radius of influence**, defined to be the length scale below which a BH will affect the dynamics of stellar orbits. Consider a concentration of mass  $M_{\text{BH}}$  in the centre of a galaxy where the characteristic velocity dispersion of stars (or gas) is  $\sigma$ . We compare this velocity dispersion with the characteristic velocity (e.g., Keplerian rotational velocity) around a SMBH at a distance  $r$ , given by  $\sqrt{GM_{\text{BH}}/r}$ . From this it follows that, for distances smaller than

$$r_{\text{BH}} = \frac{GM_{\text{BH}}}{\sigma^2} \sim 0.4 \text{ pc} \left( \frac{M_{\text{BH}}}{10^6 M_\odot} \right) \left( \frac{\sigma}{100 \text{ km/s}} \right)^{-2}, \quad (177)$$

the SMBH will significantly affect the kinematics of stars and gas in the galaxy. The corresponding angular scale is

$$\theta_{\text{BH}} = \frac{r_{\text{BH}}}{D} \sim 0.1 \text{ arcsec} \left( \frac{M_{\text{BH}}}{10^6 M_\odot} \right) \left( \frac{\sigma}{100 \text{ km/s}} \right)^{-2} \left( \frac{D}{1 \text{ Mpc}} \right)^{-1} \quad (178)$$

where  $D$  is the distance to the galaxy. From this we immediately conclude that our success in finding SMBHs will depend heavily on the achievable angular resolution. The HST enabled scientists to make huge progress in this field<sup>14</sup>. The search for SMBHs promises to be successful only in relatively nearby galaxies. The presence of a SMBH inside  $r_{\text{BH}}$  is revealed by an increase in the velocity dispersion for  $r \lesssim r_{\text{BH}}$ , which should then behave as  $\sigma \propto r^{-1/2}$  for  $r \lesssim r_{\text{BH}}$ . If the inner region of the galaxy rotates, one expects, in addition, that the rotational velocity  $v_{\text{rot}}$  should also increase inwards  $\propto r^{-1/2}$  (Schneider 2002, pg. 110).

In the case of the SMBH in the GC, stars have been observed with proper motions of more than 1000 km s<sup>-1</sup> and by combining the velocity dispersions in radial and tangential directions reveals the proper motion of stars to be increasing according to the Kepler law for the presence of a point mass,  $\sigma \propto r^{-1/2}$  down to  $r \sim 0.01$  pc. Moreover, the *acceleration* of some stars have also been measured, and from these measurements Sgr A\* (i.e., the observed compact, strong radio source) indeed emerges as the focus of the orbits and thus as the centre of mass. From the observed stellar kinematics, the enclosed mass  $M(r)$  as a function of  $r$  can be calculated. The corresponding analysis yields that  $M(r)$  is basically constant over the range  $0.01 \text{ pc} \lesssim r \lesssim 0.5 \text{ pc}$ . This exciting result clearly indicates the presence of a point mass, for which a mass of  $M \approx 10^6 M_\odot$  is determined (Schneider 2002, pg. 81).

### OBSERVATIONS

The practical problems in observing a SMBH have already been mentioned above. One problem is the angular resolution. To measure an increase in the velocities for small radii, the angular resolution needs to be better than  $\theta_{\text{BH}}$ . Furthermore, projection effects play a role because only the velocity dispersion of the projected stellar distribution, weighted by the luminosity of the stars, is measured. Added to this, the kinematics of stars can be rather complicated, so that the observed values for  $\sigma$  and  $v_{\text{rot}}$  depend on the distribution of orbits and on the geometry of the distribution. Despite these difficulties, the detection of SMBHs has been achieved in recent years, largely due to the much improved angular resolution of optical telescopes (like the HST) and to improved kinematic models (Schneider 2002, pg. 110).

Figure 32 shows an example for the kinematical method discussed above. A long-slit spectrum across the nucleus of the galaxy M84 clearly shows that, near the nucleus, both the rotational velocity and the velocity dispersion change; both increase dramatically towards the centre. Figure 32 also illustrates how strongly the measurability of the kinematical evidence for a SMBH depends on the achievable angular resolution of the observation. For this example of NGC 3115, observing with the resolution offered by space-based spectroscopy yields much higher measured velocities than is possible from the ground. Particularly interesting is the observation of the rotation curve very close to the centre. Another impressive example is the central region of M87, the central galaxy of the Virgo Cluster. The increase of the rotation curve and the broadening of the [O II]-line (a spectral line of singly-ionized oxygen) at  $\lambda = 3727$  Å towards the centre are displayed in Figure 32 and argue very convincingly for a SMBH with  $M_{\text{BH}} \approx 10^9 M_\odot$  (Schneider 2002, pg. 111).

<sup>14</sup> In this case we are limited by diffraction. For the HST at  $\lambda \sim 500$  nm, the resolution limit is  $\theta \sim 0.04$  arcsec (Etsuko)

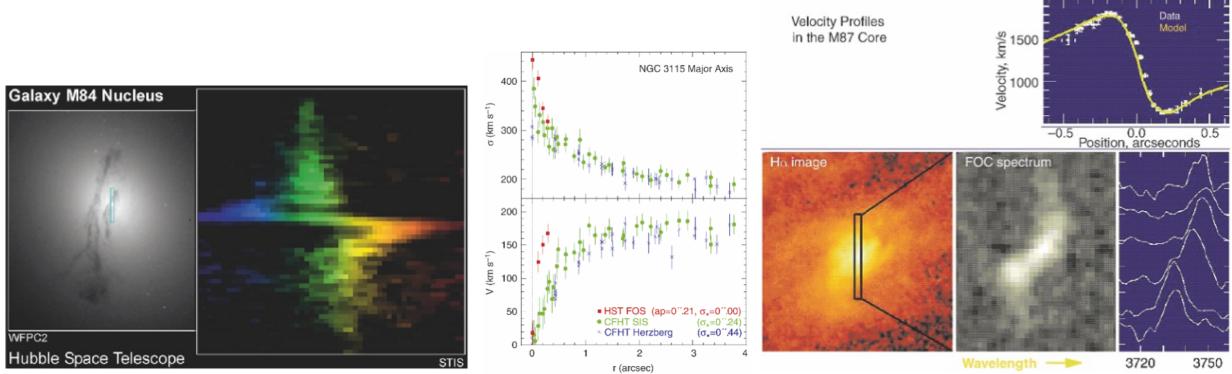


FIG. 32.— (left) An HST image of the nucleus of the galaxy M84 is shown in the left-hand panel. M84 is a member of the Virgo Cluster, about 15 Mpc away from us. The small rectangle depicts the position of the slit used by the STIS (Space Telescope Imaging Spectrograph) instrument on-board the HST to obtain a spectrum of the central region. This long-slit spectrum is shown in the right-hand panel; the position along the slit is plotted vertically, the wavelength of the light horizontally, also illustrated by colours. Near the centre of the galaxy the wavelength suddenly changes because the rotational velocity steeply increases inwards and then changes sign on the other side of the centre. This shows the Kepler rotation in the central gravitational field of a SMBH, whose mass can be estimated as  $M_{\text{BH}} \sim 10^8 M_{\odot}$ . (middle) Rotational velocity (bottom) and velocity dispersion (top), as functions of the distance from the centre along the major axis of the galaxy NGC 3115. Colours of the symbols mark observations with different instruments. Even more dramatic is the impact of resolution on measurements of the rotational velocity. Due to projection effects, the measured central velocity dispersion is smaller than the real one; this effect can be corrected for. After correction, a central value of  $\sigma \sim 600 \text{ km s}^{-1}$  is found. This value is much higher than the escape velocity from the central star cluster if it were to consist solely of stars – it would dissolve within  $10^4$  yrs. Therefore, an additional compact mass component of  $M_{\text{BH}} \sim 10^9 M_{\odot}$  must exist. (right) M87 has long been one of the most promising candidates for harbouring an SMBH in its centre. In this figure, the position of the slit is shown superimposed on an H $\alpha$  image of the galaxy (lower left) together with the spectrum of the [O II] line along this slit (bottom, centre), and six spectra corresponding to six different positions along the slit, separated by 0.14 arcsec each (lower right). In the upper right panel the rotation curve extracted from the data using a kinematical model is displayed. These results show that a central mass concentration with  $\sim 10^9 M_{\odot}$  must be present, confined to a region less than 3 pc across – indeed leaving basically no alternative but a SMBH. Images taken from Schneider (2002).

All these observations are of course no proof of the existence of a SMBH in these galaxies because the sources from which we obtain the kinematic evidence are still too far away from the Schwarzschild radius. The conclusion of the presence of SMBHs is rather that of a missing alternative; we have no other plausible model for the mass concentrations detected. As for the case of the SMBH in the MW, an ultra-compact star cluster might be postulated, but such a cluster would not be stable over a long period of time; it would dissolve within 10 Myr due to frequent stellar collisions. Based on the existence of a SMBH in our Galaxy and in AGNs, the SMBH hypothesis is the only plausible explanation for these mass concentrations (Schneider 2002, pg. 111).

#### AGN AS BHs

In order to provide an argument that AGNs must be powered by a central BH we present some relevant observational facts:

- The extent of some radio sources in AGNs may reach  $\gtrsim 1$  Mpc. From this length-scale a minimum lifetime for the activity in the nucleus of these objects can be derived, since even if the radio source expands outwards from the core with the speed of light, the age of such a source would need to be  $\tau \gtrsim 10$  Myr for the source to appear stable (i.e., in order for radio lobes to reach that far and remain filled out the source must be continually supplying
- Luminous QSOs have a luminosity of up to  $L_{\text{bol}} \sim 10^{47} \text{ erg s}^{-1}$ . Assuming that the luminosity does not change substantially over the lifetime of the source, a total energy can be estimated from the luminosity and the minimum age,  $E \sim 10^{61} \text{ erg}$ , however, the assumption of an essentially constant luminosity is not necessarily justified.
- The luminosity of some AGNs varies by more than 50% on time-scales of a day. From this variability time-scale, an upper limit for the spatial extent of the source can be determined, because the source luminosity can change substantially only on such time-scales where the source as a whole, or at least a major part of the emitting region, is in causal contact. Otherwise “one end” of the source does not know that the “other end” is about to vary. This yields a characteristic extent of the central source of  $R \lesssim 1$  lightday.

Basically, the argument in favour of BHs is found by combining the above three pieces of evidence and showing that the only suitable energy production rate *must* be gravitational in nature; thermonuclear energy production is much too low to produce that much energy in such a small volume (Schneider 2002, pg. 186).

Through the infall of matter onto a central BH, potential energy is converted into kinetic energy. If it is possible to convert part of this inward-directed kinetic energy into internal energy (heat) and subsequently emit this in the form of radiation,  $\epsilon$  (energy generation efficiency) can be larger than that of thermonuclear processes. From the theory of accretion onto black holes, a maximum efficiency of  $\epsilon \sim 6\%$  for accretion onto a non-rotating BH is derived. A black hole with the maximum allowed angular momentum can have an efficiency of  $\epsilon \sim 29\%$  (Schneider 2002, pg. 186).

#### CORRELATION WITH GALAXY PROPERTIES

Currently, strong indications of SMBHs have been found in about 35 normal galaxies, and their masses have been estimated. This permits us to examine whether, and in what way,  $M_{\text{BH}}$  is related to the properties of the host galaxy. This leads us to the

discovery of a remarkable correlation; it is found that  $M_{\text{BH}}$  is correlated with the absolute magnitude of the bulge component (or the spheroidal component) of the galaxy in which the SMBH is located. Here, the bulge component is either the bulge of a spiral galaxy or an elliptical galaxy as a whole. This correlation is described by

$$M_{\text{BH}} L_{\text{B,bulge}}^{1.11} \Leftrightarrow M_{\text{BH}} \propto M_{\text{bulge}}^{0.9}. \quad (179)$$

An even better correlation exists between  $M_{\text{BH}}$  and the velocity dispersion of the bulge component:

$$M_{\text{BH}} \propto \sigma^4, \quad (180)$$

known as the  $M - \sigma$  relation. To date, the physical origin of this very close relation has not been understood in detail. The most obvious apparent explanation – that in the vicinity of a SMBH with a very large mass the stars are moving faster than around a smaller-mass SMBH – is not conclusive: the mass of the SMBH is significantly less than one percent of the mass of the bulge component. We can therefore disregard its contribution to the gravitational field in which the stars are orbiting. Instead, this correlation has to be linked to the fact that the spheroidal component of a galaxy evolves together with the SMBH. A better understanding of this relation can only be found from models of galaxy evolution (Schneider 2002, pg. 113).

Since SMBHs appear to be closely linked with some bulk properties of galaxies, this implies an important connection between galaxy formation and the formation of SMBHs. How these behemoths formed, however, remains an open question. One popular suggestion is that they formed from collisions between galaxies; another is that they formed as an extension of the formation process of intermediate-mass black holes (IMBHs). It is still unclear of how the latter form, although the correlation of IMBHs with the cores of globular clusters and low-mass galaxies suggests that they develop in the dense stellar environments either by the mergers of stars to form a supermassive star that then core-collapses, or by the merger of stellar-mass BHs (Carroll & Ostlie 2007, pg. 639).

**QUESTION 6**

**Define and describe globular clusters. Where are they located? What are typical ages of globular clusters. How is this determined?**

## QUESTION 6

**Define and describe globular clusters. Where are they located? What are typical ages of globular clusters. How is this determined?**

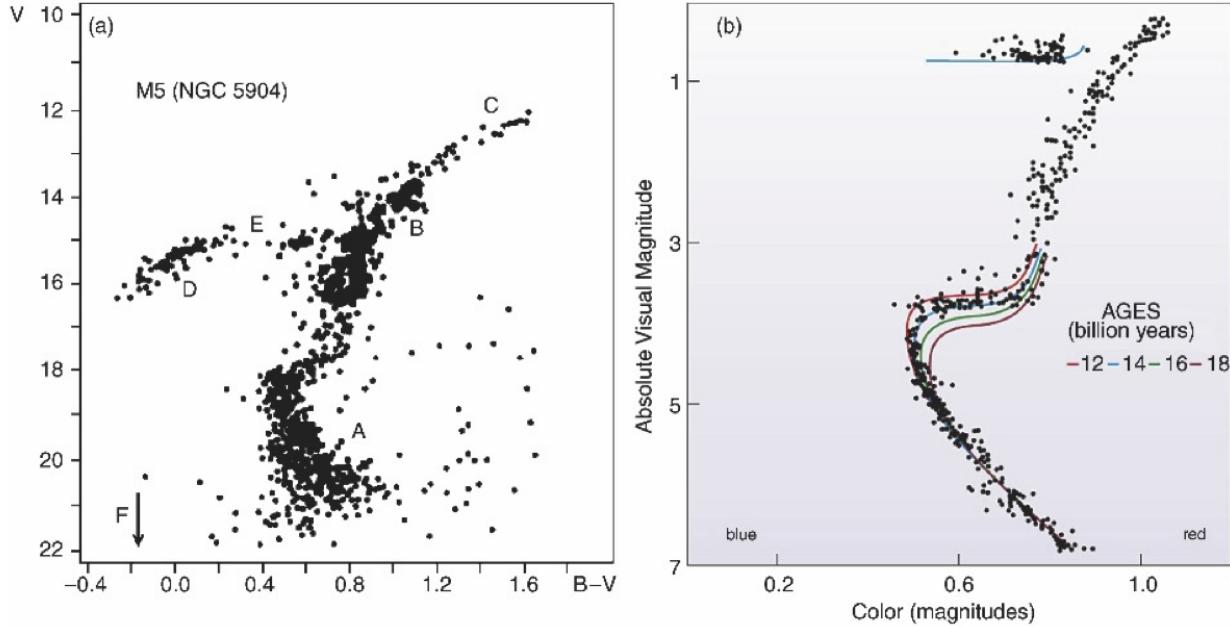


FIG. 33.— The left panel shows the colour-magnitude diagram for globular cluster M5. The different labels denote evolutionary stages: A is the main sequence; B is the red giant branch; C is the point of helium flash; D is horizontal branch; E is the Schwarzschild gap in the horizontal branch; F denotes white dwarfs (below the arrow). The turn-off point is located where the main sequence turns over to the red giant branch. The right panel shows the colour-magnitude diagram of the globular cluster 47 Tucanae compared to isochrone models parameterized by age (denoted by solid coloured lines). The ages plotted in this diagram should actually be decreased by 2 Gyr since those shown here were constructed from incorrect cluster distance measures (see text). Image taken from Schneider:2002.

A main component of the MW is the Galactic halo – a roughly spherical distribution of stars and globular clusters that surrounds the disk of the galaxy. There are about 150 of such clusters in the MW and they tend to have large velocity components perpendicular to the Galactic plane. Globular clusters are characterized as a large collection of several hundred thousand stars contained within a spherical region of radius  $\sim 20$  pc. The collection of stars are gravitationally bound and orbit within their common gravitational field (Schneider 2002, pg. 55).

Globular clusters are believed to have formed from the collapse and fragmentation of a single molecular cloud. As a result, the stars within the cluster formed with essentially identical compositions and within a relatively short period of time. From the Vogt-Russell theorem<sup>15</sup> this implies that the evolutionary tracks of the individual stars are based solely on their initial masses (Carroll & Ostlie 2007, pg. 474).

This property of globular clusters allows for an easy determination of their distance and age. It is straightforward to construct a modified version of an HR diagram of the stars in the cluster by plotting their apparent magnitude rather than absolute magnitude. Since the size of the cluster is small compared to its distance from the Earth, plotting their apparent magnitudes only amounts to shifting the vertical axis by some constant factor. The distance to the cluster can then be determined by matching the apparent main sequence of the cluster to a calibrated main sequence in absolute magnitude. This is known as spectroscopic parallax or MS fitting (Carroll & Ostlie 2007, pg. 475). Most globular clusters are found within 35 kpc from the Galactic centre although some are also found further than 60 kpc (Schneider 2002, pg. 55). At such large distances it is difficult to judge whether they are actually part of the MW or were captured from satellite galaxies.

Rather than determining the effective temperature of each star in the cluster by analyzing their spectra, it is more convenient to plot their apparent magnitude against their colour indices ( $B-V$ ). This plot is known as a colour-magnitude diagram (see Figure 33) and provides a way of determining the cluster age. At the onset of the cluster's formation the most massive stars arrive on the main sequence and evolve rapidly, possibly evolving into red giants and going supernova before the lowest-mass stars have even reached the main sequence. Since core hydrogen-burning lifetimes are inversely related to mass<sup>16</sup> the MS turn-off point (the point at which stars in the cluster are currently evolving off the MS) becomes redder and less luminous with time. Therefore, the age of the globular cluster can be determined from the position of the turn-off point by comparing it with models of stellar evolution (Carroll & Ostlie 2007, pg. 476).

In order to determine the age of the cluster in this way one compares the colour-magnitude diagram to isochrone models. An **isochrone** is simply a snapshot in time of an evolving HR diagram. The distribution of stars on the isochrone depends on the

<sup>15</sup> This states that the evolution of a star is uniquely determined by its mass and internal composition (Carroll & Ostlie 2007, pg. 333).

<sup>16</sup> To see this we note that since  $E = Mc^2$  the energy content of a star is  $E \propto M$ . For main-sequence stars we roughly observe that  $L \propto M^{3.5}$  so the time it takes them to consume their fuel goes like  $t = E/L \propto M^{-2.5}$  (Schneider 2002, pg. 429).

number of stars in each mass range within the cluster (i.e., through the IMF), combined with the different rates of evolution during each phase of stellar evolution. As we look at isochrones of later and later times we will see fewer and fewer massive stars since they will have evolved off the main sequence and ended up as SN or WDs (Carroll & Ostlie 2007, pg. 476). The colour-magnitude diagram of the globular cluster 47 Tucanae is compared to isochrones in Figure 33.

The mean age of the oldest globular clusters is  $11.5 \pm 1.3$  Gyr (Chaboyer et al. 1998). Prior to the work by Chaboyer et al. (1998) this age estimate was about 3 Gyr higher due to erroneous distance measurements to globular clusters. These distance measures were coupled to theoretical isochrone models in order to calibrate the relationship between age and MS turn-off. The distance measures were obtained from using RR Lyrae stars in the globular clusters as standard candles; uncertainties with this method were what plagued the original mass estimates. Chaboyer et al. (1998) used the Hipparcos catalogue (parallaxes of stars) along with other independent distance estimates to show that the previous distance measures were too low, thereby decreasing the age estimates. Using parallaxes of nearby field stars it is possible to define the position of the ZAMS and, through a comparison to a cluster colour-magnitude diagram, obtain a direct distance to the cluster. Since the location of the ZAMS is sensitive to metallicity only metal-poor subdwarfs were used for this.

An interesting feature evident in the colour-magnitude diagrams of young globular clusters is the **Hertzsprung Gap**. This is the paucity of stars located between the turn-off point of the MS and the RGB. This feature arises because of the rapid evolution that occurs just after leaving the MS. The evolution of this stage is largely damped for low mass stars ( $1.24 M_{\odot}$ ) and so this gap does not appear for old globular clusters where the turn-off point is located at low mass scales. Another interesting feature seen in some globular clusters is the existence of a group of stars, known as **blue stragglers**, that can be found above the turn-off point (i.e., still on the MS). Although our understanding of these stars is incomplete, it appears that their tardiness in leaving the MS is due to some unusual aspects of their evolution. The most likely scenarios appear to be mass exchange with a binary star companion, or collisions between two stars, extending the star's MS lifetime (Carroll & Ostlie 2007, pg. 478).

**QUESTION 7**

**What is the X-ray background and how is it produced?**

## QUESTION 7

**What is the X-ray background and how is it produced?**

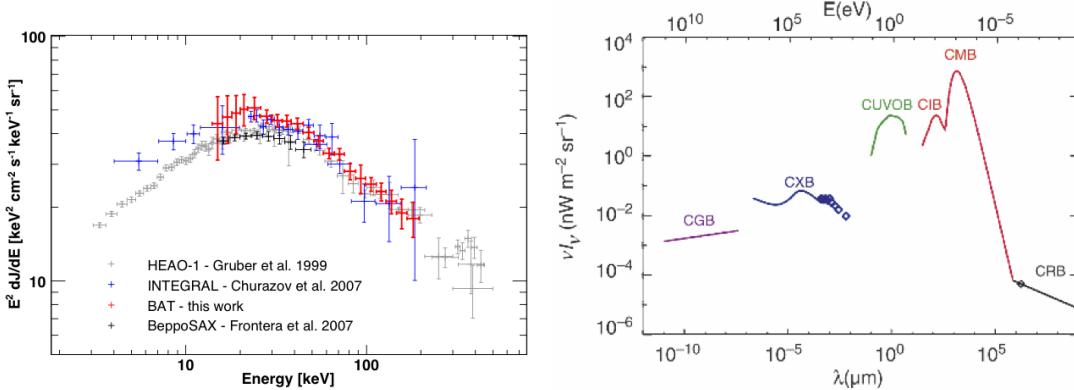


FIG. 34.— (left) Comparison of measurements of the CXB. Here  $E^2 dF/dE$  is plotted where  $F$  is the measured flux; dimensionally equivalent to  $\lambda F_\lambda$ . If we instead plot  $F_\lambda$  then the shape will be modified. In particular, a linear relationship in  $\lambda F_\lambda \propto \lambda$  is equivalent to a flat relationship  $F_\lambda = \text{const}$ . Image taken from ajello:2009. (right) Spectrum of cosmic background radiation, plotted as  $\nu I_\nu$  versus  $\lambda$ . Besides the CMB, background radiation exists in the radio domain (cosmic radio background, CRB), in the infrared (CIB), in the optical/UV (CUVOB), in the X-ray (CXB), and at gamma-ray energies (CGB). With the exception of the CMB, probably all of these backgrounds can be understood as a superposition of the emission from discrete sources. Furthermore, this figure shows that the energy density in the CMB exceeds that of other radiation components. Image taken from Schneider (2002).

In the 1970s, the first X-ray satellites discovered not only a number of extragalactic X-ray sources (such as AGNs and clusters of galaxies), but also an apparently isotropic radiation component, the **CXB**. Its spectrum is a very hard (i.e., flat) power law, cut off at an energy above  $E_0 \sim 40$  keV, which can roughly be described by

$$I_\nu \propto E^{-0.3} \exp\left(-\frac{E}{E_0}\right). \quad (181)$$

This spectrum is shown in Figure 34. Initially, the origin of this radiation was unknown, since its spectral shape was different from the spectra of sources that were known at that time. For example, it was not possible to obtain this spectrum by a superposition of the spectra of known AGNs (Schneider 2002, pg. 380).

ROSAT, with its substantially improved angular resolution compared to earlier satellites (such as the Einstein observatory), conducted source counts at much lower fluxes, based on some very deep images. From this, it was shown that at least 80% of the CXB in the energy range between 0.5 keV and 2 keV is emitted by discrete sources, of which the majority are AGNs. Hence it is natural to assume that the total CXB at these low X-ray energies originates from discrete sources, and observations by XMM-Newton seem to confirm this. However, the X-ray spectrum of normal AGNs is different from equation (181), namely it is considerably steeper (about  $S_\nu \propto \nu^{-0.7}$ ). Therefore, if these AGNs contribute the major part of the CXB at low energies, the CXB at higher energies cannot possibly be produced by the same AGNs. Subtracting the spectral energy of the AGNs found by ROSAT from equation (181), one obtains an even harder spectrum, resembling very closely that of thermal bremsstrahlung. Therefore, it was supposed for a long time that the CXB is, at higher energies, produced by a hot intergalactic gas at temperatures of  $kT \sim 30$  keV (Schneider 2002, pg. 380).

This model was excluded, however, by the precise measurement of the thermal spectrum of the CMB by COBE, showing that the CMB has a perfect blackbody spectrum. If a postulated hot intergalactic gas were able to produce the CXB, it would cause significant deviations of the CMB from the Planck spectrum, namely by the inverse Compton effect (the same effect that causes the SZ effect in clusters of galaxies). Thus, the COBE results clearly ruled out this possibility (Schneider 2002, pg. 381).

By now, the nature of the CXB at higher energies has also essentially been determined mainly through very deep observations with the Chandra satellite. About 75% of the CXB in the energy range of  $2 \text{ keV} \leq E \leq 10 \text{ keV}$  could be resolved into discrete sources. Again, most of these sources are AGNs, but typically with a significantly harder (i.e., flatter) spectrum than the AGNs that are producing the low-energy CXB. Such a flat X-ray spectrum can be produced by photoelectric absorption of an intrinsically steep (i.e., softer) power-law spectrum, where photons closer to the ionization energy are more efficiently absorbed than those at higher energy. According to the classification scheme of AGNs, these are Type 2 AGNs, thus Seyfert 2 galaxies and QSOs with strong intrinsic self-absorption (Schneider 2002, pg. 381).

There is evidence of the presence of a population of AGNs which is still currently undetected even in the deepest surveys. The analysis of the unresolved component revealed that it might be consistent with the integrated emission of a population of very absorbed, Compton-thick ( $\tau = N_H \sigma_T \sim 1$ ) AGNs. Given the fact that their emission is suppressed below 10 keV, detecting these objects is extremely difficult at soft X-rays and until a few years ago only a handful of Compton-thick AGNs were known (Ajello 2009).

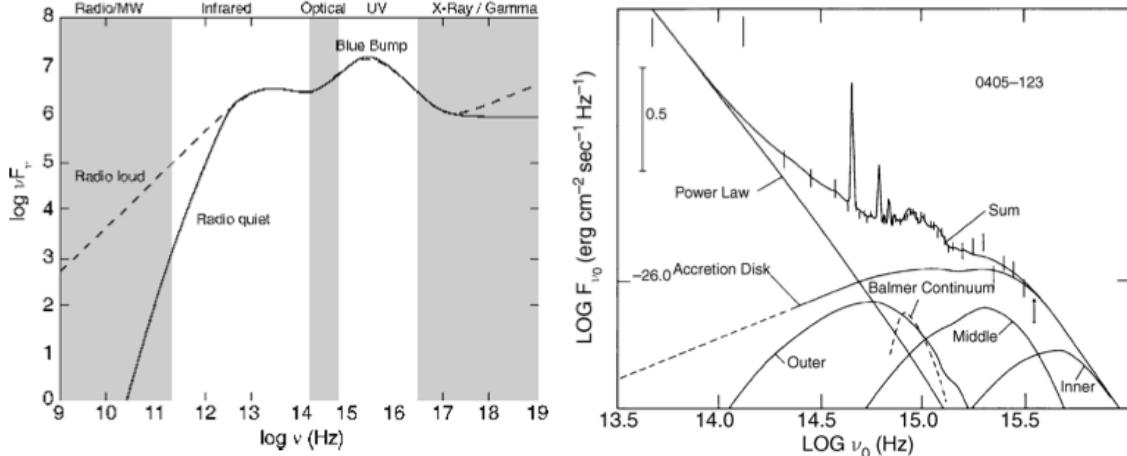


FIG. 35.— (left) Sketch of the characteristic spectral behaviour of a QSO. We distinguish between radio-loud (dashed curve) and radio-quiet (solid curve) QSOs. Plotted is  $\nu S_\nu$  (in arbitrary units), so that flat sections in the spectrum correspond to equal energy per logarithmic frequency interval. The most prominent feature is the big blue bump, a broad maximum in the UV up to the soft X-ray domain of the spectrum. Besides this maximum, a less prominent secondary maximum is found in the IR and is associated with warm dust  $T \sim 2000$  K. The spectrum increases towards higher energies in the X-ray domain of the spectrum – typically  $\sim 10\%$  of the total energy is emitted as X-rays. (right) Spectrum of a quasar at  $z = 0.57$  (data points with error bars) from the NIR and the optical up to the UV spectral region, plus a model for this spectrum (solid curve). The latter combines various components: (1) the radiation from an accretion disk that causes the big blue bump and whose spectrum is also shown for three individual radius ranges, (2) the Balmer continuum created by a forest of H line emission from  $n = 2$  to higher energy levels, and (3) an underlying power law which may have its origin in synchrotron emission of the accretion jets. Images taken from Schneider (2002).

AGN intrinsic emission is shaped by the AGN's various different regions. Thermal emission from the accretion disk provides the broad IR to UV continuum that peaks in the “big blue bump” in the UV (this bump is usually not directly observed; its empirical existence is extrapolated from UV and soft X-ray observations). For blazars, the optical continuum is dominated by power-law continuum emission, likely due to synchrotron radiation from the accretion jets. Warm dust contributes a second bump in the mid-IR (Charles).

The broad IR to UV excess is caused by emission from different parts of the accretion disk. Suppose a parcel of mass  $dm$  at a radial distance  $r$  from the black hole of mass  $M$  drops a distance  $dr$ . Taylor expansion of the gravitational potential energy gives  $dE = GMmdr/r^2$ . If half (from the virial theorem) of this liberated energy is expelled as heat (i.e., radiation), this yields a luminosity  $dL = GM\dot{m}dr/2r^2$ , where  $\dot{m}$  denotes the accretion rate. We can assume that the disk is a blackbody, radiating from the top and bottom surfaces, in which case  $dL = (2)\pi r dr \sigma_{SB} T^4$ . Equating the two  $dL$ 's gives us

$$T(r) = \left( \frac{3GM\dot{m}}{8\pi\sigma_{SB}r^3} \right)^{1/4} \Leftrightarrow T(r) = \left( \frac{3c^6}{64\pi\sigma_{SB}G^2} \right)^{1/4} \dot{m}^{1/4} M^{-1/2} \left( \frac{r}{r_s} \right)^{-3/4}. \quad (182)$$

From this analysis, we can immediately draw a number of conclusions. The most surprising one may be the independence of the temperature profile of the disk from the detailed mechanism of the dissipation because the equations do not explicitly contain the viscosity<sup>17</sup>. This fact allows us to obtain quantitative predictions based on the model of a *geometrically thin, optically thick accretion disk*. The temperature in the disk increases inwards  $\propto r^{-3/4}$ , as expected. Therefore, the total emission of the disk is, to a first approximation, a superposition of black bodies consisting of rings with different radii at different temperatures. For this reason, the resulting spectrum does not have a Planck shape but instead shows a much broader energy distribution; see Figure 35 for a complete spectrum (Schneider 2002, pg. 187).

For any fixed ratio  $r/r_s$ , equation (182) shows that the temperature increases with the accretion rate. This again was expected: since the local emission is  $\propto T^4$  and the locally dissipated energy is  $\propto \dot{m}$ , it must be  $T \propto \dot{m}^{1/4}$ . Furthermore, at fixed ratio  $r/r_s$ , the temperature decreases with increasing mass  $M$  of the BH. This implies that the maximum temperature attained in the disk is lower for more massive black holes. This may be unexpected, but it is explained by a decrease of the tidal forces, at fixed  $r/r_s$ , with increasing  $M$ . In particular, it implies that the maximum temperature of the disk in an AGN is much lower than in accretion disks around stellar sources. Accretion disks around NSs and stellar-mass BHs emit in the hard X-ray part of the spectrum and are known as X-ray binaries. In contrast, the thermal radiation of the disk of an AGN extends to the UV range only (Schneider 2002, pg. 188).

#### OTHER BACKGROUNDS

Apparently isotropic radiation has been found in wavelength domains other than the microwave and X-ray regimes; see Figure 34. Following the terminology of the CMB, these are called background radiation as well. However, the name should not imply that it is a background radiation of cosmological origin, in the same sense as the CMB. In the present context, we simply denote the flux in a specific frequency domain, averaged over sky position at high Galactic latitudes, as background radiation. Thus,

<sup>17</sup> The physical mechanism that is responsible for the viscosity is unknown. The molecular viscosity is far too small to be considered as the primary process. Rather, the viscosity is probably produced by turbulent flows in the disk or by magnetic fields, which become spun up by differential rotation and thus amplified, so that these fields may act as an effective friction (Schneider 2002, pg. 187).

when talking about an optical background here, we refer to the sum of the radiation of all galaxies and AGNs per solid angle (Schneider 2002, pg. 379).

Observations of background radiation in the infrared are very difficult to accomplish. First, it is problematic to measure absolute fluxes due to the thermal emission of the detector. In addition, the emission by interplanetary dust (and by the ISM in our MW) is much more intense than the infrared flux from extragalactic sources. For these reasons, the absolute level of the **CIB** has been determined only with relatively large uncertainties. The ISO (Infrared Space Observatory) satellite was able to resolve about 10% of the CIB at  $\lambda = 175 \mu\text{m}$  into discrete sources. Also in the sub-mm range (at about  $850 \mu\text{m}$ ) almost all of the CIB seems to originate from discrete sources which consist mainly of dust-rich star-formation regions. In any case, no indication has yet been found that the origin of the CIB is different from the emission by a population of discrete sources, in particular of high-redshift starburst galaxies. Further resolving the background radiation into discrete sources will become possible by future FIR satellites such as, for instance, Herschel (Schneider 2002, pg. 380).

**QUESTION 8**

**Describe the currently accepted model for the formation of the various types of galaxies. What are the lines of evidence to support this model?**

## QUESTION 8

**Describe the currently accepted model for the formation of the various types of galaxies. What are the lines of evidence to support this model?**

### *SPIRAL GALAXIES*

In a CDM model, halos of lower mass form first; only later can more massive halos form. This “bottom-up” scenario of structure formation follows from the shape of the power spectrum of density fluctuations, which itself is defined by the nature of dark matter – namely cold dark matter. The formation of halos of increasingly higher mass then happens by the merging of lower-mass halos. Such merging processes are directly observable in galaxy collisions. Merging should be particularly frequent in regions where the galaxy density is high, in clusters of galaxies for instance (Schneider 2002, pg. 391). The fact that we observe frequent mergers and smaller, more irregularly shaped, bluer galaxies at high redshift, in contrast to the larger, more organized galaxies in the local universe, acts in favour of hierarchical buildup.

If the gas in a halo can cool efficiently, stars may form. Since cooling is a two-body process (i.e., the cooling rate per volume element is  $\propto \rho^2$ ), only dense gas can cool efficiently. One expects that the gas, having a finite amount of angular momentum like the dark matter halo itself, will initially accumulate in a disk, as a consequence of its own dissipation. The gas in the disk then reaches densities at which efficient star formation can set in. In this way, the formation of disk galaxies, thus of spirals, can be understood qualitatively. It is generally believed that small spiral galaxies form first and are progressively built up by a series of **minor mergers**. A minor merger is one for which a smaller galaxy merges with a massive one, so that the properties of the dominating galaxy are expected to change only marginally (e.g., Sagittarius dwarf galaxy and the MW) (Schneider 2002, pg. 392).

However the hierarchical nature of structure growth implies that due to subsequent merging events, the disks can be significantly perturbed or even destroyed. Furthermore, the disks can lose angular momentum in the course of galaxy collisions. It is likely that an understanding of the formation of disk galaxies requires additional ingredients; for example, disks may form as a result of gas-rich mergers, where the resulting angular momentum of the baryons is sufficient to form a rotating and flat structure through dissipation (Schneider 2002, pg. 392).

### *ELLIPTICAL GALAXIES*

The question of the formation of ellipticals is considerably more difficult to answer. Stars in ellipticals feature a very high velocity dispersion, indicating that the gas out of which they have formed cannot have kinematically cooled down beforehand into a disk by dissipation. On the other hand, it is hard to comprehend how star formation may proceed without gas compression induced by dissipation and cooling (Schneider 2002, pg. 392).

A simple model is capable of coherently describing the features of elliptical galaxies (e.g., fundamental plane which implies a passive evolution of the stellar population in ellipticals, and the colour-magnitude of ellipticals which suggests that the stellar populations of ellipticals at a given redshift all have a similar age); this model is known as **monolithic collapse**. According to this description, the gas in a halo is nearly instantaneously transformed into stars. In this process, most of the gas is consumed, so that no further generations of stars can form later. For all ellipticals with the same redshift to have nearly identical colours, this formation must have taken place at relatively high redshift,  $z \gtrsim 2$ , so that the ellipticals are all of essentially the same age. This scenario thus requires the formation of stars to happen quickly enough, before the gas can accumulate in a disk. The process of star formation remains unexplained in this picture, however (Schneider 2002, pg. 392).

We rather expect, according to the model of hierarchical structure formation, that massive galaxies form by a series of **minor and major mergers**. A major merger is one for which both partners have a comparable mass. In this case, both galaxies will change completely. The disks may be destroyed if the disk population attains a high enough velocity dispersion to transform it into a spheroidal component. Furthermore, the gas orbits are perturbed, which may trigger massive starbursts. By means of this perturbation of gas orbits, the SMBH in the centres of the galaxies can be fed, initiating AGN activity. Due to the violence of the interaction, part of the matter is ejected from the galaxies. These stars and the respective gas are observable as tidal tails in optical images or by the 21-cm emission of neutral hydrogen. From these arguments, which are also confirmed by numerical simulations, one expects that in a “major merger” an elliptical galaxy may form. In the violent interaction, the gas is either ejected, or heated so strongly that any further star formation is suppressed (Schneider 2002, pg. 393).

This scenario for the formation of ellipticals is expected from models of structure formation. Thus far it has been quite successful. For instance, it provides a straightforward explanation for the **Butcher-Oemler effect**, which states that clusters of galaxies at higher redshift contain a larger fraction of blue galaxies. Because of the particularly frequent mergers in clusters, due to the high galaxy density, such blue galaxies are transformed more and more into early-type galaxies. However, we note that galaxies in clusters may also lose their gas in their motion through the hot IGM by which the gas is ripped out due to the so-called **ram pressure**. In this case, the gas of the disk is stripped, no further star formation takes place, and the spiral galaxy is changed into a disk galaxy without any current star formation. The fact that the fraction of ellipticals in a cluster remains rather constant as a function of redshift, whereas the abundance of S0 galaxies increases with decreasing  $z$ , indicates the importance of this process as an explanation of the Butcher-Oemler effect. On the other hand, many ellipticals show signs of complex evolution (e.g., shells and ripples in isophotes) which can be interpreted as the consequence of such mergers. Therefore, it is quite possible that the formation of ellipticals in galaxy groups happens by violent merger processes, and that these then contribute to the cluster populations by the merging of groups into clusters (Schneider 2002, pg. 393).

This model also has its problems, though. One of these is that merger processes of galaxies are also observed to occur at lower redshifts. Ellipticals formed in these mergers would be relatively young (in terms of their stars, not the galaxy as a whole), which is hardly compatible with the above finding of a consistently old age of ellipticals. However, ellipticals are predominantly

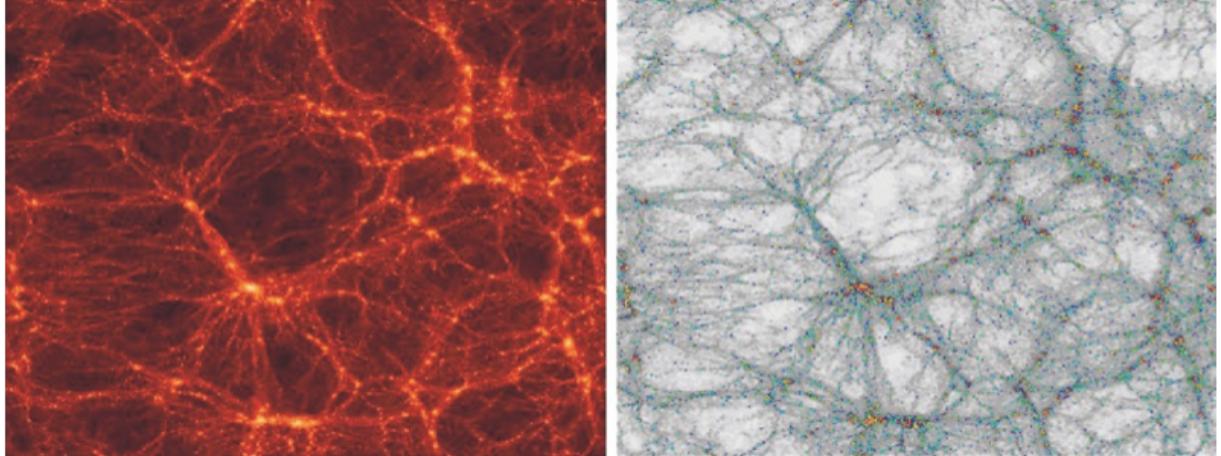


FIG. 36.— On the left, the distribution of DM resulting from an N-body simulation is shown. The DM halos identified in this mass distribution were then modelled as the location of galaxy formation – the formation of halos and their merger history can be followed explicitly in the simulations. Semi-analytic models describe the processes which are most important for the gas and the formation of stars in halos, from which a model for the distribution of galaxies results. In the panel on the right, the resulting distribution of model galaxies is represented by coloured dots, where the colour indicates the spectral energy distribution of the respective galaxy: blue indicates galaxies with active star formation, red are galaxies which are presently not forming any new stars. The latter are particularly abundant in clusters of galaxies – in agreement with observations. Image taken from Schneider (2002).

located in galaxy clusters whose members are already galaxies with a low gas content. In the merging process of such galaxies, the outcome will be an elliptical, but no starburst will be induced by merging because of the lack of gas – such mergers are sometimes called *dry* mergers (Schneider 2002, pg. 393).

#### RESULTS FROM SEMI-ANALYTIC MODELLING

Since galaxies preferentially form in filaments of large-scale structure, the accretion of smaller mass halos onto a high-mass halo occurs predominantly in the direction of the filament. The most massive sub-halos therefore tend to form a planar distribution, not unlike the one seen in the MW's satellite distribution in which the 11 satellites form a nearly flattened system oriented perpendicular to the disk of the galaxy (Schneider 2002, pg. 397).

In the framework of semi-analytic models, a spheroidal stellar population may form in a major merger, which may be defined in terms of the mass ratio of the merging halos (e.g., larger than 1:3) – the disk populations of the two merging galaxies are dynamically heated to commonly form an elliptical galaxy. The gas in the two components is heated by shocks to the virial temperature of the resulting halo, which suppresses future star formation. If the masses of the two components in a merger are very different, the gas of the smaller component will basically be accreted onto the more massive halo, where it can cool again and form new stars. By this process, a new disk population may form. In this model, a spiral galaxy is created by forming a bulge in a major merger at earlier times, with the disk of stars and gas being formed later in minor mergers and by the accretion of gas. Hence the bulge of a spiral is, in this picture, nothing but a small elliptical galaxy, which is also suggested by the very similar characteristics of bulges and ellipticals, including the fact that both types of object seem to follow the same relation between BH mass and velocity dispersion (Schneider 2002, pg. 397).

Semi-analytic modelling predicts that galaxies in clusters basically consist of old stellar populations, because here the merger processes were already concluded quite early in cosmic history. Therefore, at later times gas was no longer available for the formation of stars. Figure 36 shows the outcome of such a model in which the merger history of the individual halos has been taken straight from the numerical N-body simulation, hence the spatial locations of the individual galaxies are also described by these simulations (Schneider 2002, pg. 398). This is consistent with the **morphology-density relation** for which elliptical galaxies are much more abundant relative to spirals in the centres of dense, rich clusters of galaxies, whereas spirals dominate in less dense clusters and near the periphery of rich clusters. This effect may be partly explained by the increased likelihood of interactions in regions where galaxies are more tightly packed, destroying spirals and forming ellipticals (Carroll & Ostlie 2007, pg. 1028).

However, a competing hypothesis has also been suggested – namely, that ellipticals tend to develop preferentially near the bottoms of deep gravitational potential wells, even in the absence of interactions. Lower mass-density fluctuations in the early universe may have resulted in spiral galaxies, and the smallest fluctuations led to the formation of dwarf systems. If this is the case, then the large number of dSph's and dE's that exist has a natural explanation in the much larger number of smaller fluctuations that formed in the early universe. This mechanism could help explain galactic morphology if the initial density fluctuations in the early universe were largest in what later became the centres of rich clusters. Because the gravitational potential well in those regions would have been deeper, the probability of collisions between protogalactic clouds should have been correspondingly greater as well (Carroll & Ostlie 2007, pg. 1029).

#### COSMIC DOWNSIZING

The hierarchical model of structure formation predicts that smaller-mass objects are formed first, with more massive systems forming later in the cosmic evolution. There is ample evidence for this to be the case (e.g., galaxies are in place early in the

cosmic history, whereas clusters are abundant only at redshifts  $z \lesssim 1$ ). However, looking more closely into the issue, apparent contradictions are discovered. For example, the most massive galaxies in the local Universe, the massive ellipticals, contain the oldest population of stars, although their formation should have occurred later than those of less massive galaxies. In turn, most of the star formation in the local Universe seems to be associated with low- or intermediate-mass galaxies, whereas the most massive ones are passively evolving. Now turning to high redshift: for  $z \sim 3$ , the bulk of star formation seems to occur in the LBGs (Lyman-Break Galaxies), which are associated with high-mass halos. The study of passively evolving EROs (Extremely Red Objects) indicates that massive old galaxies were in place as early as  $z \sim 2$ , hence they must have formed very early in the cosmic history. The phenomenon that massive galaxies form their stars in the high-redshift universe, whereas most of the current star formation occurs in galaxies of lower mass, has been termed **cosmic downsizing** (Schneider 2002, pg. 401).

This downsizing can be studied in more detail using redshift surveys of galaxies. The observed line width of the galaxies yields a measure of the characteristic velocity and thus the mass of the galaxies (and their halos). Such studies have been carried out in the local universe, showing that local galaxies have a bimodal distribution in colour, which in turn is related to a bimodal distribution in the specific star-formation rate. Extending such studies to higher redshifts, by spectroscopic surveys at fainter magnitudes, we can study whether this bimodal distribution changes over time. In fact, such studies reveal that the characteristic mass separating the star-forming galaxies from the passive ones evolves with redshift, such that this dividing mass increases with  $z$ . Hence, the mass scale above which most galaxies are passively evolving decreases over time, restricting star formation to increasingly lower-mass galaxies (Schneider 2002, pg. 401).

Another problem with which models of galaxy formation are faced is the absence of very massive galaxies today. The luminosity function of galaxies is described reasonably well by a Schechter luminosity function (i.e., there is a luminosity scale  $L_*$  above which the number density of galaxies decreases exponentially), but this is in conflict with the distribution of DM halos. *Why, then, is there some kind of maximum luminosity (or stellar mass) for galaxies?* It has been suggested that the value of  $L_*$  is related to the ability of gas in a dark matter halo to cool; if the mass is too high, the corresponding virial temperature of the gas is large and the gas density low, so that the cooling times are too large to make gas cooling, and thus star formation, efficient. With a relatively high cosmic baryon density of  $\Omega_{\text{bary}} = 0.045$ , however, this argument fails to provide a valid quantitative explanation (Schneider 2002, pg. 401).

The clue to the solution of these problems may come from the absence of cooling flows in galaxy clusters. The gas density in the inner regions of clusters is large enough for the gas to cool in much less than a Hubble time. However, in spite of this fact, the gas seems to be unable to cool, for otherwise the cool gas would be observable by means of intense line radiation. This situation resembles that of the massive galaxies: if they were already in place at high redshifts, why has additional gas in their halos (visible, e.g., through its X-ray emission, and expected in structure formation models to accrete onto the host halo) not cooled and formed stars? The solution for this problem in galaxy clusters was the hypothesis that AGN activity in their central galaxy puts out enough energy to heat the gas and prevent it from cooling to low temperatures (Schneider 2002, pg. 401).

A similar mechanism may occur in galaxies as well. We observe that galaxies host a SMBH whose mass scales with the velocity dispersion of the spheroidal stellar component and thus, in elliptical galaxies, with the mass of the DM halo. If gas accretes onto these halos, it may cool and, on the one hand, form stars; on the other hand, this process will lead to accretion of gas onto the central BH and make it active again. This activity can then heat the gas and thus prevent further star formation. If the time needed to cool the gas and form stars is shorter than the free-fall time to the centre of the galaxy, stars can form before the AGN activity is switched on. In the opposite case, star formation is prevented. A quantitative analysis of these two time-scales shows that they are about equal for a halo of mass  $\sim 10^{11} M_\odot$ , about the right mass-scale for explaining the cut-off luminosity  $L_*$  in the Schechter function (Schneider 2002, pg. 402).

This also provides a mechanism for understanding the downsizing problem. In fact, since the mass of the central BH was accumulated by accretion, and since the total energy output that can be generated in the course of growing a BH to a mass of  $\sim 10^8 M_\odot$  is very large, it should not be too surprising that this nuclear activity has a profound impact on the galaxy hosting the SMBH. The fact that the hosts of luminous QSOs show no signs of strong star formation may be another indication that the AGN luminosity prevents efficient star formation in its local environment (Schneider 2002, pg. 402).

**QUESTION 9**

**Describe three different methods used in the determination of the mass of a galaxy cluster.**

## QUESTION 9

**Describe three different methods used in the determination of the mass of a galaxy cluster.**

We know that galaxies are not randomly distributed in space, but are rather preferentially located in galaxy groups and clusters. The distinction between a group and a cluster is based on the number of constituents with groups tending to have less than 50 members in a region 2 Mpc across whereas clusters can have anywhere from 50 members (poor cluster) to several thousand (rich cluster) within a region 8 across. A typical galaxy group has a velocity dispersion of 150 km/s amongst its members, a total mass of the order  $10^{13} M_{\odot}$  and mass-to-light ratios of about  $260 M_{\odot}/L_{\odot}$ . Galaxy clusters, on the other hand, have velocity dispersions around 1000 km/s, masses of the order  $10^{15} M_{\odot}$ , and mass-to-light ratios around  $400 M_{\odot}/L_{\odot}$  (Carroll & Ostlie 2007, pg. 1058).

### VIRIAL THEOREM

The mass of a galaxy cluster can be approximated using the virial theorem. The use of the virial theorem is justified by noting that the dynamical time-scale of clusters ( $t_{\text{cross}}$ ) is much shorter than the age of the universe. The dynamical time-scale is determined by the amount of time it would take a typical galaxy to traverse the length of the cluster once:  $t_{\text{cross}} \sim R/\sigma_v$  which is roughly  $10^9$  years for a velocity dispersion of 1000 km/s. Since  $t_{\text{cross}} \ll t_0$  we conclude that clusters are gravitationally bound (otherwise they would have dissolved on a timescale  $t_{\text{cross}}$ ) and also that the virial theorem applies so that

$$2E_{\text{kin}} = -E_{\text{pot}} \quad \Rightarrow \quad \sum_i m_i v_i^2 = \sum_{i < j} \frac{Gm_i m_j}{r_{ij}}, \quad (183)$$

where indices label galaxy members of the cluster. If we take  $M$  to be the total mass of the cluster then we define the velocity dispersion weighted by mass and the gravitational radius to be, respectively:

$$\langle v^2 \rangle \equiv \frac{1}{M} \sum_i m_i v_i^2 \quad r_G \equiv M^2 \left( \sum_{i < j} \frac{m_i m_j}{r_{ij}} \right)^{-1}. \quad (184)$$

Plugging these into equation (183) and solving for  $M$  yields the mass estimate

$$\frac{M}{2} \langle v^2 \rangle = \frac{GM^2}{r_G} \quad \Rightarrow \quad M = \frac{r_G \langle v^2 \rangle}{G}. \quad (185)$$

This derivation used the three-dimensional separation vectors  $r_i$  between member galaxies and the cluster centre, which are, of course, not observable quantities. To make the mass estimate in equation (185) more practical we transform to projected coordinates. If the galaxies positions and velocities are uncorrelated (i.e., isotropic velocity distribution) then we have that

$$\langle v^2 \rangle = 3\sigma_v^2 \quad \text{and} \quad r_G = \pi/2R_G \quad \text{with} \quad R_G \equiv M^2 \left( \sum_{i < j} \frac{m_i m_j}{R_{ij}} \right)^{-1}, \quad (186)$$

where  $R_{ij}$  denotes the projected separation between the  $i^{\text{th}}$  and  $j^{\text{th}}$  galaxies (Schneider 2002, pg. 234). The parameters  $\sigma_v$  and  $R_G$  are direct observables and allow us to write the mass estimate of the cluster as

$$M = \frac{3\pi R_G \sigma_v^2}{2G} = 1.1 \times 10^{15} M_{\odot} \left( \frac{\sigma_v}{1000 \text{km/s}} \right)^2 \left( \frac{R_G}{1 \text{Mpc}} \right). \quad (187)$$

This was the original technique used by Fritz Zwicky in 1933 to arrive at a large mass-to-light ratio of the Coma cluster. From this he concluded that clusters must contain considerably more mass than indicated by their stars for else they would have dispersed long ago (Schneider 2002, pg. 234).

Given the above line of argument, the question of course arises as to whether the application of the virial theorem is still justified if the main fraction of mass is not contained in galaxies. The derivation remains valid in this form as long as the spatial distribution of galaxies follows the total mass distribution. The dynamical mass determination can be affected by an anisotropic velocity distribution of the cluster galaxies and by the possibly non-spherical cluster mass distribution. In both cases, projection effects, which are dealt with relatively easily in the spherically-symmetric case, obviously become more complicated. This is also one of the reasons for the necessity to consider alternative methods of mass determination (Schneider 2002, pg. 235).

### X-RAY EMISSION

Measurements in the late 70's revealed that galaxy clusters are strong sources of X-ray radiation which is emitted by a hot gas ( $T \sim 10^7 K$ ) located between the galaxies. This gas forms part of the intercluster medium (ICM) that also contains a diffuse, irregular distribution of stars. It turns out that the mass of clusters of galaxies consist of roughly 3% stars, 15% intergalactic gas, and 80% dark matter (Schneider 2002, pg. 248).

We will now analyze how we can arrive at a mass estimate of the cluster based on X-ray observations of the intercluster gas. The sound speed of the gas is given by

$$c_s = \sqrt{\frac{P}{\rho_g}} = \sqrt{\frac{k_B T}{\mu m_p}} \sim 1000 \text{km/s}, \quad (188)$$

where  $P$  and  $\rho_g$  are the gas pressure and density and with  $\mu = 0.63$  and  $T = 10^7 K$  (Schneider 2002, pg. 246). The sound-crossing time for the cluster is thus  $t_{sc} = 2R/c_s \sim 10^8$  years which is considerably shorter than the lifetime of the cluster (approximately the age of the universe).

Since the sound-crossing time defines the time-scale on which deviations from the pressure equilibrium are evened out, the gas can be in hydrostatic equilibrium. In this case we have that  $\nabla P = -\rho_g \nabla \Phi$  where  $\Phi$  is the gravitational potential. This relation describes how the gravitational force is balanced by the pressure force. In a spherically symmetric case it leads to the result

$$\frac{1}{\rho_g} \frac{dP}{dr} = -\frac{d\Phi}{dr} = -\frac{GM(r)}{r^2}, \quad (189)$$

where  $M(r)$  is the total mass contained within the radius  $r$  which contains contributions from all forms of matter as this is how the potential  $\Phi$  is determined. Plugging  $P = nk_B T = \rho_g k_B T / (\mu m_p)$  into this equation yields

$$M(r) = -\frac{k_B Tr^2}{G\mu m_p} \left( \frac{d \ln \rho_g}{dr} + \frac{d \ln T}{dr} \right). \quad (190)$$

This result tells us that we can determine the mass profile  $M(r)$  by measuring the radial profiles of  $\rho_g$  and  $T$ . Of course, the method by which this must be accomplished is through measurements of the projected surface brightness which is a function of the emissivity of the gas. The emissivity, in turn, is determined by the process of thermal bremsstrahlung which is causing the X-ray emissions. Therefore, measuring the surface brightness of the intercluster X-ray emissions, inverting this to obtain emissivity, fitting this to bremsstrahlung emissivity to determine  $T$  and  $\rho_g$ , and finally using equation (190) allows us to arrive at a mass estimate for the given cluster (Schneider 2002, pg. 247).

In examining the IGM, we have assumed hydrostatic equilibrium, but we have disregarded the fact that the gas cools by its emission, thus it will lose internal energy. For this reason, once established, a hydrostatic equilibrium cannot be maintained over arbitrarily long times. To decide whether this gas cooling is important for the dynamics of the system, the cooling time-scale needs to be considered. This cooling time turns out to be very long,  $t_{cool} \sim 10^{11}$  yr, for gas cooling through free-free emission. Hence, the cooling time is longer than the Hubble time nearly everywhere in the cluster, which allows a hydrostatic equilibrium to be established (Schneider 2002, pg. 249).

In the centres of clusters, however, the density may be sufficiently large to yield  $t_{cool} \lesssim H_0^{-1}$ . Here, the gas can cool quite efficiently, by which its pressure decreases. This then implies that, at least close to the centre, the hydrostatic equilibrium can no longer be maintained. To re-establish pressure equilibrium, gas needs to flow inwards and is thus compressed. Hence, an inward-directed mass flow should establish itself. The corresponding density increase will further accelerate the cooling process. Since the emissivity of a relatively cool gas increases with decreasing temperature, this process should then very quickly lead to a strong compression and cooling of the gas in the centres of dense clusters. In parallel to this increase in density, the X-ray emission will strongly increase, because  $\epsilon^{ff} \propto n_e^2$ . As a result of this process, a radial density and temperature distribution should be established with a nearly unchanged pressure distribution. These so-called **cooling flows** have indeed been observed in the centres of massive clusters, in the form of a sharp central peak in X-ray emission (Schneider 2002, pg. 249).

#### GRAVITATIONAL LENSING

The third independent method of estimating the mass of a galaxy cluster is to take advantage of any gravitational lensing that it produces. This effect typically shows up as large luminous arcs centred around the cluster. This method is particularly nice since gravitational light deflection is independent of the nature and state of the deflecting matter. It thus requires neither any assumptions about the state of equilibrium of the matter nor relations between the luminous and dark matter.

In order to arrive at a mass estimate for gravitational lensing by a galaxy cluster, there are a couple of simplifying assumptions that need to be made. Firstly, we assume that the gravitational field producing the lensing is weak. This is a good assumption since the strength of a gravitational field can be quantized using the virial theorem: if a mass distribution is in virial equilibrium then  $v^2 \sim \Phi$  where  $\Phi$  is the gravitational potential. Weak fields are then characterized by  $v^2/c^2 \ll 1$  and with  $v \sim 1000$  km/s for galaxies in clusters this is well justified (Schneider 2002, pg. 122). Next, we assume that the deflecting mass has a small extent along the line-of-sight, as compared to the distances between the observer and the lens and the lens and the source. A system satisfying this configuration is known as a geometrically thin lens. Since clusters are typically 8 Mpc across, this assumption is justified for sufficiently distant clusters.

For continuous mass distributions we can divide the lens into mass elements of mass  $dm = \Sigma(\vec{\xi}) d^2 \vec{\xi}$ , where  $\Sigma(\vec{\xi})$  describes the surface mass density of the lens at position  $\vec{\xi}$  obtained by projecting the three-dimensional mass density  $\rho$  along the line-of-sight. The simplest models for gravitational lenses are those that are axially symmetric so that  $\Sigma(\vec{\xi}) = \Sigma(\xi)$  where  $\xi = |\vec{\xi}|$ . In this case the deflection angle  $\alpha$  is directed radially inwards and we have that

$$\alpha = \frac{4GM(\xi)}{c^2 \xi}, \quad (191)$$

where  $M(\xi)$  is the mass contained within radius  $\xi$  (Schneider 2002, pg. 122).

The simplest model for a galaxy cluster as a lens is the singular isothermal sphere (SIS). This has a mass density profile that goes like  $\rho = \sigma_v^2 / 2\pi G r^2$  so that the surface mass density along the line-of-sight is  $\Sigma(\xi) = \sigma_v^2 / 2G\xi$ . This yields a projected mass

of  $M(\xi) = \pi\sigma_v^2\xi/G$  which yields the deflection angle

$$\theta_E = 4\pi \frac{\sigma_v^2}{c^2} \frac{D_{ds}}{D_s}, \quad (192)$$

where  $D_{ds}$  and  $D_s$  are the distances between the lens and the source and the lens and the observer, respectively.  $\theta_E$  is known as the Einstein angle of the SIS and has a characteristic scale of

$$\theta_E = 29'' \left( \frac{\sigma_v}{1000 \text{km/s}} \right)^2 \left( \frac{D_{ds}}{D_d} \right). \quad (193)$$

Very high magnifications and distortions of images can only occur very close to the Einstein radius. This yields an immediate mass estimate for a galaxy cluster by assuming that the luminous arc seen around the cluster is at an angular scale of the Einstein radius from the centre of the cluster. This yields

$$M(\theta_E) = \pi(D_d\theta_E)^2\Sigma_{cr}, \quad (194)$$

where  $\Sigma_{cr}$  is the critical surface mass density given by equation (3.52) of Schneider (2002).

Of course, clusters are not generally spherically symmetric so that the separation of the arc from the cluster's centre can deviate from the Einstein radius. Instead, models with asymmetric mass distributions predict a variety of configurations for arcs and positions of multiple images. If several arcs are discovered in a cluster or several images of the source of an arc, then detailed mass models of the cluster can be investigated (Schneider 2002, pg. 261). This can lead to very well-determined mass profiles of the cluster.

**QUESTION 10**

**What is the density-morphology relation for galaxies? How is that related to what we know about the relationship between galaxy density and star formation rates in galaxies?**

## QUESTION 10

**What is the density-morphology relation for galaxies? How is that related to what we know about the relationship between galaxy density and star formation rates in galaxies?**

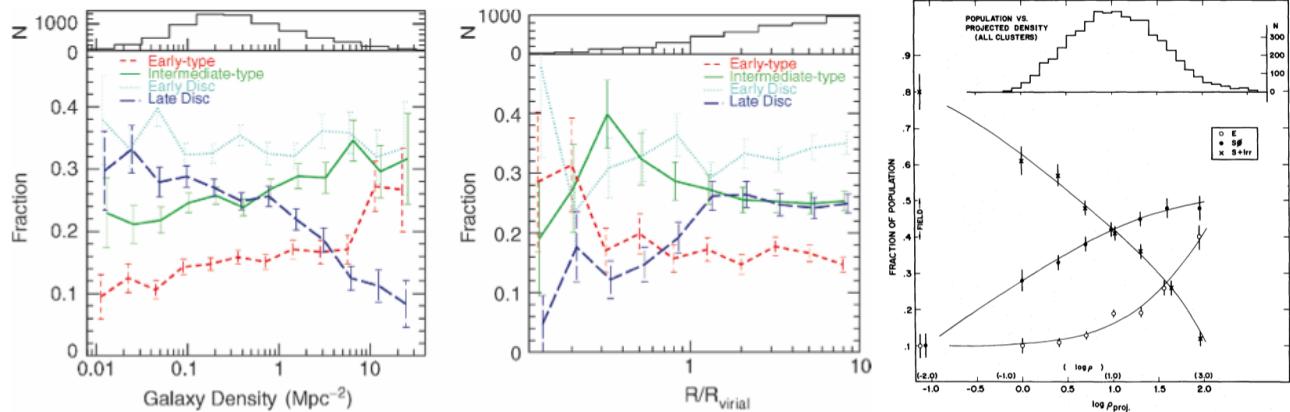


FIG. 37.— (left) The number fraction of galaxies of different morphologies is plotted as a function of the local galaxy density (left panel), and for galaxies in clusters as a function of the distance from the cluster centre, scaled by the corresponding virial radius (right panel). Galaxies have been divided into four different classes. “Early-types” contain mainly ellipticals, “intermediates” are mainly S0 galaxies, “early and late disks” are predominantly Sa and Sc spirals, respectively. In both representations, a clear dependence of the galaxy mix on the density or on the distance from the cluster centre, respectively, is visible. In the histograms at the top of each panel, the number of galaxies in the various bins is plotted. Image taken from Schneider (2002). (right) The fraction of elliptical (E), lenticular (S0), and spiral plus irregular (S+Irr) galaxies as a function of the projected density in units of galaxies per  $\text{Mpc}^2$ . The data shown are for a collection of cluster members and field galaxies. The upper histogram shows the number distribution of the galaxies over the bins of projected density. Image taken from Dressler (1980).

The **density-morphology relation** is the observation that in denser environments, such as galaxy clusters, there is a higher ratio of elliptical galaxies versus spiral compared with field galaxies. Also, the higher the spatial density of galaxy in a given region, the stronger this effect is. Another effect is that in a given cluster, as one moves away from the centre the fraction of spiral galaxies increases, compared with the centre fraction (Yevgeni). These features are displayed in Figure 37. The density-morphology relation is also seen in galaxy groups. The fraction of late-type galaxies decreases with increasing group mass. Furthermore, the fraction of early-type galaxies increases with decreasing distance from the group center, as is also the case in clusters (Schneider 2002, pg. 240).

### INTERPRETATION

A closer examination of the middle panel of Figure 37 may provide a clue as to what physical processes are responsible for the dependence of the morphological mix on the local number density. Three different regimes in radius can be identified: for  $R \gtrsim R_{\text{vir}}$ , the fraction of the different galaxy types remains basically constant. In the intermediate regime,  $0.3 \lesssim R/R_{\text{vir}} \lesssim 1$ , the fraction of S0 galaxies strongly increases inwards, whereas the fraction of late-type spirals decreases accordingly. This result is compatible with the interpretation that in the outer regions of galaxy clusters spirals lose gas and these galaxies then transform into passive S0 galaxies (Schneider 2002, pg. 240). Possible explanations for the loss of gas come from galaxy-galaxy merger, ram pressure stripping, and gas evaporation. Galaxy mergers can strip gas from galaxies through the addition of thermal energy that heats the gas above the virial temperature of the galaxy. In addition, any clumping of gas that may occur during the merger will spark star formation, consuming much of the remaining gas. Ram pressure stripping is caused by the fast motion of a galaxy through the ICM, which causes a pressure front to build up at the leading edge of the galaxy. Finally, gas evaporation is induced by the pressure gradient between the surrounding hot ICM and the cool internal ISM (Yevgeni).

Furthermore, we see that below  $R \lesssim 0.3R_{\text{vir}}$ , the fraction of S0 galaxies decreases strongly, and the fraction of ellipticals increases substantially. In fact, the ratio of the number densities of S0 galaxies and ellipticals, for  $R \lesssim 0.3R_{\text{vir}}$ , strongly decreases as  $R$  decreases. This may hint at a morphological transformation in which S0 galaxies are turned into ellipticals, probably by mergers. Such gas-free mergers, also called “dry mergers”, may be the preferred explanation for the generation of elliptical galaxies. One of the nice properties of dry mergers is that such a merging process would not be accompanied by a burst of star formation, unlike the case of gas-rich collisions of galaxies. The existence of a population of newly-born stars in ellipticals would be difficult to reconcile with the generally old stellar population actually observed in these galaxies (Schneider 2002, pg. 240).

One clue as to the origin of the morphological transformation of galaxies in clusters, as a function of distance from the cluster centre, comes from the observation that the velocity dispersion of very bright cluster galaxies seems to be significantly smaller than that of less luminous ones. Assuming that the mass-to-light ratio does not vary substantially among cluster members, this then indicates that the most massive galaxies have smaller velocity dispersions. One way to achieve this trend in the course of cluster evolution is by dynamical interactions between cluster galaxies. Such interactions tend to “thermalize” the velocity distribution of galaxies, so that the mean kinetic energy of galaxies tends to become similar. This then causes more massive galaxies to become slower on average:  $v \propto m^{-1/2}$ . If this interpretation holds, then the density-morphology relation may be

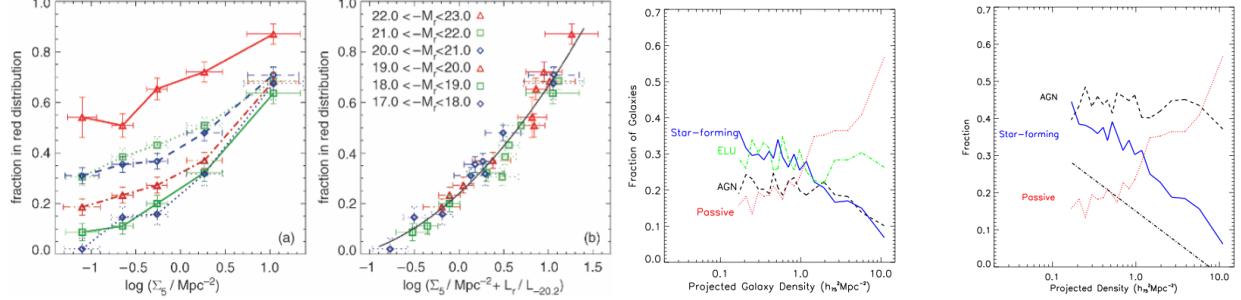


FIG. 38.— (left) Left: the fraction of galaxies in the red distribution is shown as a function of  $\Sigma_5$ , an estimator of the local galaxy number density based on the projected distance of the fifth-nearest spectroscopically confirmed neighbour galaxy within  $\pm 1000 \text{ km s}^{-1}$ . Different symbols correspond to different luminosity bins, as indicated. Right: the same red fraction is plotted against a combination of the local galaxy density  $\Sigma_5$  and the luminosity of the galaxy. Image taken from Schneider (2002). (right) The fraction of galaxies as a function of the different spectral classifications, in our pseudo volume-limited sample, as a function of density. We see an increase in the fraction of passive galaxies with density, and a decrease in the fraction of star-forming galaxies with density. This is the SFR-density and morphology-density relations. Image taken from Miller (2004).

attributed to these dynamical interactions, rather than to things like ram-pressure stripping of the ISM as the galaxies move through the ICM (Schneider 2002, pg. 241).

Recently, Bekki & Couch (2011) argued that galaxy interactions and mergers are the major processes responsible for transforming spiral galaxies into S0's. Recent observational studies on the evolution of the S0 fraction in groups and clusters have found that S0 evolution is significantly more dramatic in groups than in clusters. These observations suggest that cluster-related physical processes such as ram pressure stripping are not responsible for the formation of the majority of S0's. Also, it is well known that galaxy bulges in S0's are systematically more luminous than those in spirals, which implies that bulges in spirals need to grow significantly in order to be transformed into S0's: disc fading alone, due to the truncation of star formation, cannot be the main mechanism of S0 formation. Thus, galaxy interactions and merging, which are highly likely to occur in groups (**Shouldn't this be more likely in clusters?**), are a promising mechanism for S0 formation. Multiple slow tidal encounters with group member galaxies can dynamically heat up the discs of spirals so that their initially thin discs can be transformed into thick ones during the morphological transformation from spirals into S0's. Such tidal interaction can also trigger moderately strong starbursts in the inner regions of their bulges so that bulges can grow significantly during S0 formation.

#### COLOUR-DENSITY RELATION

The classification of galaxies by morphology, given by the Hubble classification scheme, has the disadvantage that morphologies of galaxies are not easy to quantify. An alternative to classify galaxies is provided by their colour. We expect that early-type galaxies are red, whereas late-type galaxies are considerably bluer. Using photometric measurements and spectroscopy from the SDSS the colours and absolute magnitudes of  $\sim 10^5$  low-redshift galaxies has been studied. From a colour-magnitude diagram of these colours we find a bimodal distribution: one peak at high luminosities and red colour, the other at significantly fainter absolute magnitudes and much bluer colour. It appears that the galaxies are distributed at and around these two density peaks, hence galaxies tend to be either luminous and red, or less luminous and blue (Schneider 2002, pg. 120). It is therefore useful to make a similar comparison to see how galaxy colour trends with environment; the result is shown in Figure 38. The fraction of red galaxies increases towards higher local number density, and the relative increase is stronger for the less luminous galaxies. Surprisingly, the fraction of galaxies in the red sample seems to be a function of a combination of the local galaxy density and the luminosity of the galaxy.

#### SFR-DENSITY RELATION

The star formation rates of galaxies can be indirectly determined in various ways. The most successful methods rely on correlations of the star formation rate with measurements of the FIR luminosity (which arises from dust heated by star formation), the radio luminosity (which results from synchrotron emission associated with supernovae) and indicators that are sensitive to the ionizing flux from massive stars. The last category includes measurements of the UV continuum and the fluxes of nebular emission lines like  $\text{H}\alpha$  (Gómez et al. 2003). By measuring the SFR of a large collection of galaxies and comparing this to their local density (e.g., through the fifth-nearest neighbour approach described above), we can learn about any SFR-density relation that emerges; this is shown in Figure 38.

As was mentioned above galaxies in denser environments will interact with each other in many different ways (e.g., mergers, tidal interactions, ram-pressure stripping, etc.). These interactions result in a loss of gas from their parent galaxies, therefore suppressing subsequent star formation within them. In the short run, star formation may be induced due to clumping of gas in the galaxy, but once this burst is complete, further star formation is quenched by the lack of gas. This explains why elliptical galaxies have very old stellar populations, and S0 galaxies have no star formation in them (Yevgeni).

**QUESTION 11**

**Draw the spectral energy distribution (SED) of a galaxy formed by a single burst of star formation at the ages of 10 Myrs, 2Gyrs, and 10 Gyr.**

### QUESTION 11

**Draw the spectral energy distribution (SED) of a galaxy formed by a single burst of star formation at the ages of 10 Myrs, 2Gyrs, and 10 Gyr.**

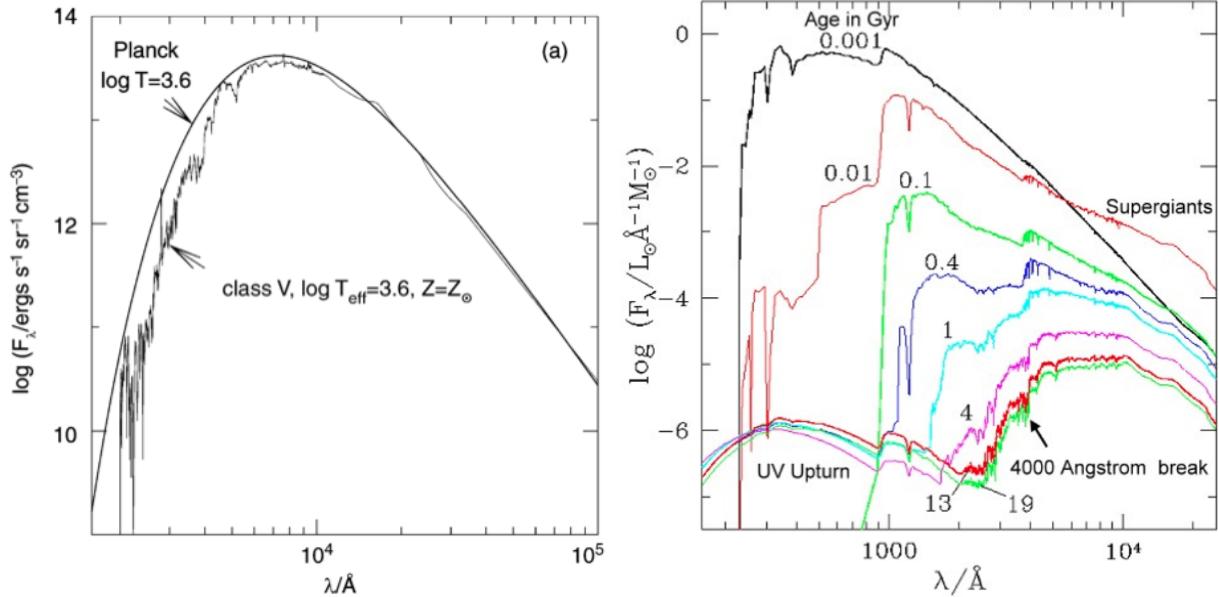


FIG. 39.— (left) Comparison of the spectrum of a MS star with a blackbody spectrum of equal effective temperature. The opacity of the stellar atmosphere causes clear deviations from the Planck spectrum in the UV/optical. Image taken from Schneider (2002). (right) Spectral evolution of stellar populations as predicted by isochrone synthesis models for instantaneous starburst. This model uses the Salpeter IMF. Image taken from Bob's AST2040 Lecture Notes.

The theory of **population synthesis** aims at interpreting the spectrum of galaxies as a superposition of stellar spectra. We have to take into account the fact that the distribution of stars changes over time, which means that the spectral distribution of the population also changes in time. The spectral energy distribution of a galaxy thus reflects its history of star formation and stellar evolution (Schneider 2002, pg. 132).

Let  $S_{\lambda,Z}(t')$  be the emitted energy per wavelength and time interval, normalized to an initial total mass of  $1 M_\odot$ , emitted by a group of stars of initial metallicity  $Z$  and age  $t'$ . The function  $S_{\lambda,Z(t-t')}(t')$ , which describes this emission at any point  $t$  in time, accounts for the different evolutionary tracks of the stars in the HR diagram. It also accounts for their initial metallicity (i.e., at time  $t-t'$ ), where the latter follows from the chemical evolution of the ISM of the corresponding galaxy. Then the total spectral luminosity of this galaxy at a time  $t$  is given by

$$F_\lambda(t) = \int_0^t dt' \psi(t-t') S_{\lambda,Z(t-t')}(t'), \quad (195)$$

thus by the convolution of the star-formation rate with the spectral energy distribution of the stellar population (Schneider 2002, pg. 133).

In order to compute  $S_{\lambda,Z(t-t')}$ , models for stellar evolution and stellar atmospheres are needed. These can then be used to determine the positions of stars at equal time on the HR diagram; known as isochrone. The spectrum  $S_{\lambda,Z(t-t')}$  is then the sum over all spectra of the stars on an isochrone; see Figure 39.

In the beginning, the spectrum and luminosity of a stellar population are dominated by the most massive O and B stars, which emit intense UV radiation. But after  $\sim 10$  Myr, the flux below 1000 Å is diminished significantly, and after 100 Myr, it hardly exists any more. At the same time, the flux in the NIR increases because the massive stars evolve into red supergiants (Schneider 2002, pg. 133).

For  $100 \text{ Myr} \lesssim t \lesssim 1 \text{ Gyr}$ , the emission in the NIR remains high, whereas short-wavelength radiation is more and more diminished. After  $\sim 1 \text{ Gyr}$ , RGB stars account for most of the NIR production. After  $\sim 3 \text{ Gyr}$  the UV radiation increases again due to blue stars on the HB into which stars evolve after the AGB phase, and due to WDs which are hot when they are born. Between an age of 4 and 13 Gyr, the spectrum of a stellar population evolves fairly little (Schneider 2002, pg. 133).

As time proceeds absorption lines, and in particular the break at 4000 Å, are becoming more prominent. The 4000 Å break is due to the accumulation of absorption lines of ionized metals; it gets larger with age due to increasing temperatures and correspondingly larger opacity, as well enhanced metallicity. The 4000 Å break is near the 3646 Å Balmer break, which marks the termination of the H Balmer series and is strongest in A stars. This break does not change strength monotonically with age, peaking in strength for a stellar population of  $\sim 1$  Gyr. Similarly, at 912 Å there is a Lyman break, past which photoionization

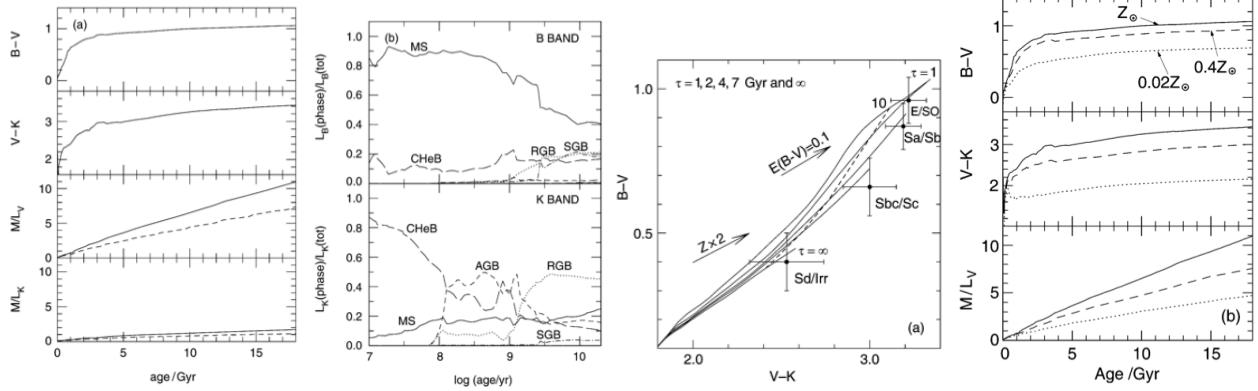


FIG. 40.— (a) For the same stellar population as in Figure 39, the upper two graphs show the colours  $B-V$  and  $V-K$  as a function of age. The lower two graphs show the mass-to-light ratio in two colour bands in Solar units. The solid curves show the total  $M/L$  (i.e., including the mass that is later returned into the ISM), whereas the dashed curves show the  $M/L$  of the stars itself. (b) The fraction of  $B$ - (top) and  $K$ -luminosity (bottom) contributed by stars in their different phases of stellar evolution (CHeB: core helium burning stars; SGB: subgiant branch). (c) Evolution of colours between  $0 \leq t \leq 20$  Gyr for a stellar population with star-formation rate given by equation (196), for five different values of the characteristic time-scale  $\tau$  ( $\tau = \infty$  is the limiting case for a constant star-formation rate) – Galactic centre see solid curves. The typical colours for four different morphological types of galaxies are plotted. For each  $\tau$ , the evolution begins at the lower left (i.e., as a blue population in both colour indices). In the case of constant star formation, the population never becomes redder than Irr's; to achieve redder colours,  $\tau$  has to be smaller. The dashed line connects points of  $t = 10$  Gyr on the different curves. Here, a Salpeter IMF and Solar metallicity was assumed. The shift in colour obtained by doubling the metallicity is indicated by an arrow, as well as that due to an extinction coefficient of  $E(B-V) = 0.1$ ; both effects will make galaxies appear redder. (d) The dependence of colours and  $M/L$  on the metallicity of the population. Images taken from Schneider (2002).

of H atoms creates significant absorption. This feature can be seen as early as 10 Myr, and becomes less prominent by 1 Gyr as fewer and fewer UV photons are created by the stellar population (Charles).

#### COLOUR EVOLUTION

It is practically far more difficult to obtain spectra of galaxies than photometric observations, especially in deep-field surveys. Quite often, then, theoretically calculated spectra are reduced to colours by multiplying them with the transmission curves of colour filters, and integrating over wavelength. Hence the spectral evolution implies a colour evolution, as is illustrated in Figure ???. For a young stellar population the colour evolution is rapid and the population becomes redder, again because the hot blue stars have a higher mass and thus evolve quickly in the HR diagram. For the same reason, the evolution is faster in  $B-V$  than in  $V-K$ . It should be mentioned that this colour evolution is also observed in star clusters of different ages. The mass-to-light ratio also increases with time because  $M$  remains constant while  $L$  decreases (Schneider 2002, pg. 135).

As shown in Figure ???, the blue light of a stellar population is always dominated by main-sequence stars, although at later stages a noticeable contribution also comes from HB stars. The NIR radiation is first dominated by stars burning helium in their centre (this class includes the supergiant phase of massive stars), later by AGB stars, and after  $\sim 1$  Gyr by red giants. MS stars never contribute more than 20% of the light in the K-band. The fact that  $M/L_K$  varies only little with time implies that the NIR luminosity is a good indicator for the total stellar mass: the NIR mass-to-light ratio is much less dependent on the age of the stellar population than that for bluer filters (Schneider 2002, pg. 135).

#### ADVANCED MODELLING

There are very few places in the universe where one can view an isolated stellar population created from a single burst of star formation – certainly, galaxies look more complex. In particular, we expect that the star-formation rate decreases over time because more and more matter is bound in stars and thus no longer available to form new stars. A standard model is to assume that the SFR is exponentially decreasing in time:

$$\psi(t) = \frac{1}{\tau} \exp \left[ -\frac{t-t_f}{\tau} \right] H(t-t_f), \quad (196)$$

where  $H(t)$  is the Heaviside function,  $t_f$  is the time of star formation, and  $\tau$  is the characteristic duration of star formation. The SEDs for a star formation history governed by equation (196) and for a constant star formation are shown in Figure 41 and can be contrasted to Figure 41.

From Figure ?? we find that the colours of the population depend strongly on  $\tau$ . Specifically, galaxies do not become very red if  $\tau$  is large because their star-formation rate, and thus the fraction of massive blue stars, does not decrease sufficiently. The colours of Sc spirals, for example, are not compatible with a constant star-formation rate – except if the total light of spirals is strongly reddened by dust absorption (but there are good reasons why this is not the case). To explain the colours of early-type galaxies we need  $\tau \lesssim 4$  Gyr. In general, one deduces from these models that a substantial evolution to redder colours occurs for  $t \gtrsim \tau$ . Since the luminosity of a stellar population in the blue spectral range decreases quickly with the age of the population, whereas increasing age affects the red luminosity much less, we conclude that the spectral distribution of galaxies is mainly determined by the ratio of the star-formation rate today to the mean star-formation rate in the past,  $\psi(\text{today})/\langle \psi \rangle$  (Schneider 2002, pg. 136).

Another physical quantity that must be considered is the metallicity  $Z$ . A small value of  $Z$  results in a bluer colour and a smaller  $M/L$  ratio. The age and metallicity of a stellar population are degenerate in the sense that they both affect the colour of a

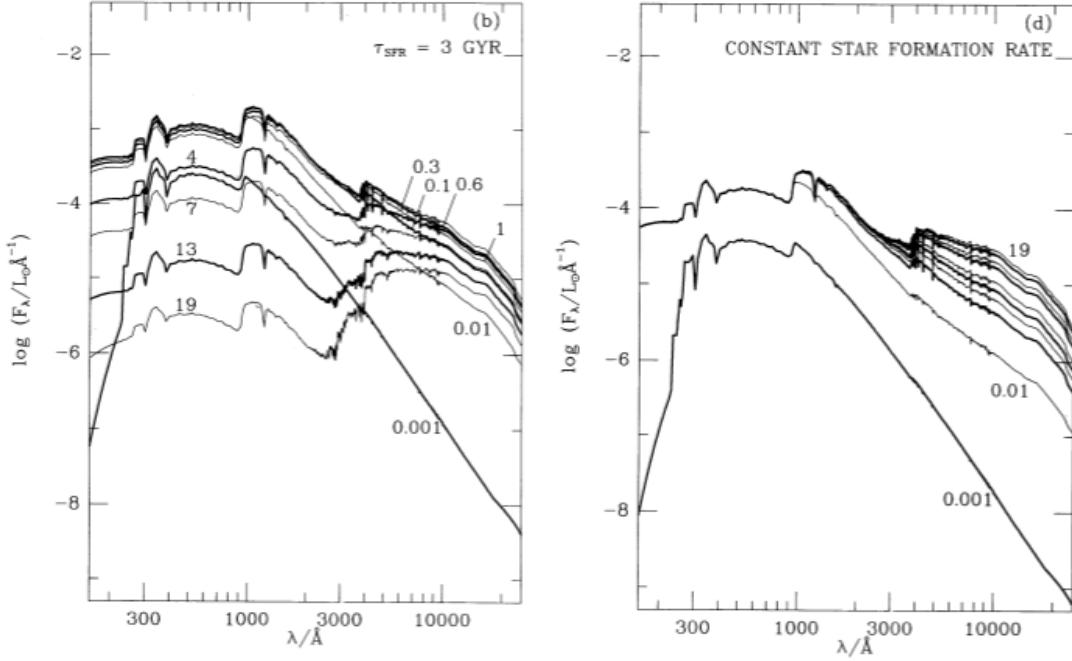


FIG. 41.— Spectral evolution of stellar populations as predicted by isochrone synthesis models for a star formation history governed by equation (196) with  $\tau = 3$  Gyr (left panel) and a constant star formation rate (right panel). Both models assume a Salpeter IMF. Image taken from Bruzual A. & Charlot (1993).

population. The age estimate of a population from colour will therefore strongly depend on the assumed value for  $Z$ . However, this degeneracy can be broken by taking several colours, or information from spectroscopy, into account. Intrinsic dust absorption will also change the colours of a population. This effect cannot be easily accounted for in the models because it depends not only on the properties of the dust but also on the geometric distribution of dust and stars. For example, it makes a difference whether the dust in a galaxy is homogeneously distributed or concentrated in a thin disk. Empirically, it is found that galaxies show strong extinction during their active phase of star formation, whereas normal galaxies are presumably not much affected by extinction, with early-type galaxies affected the least (Schneider 2002, pg. 136).

Besides stellar light, the emission by H II regions also contributes to the light of galaxies. It is found, though, that after  $\sim 10$  Myr the emission from gas nebulae only marginally contributes to the broad-band colours of galaxies. However, this nebular radiation is the origin of emission lines in the spectra of galaxies. Therefore, emission lines are used as diagnostics for the star-formation rate and the metallicity in a stellar population (Schneider 2002, pg. 136).

#### REAL GALAXIES

In general, the later the Hubble type, (1) the bluer the overall spectral distribution, (2) the stronger the emission lines, (3) the weaker the absorption lines, and (4) the smaller the 4000 Å break in the spectra. From the above discussion, we would also expect these trends if the Hubble sequence is considered an ordering of galaxy types according to the characteristic age of their stellar population or according to their SFR. Elliptical and S0 galaxies essentially have no star-formation activity, which renders their SED dominated by red stars. Furthermore, in these galaxies there are no H II regions where emission lines could be generated. The old stellar population produces a pronounced 4000 Å break, which corresponds to a jump by a factor of  $\sim 2$  in the spectra of early-type galaxies. It should be noted that the spectra of ellipticals and S0 galaxies are quite similar (Schneider 2002, pg. 137).

By contrast, Sc spirals and Irr galaxies have a spectrum which is dominated by emission lines, where the Balmer lines of H as well as N and O lines are most pronounced. The relative strength of these emission lines are characteristic for H II regions, implying that most of this line emission is produced in the ionized regions surrounding young stars. For irregular galaxies, the spectrum is nearly totally dominated by the stellar continuum light of hot stars and the emission lines from H II regions, whereas clear contributions by cooler stars can be identified in the spectra of Sc spiral galaxies. The spectra of Sa and Sb galaxies form a kind of transition between those of early-type galaxies and Sc. Their spectra can be described as a super-position of an old stellar population generating a red continuum and a young population with its blue continuum and its emission lines. This can be seen in connection with the decreasing contribution of the bulge to the galaxy luminosity towards later spiral types (Schneider 2002, pg. 138).

**QUESTION 12**

**What are Lyman-Break Galaxies and how do we find them?**

## QUESTION 12

### What are Lyman-Break Galaxies and how do we find them?

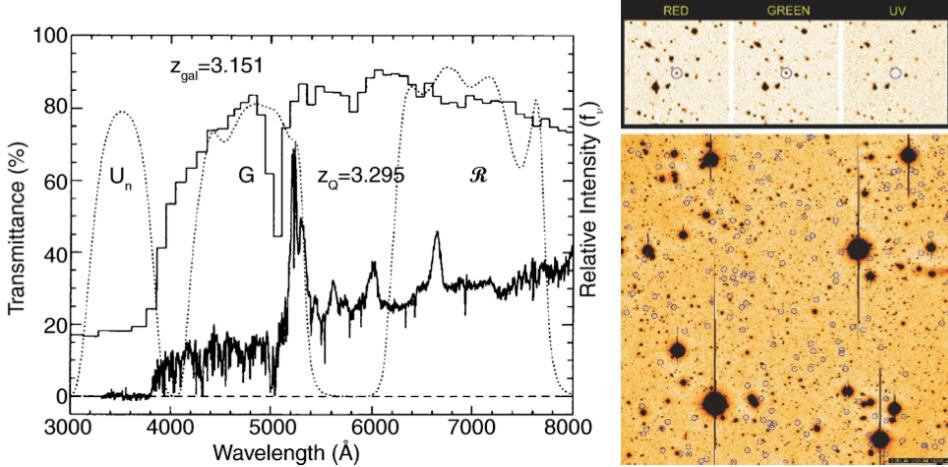


FIG. 42.— (left) Principle of the Lyman-break method. The histogram shows the synthetic spectrum of a galaxy at  $z = 3.15$ , generated by models of population synthesis; the spectrum belongs to a QSO at slightly higher redshift. Clearly, the decline of the spectrum at  $\lambda \leq 912(1+z)$  Å is noticeable. The three dotted curves are the transmission curves of three broad-band filters, chosen such that one of them ( $U_n$ ) blocks all photons with wavelengths above the Lyman-break. The colour of this galaxy would then be blue in  $G - R$ , and very red in  $U_n - G$ . (right) Top panel: a  $U$ -band drop-out galaxy. It is clearly detected in the two redder filters, but vanishes almost completely in the  $U$ -filter. Bottom panel: in a single CCD frame, a large number of candidate Lyman-break galaxies are found. They are marked with circles here; their density is about 1 per square arcminute. Images taken from Schneider (2002).

Identifying objects at high redshift is tricky business since they tend to be very faint and difficult to resolve spectroscopically. A major breakthrough in finding high- $z$  objects is the so-called **Lyman-break method**. Since hydrogen is so abundant and its ionization cross-section so large, one can expect that photons with  $\lambda < 912$  Å are very heavily absorbed by neutral hydrogen in its ground state. Therefore, photons with  $\lambda < 912$  Å have a low probability of escaping from a galaxy without being absorbed. Intergalactic absorption also contributes, as evidenced by the Ly $\alpha$  forest in QSO spectra. The intergalactic gas absorbs a large fraction of photons emitted by a high-redshift source at  $\lambda < 1216$  Å, and virtually all photons with a rest-frame wavelength  $\lambda \lesssim 912$  Å. The strength of this absorption increases with increasing redshift. Combining these facts, it is concluded that spectra of high-redshift galaxies should display a distinct feature – a “break” – at  $\lambda = 1216$  Å. Furthermore, radiation with  $\lambda \lesssim 912$  Å should be strongly suppressed by intergalactic absorption, as well as by absorption in the ISM of the galaxies themselves, so that only a very small fraction of these ionizing photons will reach us (Schneider 2002, pg. 357).

From this, a strategy for the detection of galaxies at  $z \gtrsim 3$  emerges. We consider three broad-band filters with central wavelengths  $\lambda_1 < \lambda_2 < \lambda_3$ , where their spectral ranges are chosen to not (or only marginally) overlap. If  $\lambda_1 \lesssim (1+z) 912 \lesssim \lambda_2$ , a galaxy containing young stars should appear relatively blue as measured with the filters  $\lambda_2$  and  $\lambda_3$ , and be virtually invisible in the  $\lambda_1$ -filter: because of the absorption, it will drop out of the  $\lambda_1$ -filter; see Figure 42. For this reason, galaxies that have been detected in this way are called **Lyman-break galaxies (LBG)** or **drop-outs**. An example is shown in Figure 42. LBGs are primarily detected using UV and optical filters, but progress in infrared astronomy has allowed the use of this technique at higher redshifts using infrared filters (Wikipedia).

#### PHOTOMETRIC REDSHIFT

The Lyman-break technique is a special case of a method for estimating the redshift of galaxies (and QSOs) by multicolour photometry. This technique can be employed due to the spectral break at  $\lambda = 912$  Å and  $\lambda = 1216$  Å, respectively. Spectra of galaxies also show other characteristic features; as was known, the broad-band energy distribution is basically a superposition of stellar radiation. A stellar population of age  $\gtrsim 100$  Myr features a 4000 Å break because, due to a sudden change in the opacity at this wavelength, the spectra of most stars show such a break at about 4000 Å. Hence, the radiation from a stellar population at  $\lambda < 4000$  Å is less intense than at  $\lambda > 4000$  Å; this is the case particularly for early-type galaxies (Schneider 2002, pg. 363).

If we assume that the star-formation histories of galaxies are not too diversified, galaxies will not be located at an arbitrary location in a multidimensional colour diagram; rather, they should be concentrated in certain regions. In this context the 4000 Å break and the Ly $\alpha$  break play a central role, as is illustrated in Figure 43. Once these characteristic domains in colour space where (most of) the galaxies are situated are identified, the redshift of galaxies can be estimated solely from their observed colours, since they are functions of the redshift. The corresponding estimate is called the **photometric redshift** (Schneider 2002, pg. 363).

More precisely, a number of standard spectra of galaxies (so-called templates) are used, which are either selected from observed galaxies or computed by population synthesis models. Each of these template spectra can then be redshifted in wavelength, from which a K-correction results. For each template spectrum and any redshift, the expected galaxy colours are determined by integrating the spectral energy distribution, multiplied by the transmission functions of the applied filters, over wavelength. This

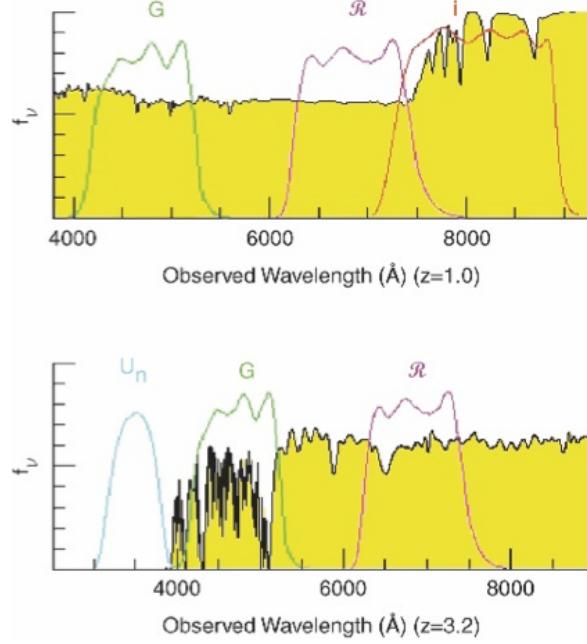


FIG. 43.— The bottom panel illustrates again the principle of the drop-out method, for a galaxy at  $z \sim 3.2$ . Whereas the Ly $\alpha$  forest absorbs part of the spectral flux between (rest-frame wavelength) 912 Å and 1216 Å, the flux below 912 Å vanishes almost completely. By using different combinations of filters (top panel), an efficient selection of galaxies at other redshifts is also possible. The example shows a galaxy at  $z = 1$  where the 4000-Å break is utilized, which occurs in stellar populations after  $\sim 10$  Myr and which is considered to be one of the most important features for the method of photometric redshift. Image taken from Schneider (2002).

set of colours can then be compared with the observed colours of galaxies, and the set best resembling the observation is taken as an estimate for not only the redshift but also the galaxy type. Depending on whether this break is identified as the Lyman-break or the 4000-Å break, the resulting redshift estimates will be very different. To break the corresponding degeneracy, a sufficiently large number of filters must be available to probe the spectral energy distribution over a wide range in wavelengths (Schneider 2002, pg. 363).

#### GALACTIC WINDS

The inferred high star-formation rates of LBGs (as seen through intense Ly $\alpha$  emission) implies an accordingly high rate of SN explosions. These release part of their energy in the form of kinetic energy to the ISM in these galaxies. This process will have two consequences. First, the ISM in these galaxies will be heated locally, which slows down (or prevents) further star formation in these regions. This thus provides a feedback effect for star formation which prevents all the gas in a galaxy from turning into stars on a very short time-scale, and is essential for understanding the formation and evolution of galaxies. Second, if the amount of energy transferred from the SNe to the ISM is large enough, a galactic wind may be launched which drives part of the ISM out of the galaxy into its halo. Evidence for such galactic winds has been found in nearby galaxies, for example from neutral hydrogen observations of edge-on spirals which show an extended gas distribution outside the disk. Furthermore, the X-ray corona of spirals is most likely linked to a galactic wind in these systems (Schneider 2002, pg. 360).

Indeed, there is now clear evidence for the presence of massive winds from LBGs. The spectra of LBGs often show strong absorption lines, which are blueshifted relative to the velocity of the emission lines in the galaxy. Such absorption can be produced by a wind moving out from the star-forming regions of the galaxy, so that its redshift is smaller than that of the emission regions. Characteristic velocities are  $\sim 200$  km s $^{-1}$ . Whereas these observations clearly show the presence of outflowing gas, it remains undetermined whether this is a fairly local phenomenon, restricted to the star-formation sites, or whether it affects the ISM of the whole galaxy (Schneider 2002, pg. 361).

**QUESTION 13**

**Draw a spectrum of a high-redshift quasar. What do quasar emission lines typically look like? Explain what we see in the spectrum at rest wavelengths bluer than 1216 Å?**

### QUESTION 13

Draw a spectrum of a high-redshift quasar. What do quasar emission lines typically look like? Explain what we see in the spectrum at rest wavelengths bluer than 1216 Å?

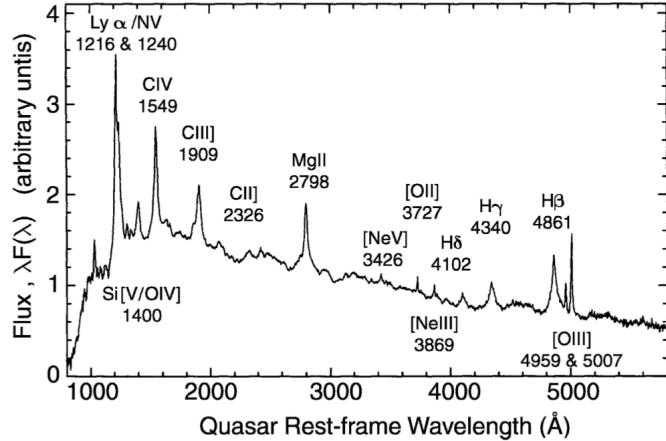


FIG. 44.— Combined spectrum of a sample of 718 individual quasars transformed into rest wavelengths of the sources. The most prominent emission lines are marked. Image taken from Schneider (2002).

Quasars belong to a class of objects known as active galactic nuclei (AGN). Quasars are the most luminous of the AGN and show significant emission in the full range of wavelengths from radio to X-ray. The luminosity of quasars can exceed the luminosities of normal galaxies by  $\sim 10^3$  and this luminosity originates from a compact core on the scale of a parsec (Schneider 2002, pg. 175).

Quasars were first discovered by identifying radio sources with point-like optical sources. The morphology of quasars in the radio regime is quite complicated and depends on the observed frequency. In most cases, the extended source is observed as a pair of radio lobes situated roughly symmetrically around the optical position of the quasar. The lobes are usually connected to the central core by jets and the entire radio source can extend up to a Mpc in scale (Schneider 2002, pg. 178).

#### CONTINUUM SPECTRUM

The continuum spectrum of a quasar can often be described, over a broad frequency range, by a power law of the form

$$S_\nu \propto \nu^{-\alpha}, \quad (197)$$

where  $\alpha$  is called the spectral index.  $\alpha = 0$  corresponds to a flat spectrum, whereas  $\alpha = 1$  describes a spectrum in which the same energy is emitted in every logarithmic frequency interval (Schneider 2002, pg. 178).

In the radio regime, observations find that the spectral index is  $\alpha = 0.7(0)$  for the extended radio source (compact core). The radiation observed in the radio band is also observed to be highly linearly polarized. This spectral form and high degree of polarization suggest that the radio emission is produced by synchrotron radiation of relativistic electrons. Synchrotron radiation is self-absorbed at low frequencies and in the limiting case of large optical depth produces a spectrum  $S_\nu \propto \nu^{2.5}$ . The spectral index of the extended source (compact core) suggests that this medium is optically thin (thick) (Schneider 2002, pg. 181).

The other side of the continuum spectrum of a quasar shows an increase towards UV wavelengths. This is consistent with the thermal radiation produced by an accretion disk. The spectrum continues up to the hard X-ray regime where it follows a power law with a spectral index of  $\alpha = 0.7$ . However, this does not extend down to soft X-rays where the spectrum is actually larger than what would be expected from the power law. The interpretation is that the (non-thermal) source of the X-ray emission produces a simple power law, and the additional flux at lower X-ray energies is thermal emission from the accretion disk (Schneider 2002, pg. 195).

Two prominent features in the continuum spectrum of a quasar are the “big blue bump” (BBB) and IR-bump. The former is described by detailed models of accretion disks with the latter being described by thermal emission of warm dust ( $T \lesssim 2000K$  (Schneider 2002, pg. 195).

#### EMISSION LINES

The optical and UV spectra of quasars are dominated by numerous strong and very broad emission lines; see Figure 44. Typically lines of Balmer series and Ly $\alpha$  of H, and metal lines of ions, are observed. Such lines are interpreted as arising from Doppler velocities and their corresponding widths suggest velocity distributions of the components in the emitting region of order  $\Delta v \lesssim 10^4 \text{ km s}^{-1}$  (or  $\Delta\lambda/\lambda \lesssim 0.03$ ). These lines cannot be a result of thermal line broadening since that would imply  $kT \sim m_p(\Delta v)^2/2 \sim 1 \text{ Mev}$  or  $T \sim 10^{10} K$  - no emission lines would be produced at such large temperatures because all atoms would be fully ionized (also at this temperature electron/positron pairs would form and we would see evidence of this in gamma

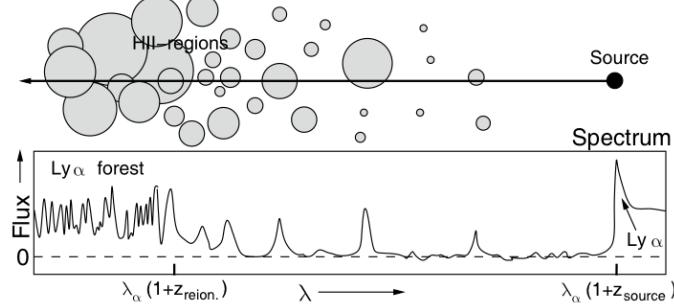


FIG. 45.— Light from a very distant quasar is absorbed by neutral hydrogen along the line-of-sight. Flux will be visible at the corresponding wavelengths where light passes through H II regions. The Ly $\alpha$  forest is produced when the H II regions overlap. Image taken from Schneider (2002). radiation at 511 keV). Such large Doppler velocities are consistent with Keplerian rotation around a SMBH at a distance of about 1000 Schwarzschild radii (Schneider 2002, pg. 196).

The region in which the broad emission lines are produced is known as the broad-line region (BLR). Examination of the heights of the emission lines allows for an estimation of the gas density within the BLR. To see this, we must first point out that both allowed and semi-forbidden transitions are found amongst the broad lines<sup>18</sup>. However, no forbidden transitions are observed among the broad lines. The absence of forbidden lines is used to derive a lower limit for the gas density while the occurrence of semi-forbidden lines is used to yield an upper bound. Typical values find  $n_e \sim 3 \times 10^9 \text{ cm}^{-3}$ . Moreover, from the ionization stages of the line-emitting elements, a temperature can be estimated which gives  $T \sim 10^4 \text{ K}$ . The density and temperature measures can then be coupled to observed line strength and distance to the quasar to determine volume-filling fraction of line-emitting gas. One typically finds that the gas in which the broad lines originate fill only  $10^{-7}$  of the BLR; they must be concentrated in clouds (Schneider 2002, pg. 197).

Quasars also show narrow emission lines in their optical and UV spectra. The typical width suggests Doppler velocities of  $\sim 400 \text{ km/s}$ , which is significantly narrower than the BLR, but still broader than normal galaxies. The region in which these lines arise is known as the narrow line region (NLR). The existence of forbidden lines in the NLR suggests that it has a much lower density than the BLR. Indeed, typical measures find that NLR has  $n_e \sim 10^3 \text{ cm}^{-3}$ ,  $T \sim 10^4 \text{ K}$ , and an extent of about 100 pc (Schneider 2002, pg. 201).

We can now glean a picture of the gas distribution around a quasar. First is the BLR which contains gas clouds with electron number densities of  $10^9 \text{ cm}^{-3}$  and velocities of order  $10^4 \text{ km/s}$ . Note that the kinematics of these clouds is not known. Cooling within these clouds occurs via the observed broad emission lines whilst heating arises from the absorption of continuum radiation from the quasar, which photoionizes the gas. The difference between the energy of the ionizing photon and the ionization potential yields the energy of the released electron, which is then thermalized by collisions and heats the gas. The NLR, on the other hand, contains lower density gas clouds that extend much further from the quasar. The NLR can actually be resolved for nearby Seyfert galaxies and it appears as two cone-shaped regions. This suggests that the ionization of the NLR from the continuum radiation of the quasar is not isotropic, but rather depends strongly on direction (Schneider 2002, pg. 201).

#### ABSORPTION LINES

Neutral hydrogen absorbs photons at a rest wavelength of  $\lambda = \lambda_{\text{Ly}\alpha} = 1216 \text{ \AA}$  which corresponds to the transition of an electron from the ground state to the next highest energy state. Photons from a quasar at redshift  $z_{\text{QSO}}$  are redshifted to this wavelength somewhere along the line-of-sight to the quasar if the photon was emitted with  $\lambda_{\text{Ly}\alpha}/(1+z_{\text{QSO}}) < \lambda < \lambda_{\text{Ly}\alpha}$ . However, if the wavelength at emission was greater than  $\lambda_{\text{Ly}\alpha}$  then the radiation cannot be absorbed by neutral hydrogen on its way to us. Hence, a jump should occur between the red and blue side of the Ly $\alpha$  emission line of the quasar; this is known as the Gunn-Peterson effect (Schneider 2002, pg. 331).

Bluewards of the Ly $\alpha$  emission line we see that the spectrum of high redshift quasars are characterized by strong absorption lines, caused by the aforementioned absorption process. This dense system of absorption lines is known as the Ly $\alpha$  forest and is subcategorized according to the strength of absorption lines. Narrow absorption lines are caused by the absorption of neutral hydrogen gas with column densities  $N_H \lesssim 10^{17} \text{ cm}^{-2}$ , known as Lyman Limit Systems (LLS). They are named this because at column densities higher than this, neutral hydrogen almost totally absorbs all the ionizing radiation at  $\lambda \leq 912 \text{ \AA}$  (in the hydrogen rest-frame). Broad absorption lines occur from hydrogen gas with column densities  $N_H \gtrsim 10^{20} \text{ cm}^{-2}$ , known as Damped Lyman Limit Systems (DLLS) (Schneider 2002, pg. 220).

The Ly $\alpha$  forest provides an interesting probe into the epoch of reionization. If we compare this spectrum to quasars at high redshifts we see an increased damping of the Ly $\alpha$  forest with increasing redshift. In fact, at redshifts greater than  $\sim 5$  we observe that the Ly $\alpha$  forest becomes almost completely depleted into the Gunn-Peterson trough. This strong damping indicates the presence of large column density neutral hydrogen absorption systems like the DLLS. This rapid suppression of the Ly $\alpha$  forest with increasing  $z$  suggests that the abundance of neutral hydrogen increases with increasing  $z$ . The transition to this phase

<sup>18</sup> Examples of the allowed transitions are Ly $\alpha$ , MgII, and CIV, whereas CIII] and NIV] are semi-forbidden transitions. An excited atom can fall to a lower energy state either through spontaneous emission of a photon or through losing energy via collisions with other atoms. The rate of the former process is described by atomic properties whereas the rate of the latter depends on the gas density. If the density of gas is high, the mean time between collisions is much shorter than the average lifetime of forbidden or semi-forbidden transitions. The corresponding line photons are then not observed. That is, to make forbidden lines visible, the gas density needs to be very low. Densities this low cannot be produced in the laboratory, thus they are ‘forbidden’.

marks the epoch of reionization. Figure 45 provides a cartoon description of this process and shows how the increased damping of the Ly $\alpha$  forest corresponds to absorption by high column density neutral hydrogen clouds. The redshift at which the Ly $\alpha$  forest becomes observable after the Gunn-Peterson trough can be used to constrain the epoch of reionization.

**QUESTION 14**

**Sketch the SED from the radio to gamma of extragalactic radiation on large angular scales. Describe the source and emission mechanism for each feature.**

### QUESTION 14

**Sketch the SED from the radio to gamma of extragalactic radiation on large angular scales. Describe the source and emission mechanism for each feature.**

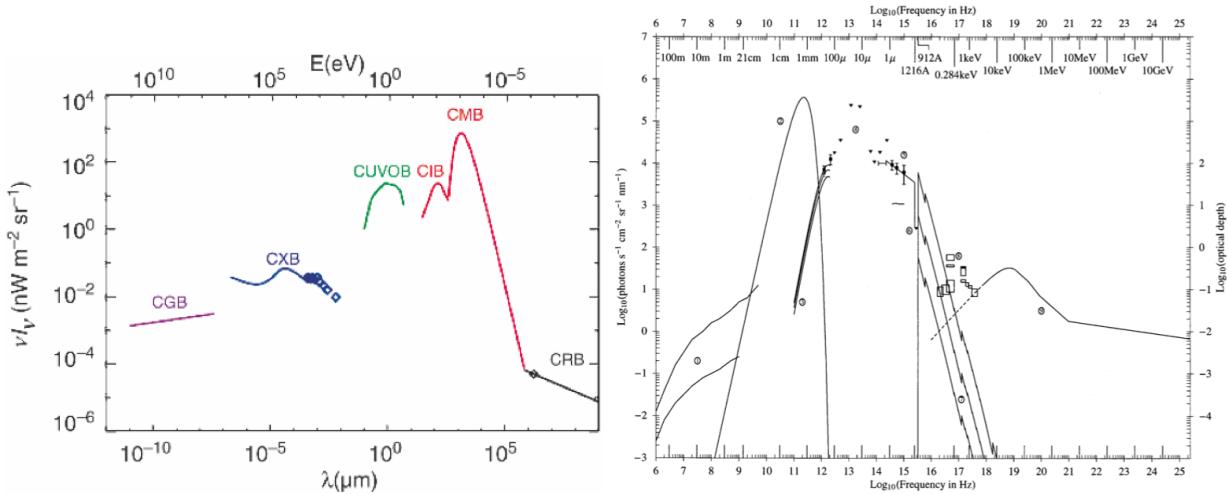


FIG. 46.— (left) Spectrum of cosmic background radiation, plotted as  $\nu I_\nu$  versus wavelength. Besides the CMB, background radiation exists in the radio domain (cosmic radio background, CRB), in the infrared (CIB), in the optical/UV (CUVOB), in the X-ray (CXB), and at gamma-ray energies (CGB). With the exception of the CMB, probably all of these backgrounds can be understood as a superposition of the emission from discrete sources. Furthermore, this figure shows that the energy density in the CMB exceeds that of other radiation components. Image taken from Schneider (2002). (right) The background radiation spectrum of the universe: (1) radio, (2) CMB, (3) FIRAS excess, (4) DIRBE background (points with error bars), (5) optical background, (6) UVB, (7) the ISM photoionization optical depth (right-hand scale) for  $10^{19}$ ,  $10^{18}$ , and  $10^{17}$  H atoms  $\text{cm}^{-2}$ , (8) soft X-ray background, and (9) high-energy background. In this diagram, an equal plotted value means an equal amount of energy per logarithmic interval of frequency. The new Voyager upper limit between 912 and 1216 Å suggests that the transition from the high background in the visible to the low background in the X-ray may occur at 1216 Å, which in turn would suggest that the UV and visible background at high galactic latitudes is redshifted Ly $\alpha$  recombination radiation. Image taken from Henry (1999).

As we have briefly discussed in another question, isotropic radiation of extragalactic origin has been found in wavelength domains from radio up to gamma. Following the terminology of the CMB, these are called background radiation as well. However, the name should not imply that it is a background radiation of cosmological origin, in the same sense as the CMB. From the thermal cosmic history, no optical or X-ray radiation is expected from the early phases of the universe. Therefore, for a long time it was unknown what the origin of these different background radiations may be. In Figure 46 we show the SED of extragalactic background radiation with the ordinates plotted in units of  $\nu I_\nu$ , so that it denotes equal energy per logarithmic mass interval (Schneider 2002, pg. 379). Although a lot of the background radiation arises from the combination of many point sources, we can still plot its spectral energy smoothly by averaging over the sky (Yevgeni).

#### COSMIC RADIO BACKGROUND

The **Cosmic Radio Background (CRB)** follows a power law in its spectral flux density,  $S_\nu \propto \nu^\alpha$ , with  $\alpha \approx 0.6$ , and has a brightness temperature of  $T = 1.2$  K at 1 GHz. This suggests that the source of emission comes from synchrotron emission, produced by relativistic electrons. This fact puts a constraint on the possible sources for the CRB. If the source relates to star formation activity, it must not overproduce the observed FIR background, whose origin comes from star formation, as discussed below. Also it should not overproduce the observed X-ray and  $\gamma$ -ray backgrounds through inverse Compton scattering of background field photons, caused by these relativistic electrons (Yevgeni). It is argued by Singal et al. (2010) that the CRB cannot be galactic in origin. Firstly, observations of other spiral galaxies are not consistent with such a bright halo component existing in the radio. Moreover, inverse Compton scattering of ambient light in the galactic halo would produce a stronger X-ray background than is actually observed in the MW.

The origin of the CRB is still a highly debated question (Yevgeni). One possibility is that it arises from an extended distribution such as hot gas contained in the IGM or ICM. The former is unlikely since the magnetic field in the IGM would be too weak to produce synchrotron radiation. Some extended low-surface-brightness radio structures are observed in galaxy clusters, but their spectra are relatively steep with  $\alpha > 1$ , and therefore diffuse ICM emission is likely not a dominant contributor to the CRB. Another possibility is that the CRB results from the integrated emission of a large distribution of point sources. Radio SNe (associated with Type II SNe) and radio-quiet quasars have been considered, but observation evidence suggests that they are not numerous enough to produce this background. Presently, a favoured candidate are star-forming galaxies since they meet the requirements of being numerous enough, possessing strong enough magnetic fields, and have reasonably flat spectra. The only problem is that the observed FIR/Radio luminosity ratio observed for star-forming galaxies (e.g., Seyferts, spirals, ULIRGS) overproduces the FIR background (by a factor  $> 10$ ) if it is to account for the strength of the CRB. Recently, it has been suggested that a redshift-evolving FIR/R ratio could reconcile this problem. Possible evolutionary trends which could explain this include

an enhanced fraction of energy generated by star formation going into relativistic electrons, a larger fraction of high mass stars and hence more SNe, increased synchrotron emission due to stronger magnetic fields, or stronger AGN activities; this is still mostly speculation, however (Singal et al. 2010).

#### COSMIC MICROWAVE BACKGROUND

The CMB originates from the cosmological epoch of recombination. Beyond the redshift of  $z \gtrsim 1100$  all the atoms in the universe were ionized, and so the mean free path of photons was very small, due to free-free scattering. However, as the universe cooled enough for atoms to combine into neutral particles, only those photons with a specific wavelength corresponding to an electron transition in the atom were able to interact, and hence most photons were able to free stream. The CMB is the remnant we see of this photon flash, after it has been redshifted from its original wavelength of  $\lambda_{\text{emit}} \approx 1 \mu\text{m}$  to its present-day wavelength  $\lambda_0 \approx 2 \text{ mm}$  (Yevgeni).

The observations that the CMB has a nearly perfect blackbody spectrum and that it is nearly isotropic (once the Doppler dipole is removed) provide strong support for the Hot Big Bang model of the universe. A background of nearly isotropic blackbody radiation is natural if the universe was once hot, dense, opaque, and nearly homogeneous, as it was in the Hot Big Bang scenario. If the universe did not go through such a phase, then any explanation of the CMB will have to be much more contrived (Ryden 2002, pg. 185).

#### COSMIC INFRARED BACKGROUND

Observations of the **Cosmic Infrared Background (CIB)** have been difficult to obtain for several reasons. Firstly, the atmosphere blocks a large fraction of IR demanding that balloon-borne experiments or space telescopes be used. In addition, our own solar system (e.g., zodiacal light in the IR results from thermal emission of interplanetary dust particles) and MW (e.g., emission from ISM) are major sources of IR radiation (usually defined in the range of  $8 \mu\text{m} - 1000 \mu\text{m}$ ). The first major achievement was accomplished by the launch of the Cosmic Background Explorer (COBE), with its instrument Diffuse Infrared Background Experiment (DIRBE) (Yevgeni).

A recent study on the origins of the CMB was performed by Younger & Hopkins (2010). They found that the main source of IR radiation comes for extragalactic star formation, where dust obscuring UV emission from recently formed massive stars reprocesses this into thermal IR radiation. Most of the emission comes from Luminous Infrared Galaxies (LIRGS) – defined as having luminosities  $10^{11} L_\odot \lesssim L_{\text{IR}} \lesssim 10^{12} L_\odot$  – with most of this coming from moderate redshifts ( $1 \lesssim z \lesssim 2$ ). Another significant contribution comes from “regular” galaxies at low redshift, and from ULIRGS – defined as having luminosities  $L_{\text{IR}} \gtrsim 10^{12} L_\odot$  – at high redshift ( $z \gtrsim 2$ ). AGNs enshrouded in dust may also contribute a small portion ( $\lesssim 1\%$ ) of the observed CIB, though they are not numerous enough to have a major effect.

#### COSMIC OPTICAL BACKGROUND

The **Cosmic Optical Background (COB)** is the optical component of the extragalactic background light, which is the integrated radiation from all light sources outside the MW. Robust detection of the COB has long been hampered by extremely bright foreground emissions including terrestrial airglow and zodiacal light, which are a few orders of magnitude brighter than the COB signal (Matsuoka et al. 2011). The former describes emission from excited atoms in the upper atmosphere of the Earth due to absorption of solar radiation, while the latter describes the scattering of solar optical and UV radiation by interplanetary dust particles (Leinert et al. 1998). The dominant contribution to the COB comes from starlight of ordinary galaxies at  $z < 10$ , while other mechanisms, such as mass accretion to SMBHs in AGN, gravitational collapse of stars, and particle decay, can contribute smaller portions to the COB brightness (Matsuoka et al. 2011).

#### COSMIC ULTRAVIOLET BACKGROUND

The **Cosmic Ultraviolet Background (CUVB)** is similar to the COB in that foreground sources including terrestrial airglow and zodiacal light make measurements of its spectrum difficult. The largest component to the CUVB is starlight scattered by interstellar dust (not to be confused with absorption and subsequent thermal emission in the IR) with other potential sources including atomic and molecular emission lines (Murthy 2009). Henry (1999) also suggest that the COB and CUVB is redshifted Ly $\alpha$  radiation originating from radiative recombination occurring within neutral pockets of the IGM.

#### COSMIC X-RAY BACKGROUND

We have already discussed the CXB in detail in another question; it is believed to be the integrated emission from a large collection of discrete AGNs.

#### COSMIC GAMMA RAY BACKGROUND

My assumption is that the **Cosmic Gamma Ray Background (CGB)** is produced by gamma-ray bursts (GRBs). Long-soft GRBs are believed to be produced during the core-collapse SN of very massive (possibly WR) stars. For such massive stars, a black hole with a surrounding debris disk will form during a SN. The collimating effect of the debris disk and associated magnetic fields would lead to a jet emanating from the centre of the SN. Since the jet material will be highly relativistic, it will appear to be further collimated. The jet will plow its way through the overlying material of the infalling stellar envelope producing bursts of gamma-rays. Short-hard GRBs, on the other hand, are thought to be the result of mergers of compact objects, either two NSs or a NS and a BH. These types of GRBs emit roughly 1000 times less energy than their long-soft counterparts (Carroll & Ostlie 2007, pg. 544).

**QUESTION 15**

**What are AGNs? Describe different observational classes of them and how they may relate to each other.**

## QUESTION 15

**What are AGNs? Describe different observational classes of them and how they may relate to each other.**

An **active galactic nucleus (AGN)** is a small region at the centre of a galaxy that shows highly luminous non-thermal emission that is strong over a broad range of the electromagnetic spectrum; often covering the range of radio to gamma. This is significant because if one were to simply superimpose the spectra of the stellar population in these galaxies, the resulting spectrum would generally only stretch from 4000 Å (UV) to 20,000 (IR) Å (modulo recent star formation, which would add more UV, and dust extinction, which would bump up the far-IR). Moreover the luminosity of these central regions are often on the same order of magnitude as all other sources of radiation in the galaxy, making the source of this luminosity unlikely to be dense stellar populations. The various classes of AGN are unified by the idea that all AGN emission is ultimately caused by accretion of material onto SMBHs. Observational differences between AGNs can be explained due to orientation and line-of-sight extinction between their respective SMBHs and observers (Charles).

### AGN ZOOLOGY

A wide range of objects are subsumed under the name AGN, all of which share some common features. These are listed below:

- Strong non-thermal emission in the core of the host galaxy ranging from radio to gamma. This is in contrast to the light of normal galaxies which is dominated by stars in the optical and NIR part of the spectrum. This is considered thermal radiation since the emitting plasma in stellar atmospheres is basically in thermodynamical equilibrium (Schneider 2002, pg. 175). As a result of their broad distributions, quasars – the most luminous class of AGN – were initially discovered by identifying radio sources with point-like optical sources (Schneider 2002, pg. 178).
- The flux of the source varies at nearly all frequencies, where the variability time-scale differs among the objects and also depends on the wavelength. In general, it is found that the variability time-scale is smaller and its amplitude larger when going to higher frequencies of the observed radiation (Schneider 2002, pg. 178).
- Extended radio emission, which can be divided into two classes. *Fanaroff-Riley Type I* (FR I) are brightest close to the core, and the surface brightness decreases outwards. In contrast, the surface brightness of *Fanaroff-Riley Type II* (FRII) sources increases outwards, and their luminosity is in general higher than that of FR I sources. In addition, FR II radio sources often have **jets**; they are extended linear structures that connect the compact core with a **radio lobe**. Jets often show internal structure such as knots and kinks. Their appearance indicates that they transport energy from the core out into the radio lobe. In some cases the jets are bipolar, while in others only one jet is visible (Schneider 2002, pg. 179). The radio emission spectral indices of the core, jet and lobes are different, and therefore AGN radio catalogues at different wavelengths will preferentially pick out AGN with certain morphological features (Charles).
- Continuum emission broadly described by a single power law, as in equation (197). The radio spectrum in the lobes has an index  $\alpha \sim 0.7$ , while the core has  $\alpha \sim 0$ . Emission is often highly polarized in the extended regions (not as much in the core) – this fact, combined with the spectral indices, suggests relativistic synchrotron emission. The spectral index will be adjusted by the optical depth of the emitting body. Radio lobes are optically thin, meaning that  $\alpha \sim 0.7$  is only due to the kinetic energy spectrum of the electrons. In contrast, the core is optically thick, which explains why core emission has an index of  $\alpha \lesssim 0$  (Charles).
- Extremely broad emission lines with Doppler velocities of order  $v \sim 10^4 \text{ km s}^{-1}$ . Narrower emission lines with  $v \sim 100 \text{ km s}^{-1}$  are also observed, though these are still broad compared to the typical velocities in normal galaxies (Schneider 2002, pg. 181).
- Low local space density: Seyferts make up  $\sim 5\%$  of spirals, while quasars are even more rare. Note that QSOs have a space density versus redshift distribution that peaks at  $z \sim 3.3$ , and tails off on either side. The reason for this is likely galactic evolution of some kind. For instance, it may be possible that at higher redshifts there are fewer large BHs whereas at lower redshifts they are not fed as efficiently through accretion (Charles).

We now summarize the different classes of AGN and identify any exceptions they make to the above list of common characteristics.

- **Quasars:** These are point-like in the optical and were therefore originally misidentified as galactic stars. They tend to have high redshift, meaning they are extremely luminous, and are characterized by a very blue spectrum with broad emission lines (Charles).
  - **Radio-quiet quasars:** These are identical to regular quasars (or “radio-loud quasars”) in the optical, but display very little radio emission. These are often referred to as quasi-stellar objects (QSOs), though in modern usage the term QSO is meant to include both quasars and radio-quiet QSOs (Schneider 2002, pg. 183).
- **Seyfert Galaxies:** These are the AGNs which were detected first. Their luminosity is considerably lower than that of QSOs. On optical images they are identified as spiral galaxies which have an extraordinarily bright core whose spectrum shows strong and broad emission lines (Schneider 2002, pg. 183).

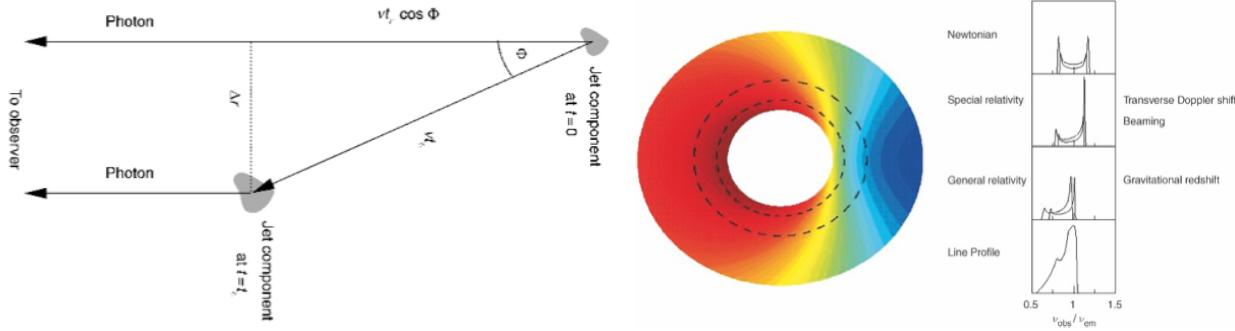


FIG. 47.— (a) Explanation of superluminal motion: a source component is moving at velocity  $v$  and at an angle  $\phi$  relative to the line-of-sight. We consider the emission of photons at two different times  $t = 0$  and  $t = t_e$ . Photons emitted at  $t = t_e$  will reach us by  $\Delta t = t_e(1 - \beta\cos\phi)$  later than those emitted at  $t = 0$ . The apparent separation of the two source components then is  $\Delta r = v_{t_e}\sin\phi$ , yielding an apparent velocity on the sky of  $v_{app} = \Delta r/\Delta t = v\sin\phi/(1 - \beta\cos\phi)$ . (b) The profile of the broad iron line is caused by a combination of Doppler shift, relativistic beaming, and gravitational redshift. On the left, the observed energy of the line as a function of position on a rotating disk is indicated by colours. Here, the energy in the right part of the disk which is moving towards us is blueshifted, whereas the left part of the disk emits redshifted radiation. Besides this Doppler effect, all radiation is redshifted because the photons must escape from the deep potential well. The smaller the radius of the emitting region, the larger this gravitational redshift. The line profile we would obtain from a ring-shaped section of the disk (dashed ellipses) is plotted in the panels on the right. The uppermost panel shows the shape of the line we would obtain if no relativistic effects occurred besides the non-relativistic Doppler effect. Below, the line profile is plotted taking the relativistic Doppler effect and beaming into account. This line profile is shifted towards smaller energies by gravitational redshift so that, in combination, the line profile shown at the bottom results. Images taken from Schneider (2002).

- **Seyfert Type 1:** Display both the broad and narrow set of emission lines characteristic of AGN. The optical spectrum of Seyfert 1 galaxies is very similar to that of QSOs. In fact, a smooth transition exists between (radio-quiet) QSOs and Seyfert 1 galaxies. Except for their core luminosity, no fundamental differences exist between the two classes, and the dividing line only occurs for historical reasons based on the different methods of detecting them.
- **Seyfert Type 2:** Contain only the narrow set of emission lines that are characteristic of AGN. A continuum actually exists between transitioning between Seyfert 1 galaxies and Seyfert 2 galaxies; often Seyfert 1.5 and Seyfert 1.8 galaxies are used.
- **Radio Galaxies:** These are elliptical galaxies with an AGN so that they display unusually strong radio emission over regular ellipticals. In a similar fashion to Seyfert galaxies, for radio galaxies we also distinguish between those with and without broad emission lines (as below). In principle, the two types of radio galaxy can be considered as radio-loud Seyfert 1 and Seyfert 2 galaxies but with a different morphology of the host galaxy (Schneider 2002, pg. 184).
  - **Broad-line radio galaxies (BLRGs):** Display both the broad and narrow set of emission lines characteristic of AGN. A smooth transition between BLRG and quasars also seems to exist, again separated by optical luminosity as for Seyfert galaxies.
  - **Narrow-line radio galaxies (NLRGs):** Contain only the narrow set of emission lines that are characteristic of AGN. A continuum actually exists between BLRGs and NLRGs.
- **Blazars:** These are very similar to OSOs, except that they appear to have very rapid (often over days) and greatly varying optical emission. This emission is also significantly more polarized compared to that of QSOs (Charles). All known blazars are radio sources. Besides the violent variability, blazars also show highly energetic and strongly variable  $\gamma$ -radiation (Schneider 2002, pg. 185).
  - **Optically violent variables (OVVs):** These are QSOs that show substantial variation in the optical over very short periods of time, and display a large degree of optical polarization.
  - **BL Lacertae objects (BL Lacs):** Resemble OVs but without strong emission and absorption lines. During their epochs of low luminosity, some BL Lacs will actually appear to morph into OVs, suggesting a continuum between the two classes.
- **LINERs (Low Ionization Nuclear Emission Regions):** These are not traditionally considered AGN but are still characterized by line strengths that are hard to reproduce with a superposition of stellar spectra, combined with a slight UV excess. LINERs appear to be lower-luminosity versions of AGN, and there perhaps is a continuum between LINERS and Seyferts or radio galaxies. They are fairly common in the universe; over half the all local spirals show LINER activity (Charles).

#### AGN POWER SOURCE

We note a few pieces of evidence that point us to a strong, compact power source for AGN. The size and structure of the jets and radio lobes of AGN requires an enormous power source that “points” in a constant direction for a time on the order of 10 Myr. Bright AGN have luminosities up to  $10^{47}$  erg s $^{-1}$ ; if this were constant for 10 Myr, a total of  $10^{61}$  erg of energy is required; an enormous amount indeed. AGN luminosity may change significantly on the timescale of hours. Since information propagates at the speed of light, global variations in emitted power must be the result of changes in a very compact source (Charles).

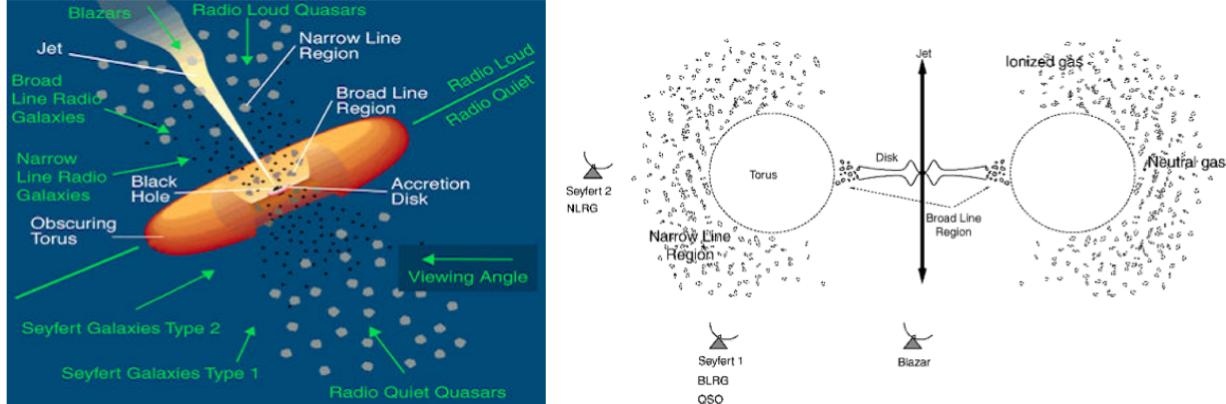


FIG. 48.— These illustrations show our current understanding of the unification of AGN types. The accretion disk is surrounded by a thick torus containing dust which thus obscures the view to the centre of the AGN. When looking from a direction near the plane of the disk, a direct view of the continuum source and the BLR is blocked, whereas it is directly visible from directions closer to the symmetry axis of the disk. The difference between Seyfert 1 (and BLRG) and Seyfert 2 (and NLRG) is therefore merely a matter of orientation relative to the line-of-sight. If an AGN is seen exactly along the jet axis, it appears as a blazar. The difference between Seyferts and radio galaxies is explained as a matter of a lack of synchrotron emitting accretion jets (perhaps due to the spin of the SMBH?). Images taken from Bob's AST2040 Lecture Notes and Schneider (2002).

It is essentially impossible for this nuclear source to come from, for example, a dense star cluster. Fusion, at best, can derive  $0.008 mpc^2$  worth of energy per nucleon – this translates to  $\gtrsim 10^9 M_\odot$  required if one assumes an AGN lifetime of 10 Myr. The Schwarzschild radius of a  $10^9 M_\odot$  system is  $r_S \approx 10^{15}$  cm, the same order of magnitude as the compact source radius. Thus, strong gravitational effects cannot be avoided when discussing AGN activity<sup>19</sup>. The only other reasonable power source is gravitational potential energy via accretion. Accretion onto a BH, for example, has an efficiency  $\epsilon$  (i.e. in  $\epsilon mc^2$ ) of anywhere between 6% (for non-rotating BHs) and 30% (BH with maximum allowed angular momentum). Luminosity from steady-state accretion is Eddington luminosity limited; if we assume  $L = L_{Ed}$ , we obtain a mass estimate of  $10^6 M_\odot$  for Seyferts and  $10^8 M_\odot$  for QSOs. Such a large mass in a region so small would require that the mass be contained in a SMBH (Charles).

An SMBH power source is also consistent with a number of other observations. Firstly, the *big blue bump* can be explained as blackbody radiation from the superheated accretion disk. The bulk motion of radio jets is also highly relativistic<sup>20</sup>, and in many astrophysical processes the velocity of ejected material is on the same order of magnitude as the escape velocity of the system<sup>21</sup>. Also, relativistic jets suggest escape from a black hole. Rotating SMBHs naturally act like gyroscopes, allowing them to create straight jets. X-ray emission line profiles can also be explained by emission conditions near the SMBH event horizon; see Figure 47.

#### AGN UNIFICATION

Since AGN are all dominated by their SMBHs, it is to be expected that all AGN can be described by some combination of accretion rate ratio  $\dot{m}/\dot{m}_{Ed}$  and mass of the BH  $M$ . The latter can be used to explain the transition from Seyfert to QSOs, and BLRGs to quasars. Furthermore, in the context of the SMBH plus accretion disk model, an AGN system will naturally be anisotropic. This suggests that absorption and scattering along the line-of-sight with some inclination angle can explain the transition between certain AGN types; see Figure 48. This is further reinforced by polarization measurements, which show narrow-line AGN will often also have fainter, highly polarized broad line emission hidden within their SEDs. This suggests that the difference between a narrow-line radio galaxy or Seyfert and their broad-line counterparts is the obscuration of the broad-line emitting regions by dust and gas. Chandra and XMM Newton recently discovered QSOs enshrouded in regions with high hydrogen column density (as a result, soft X-rays suffer a large degree of extinction); these QSOs contribute to the CXB discussed in another question (Charles).

Another issue affecting AGN observations is the relativistic **beaming** of their jets. According to SR, a moving source that is emitting isotropically in its rest-frame has an anisotropic emission pattern, with the angular distribution depending on its velocity. The radiation is emitted preferentially in the direction of the velocity vector of the source (thus, in the forward direction), so that a source will appear brighter if it is moving towards the observer. Conversely, the source will appear dimmer if it is moving away from the observer (Schneider 2002, pg. 210). Hence, relativistic beaming greatly increases the flux we receive from an incoming jet, and equivalently decreases the flux we receive from a jet moving away from us. This explains why, in general, only one jet can easily be detected. This beaming is what makes blazars so bright, and fact that blazars can be divided into OVs and BL Lacs is due to the strong dependence on this effect with respect to moving slightly away from pointing directly into the jet. Synchrotron radiation is also polarized, which explains why blazars have polarized emission. Moreover, relativistic beaming amplifies changes in luminosity, so any minute change in the AGN corresponds to a large change in the resulting emission (Charles).

How to connect radio loud and radio quiet systems is not entirely resolved, since there is more than one way to unify AGN. One suggestion is that it may be related to galactic morphology; for instance, radio galaxies are ellipticals whereas Seyferts

<sup>19</sup> This is even true if we assume the material can be ejected from the system by some hypothetical process; the energy this would require is greater than the energy generated by fusion.

<sup>20</sup> As determined through direct observations and the requirement for shock-heated electrons to produce synchrotron radiation in the radio lobes.

<sup>21</sup> In fact, the observed bulk motion is actually apparently superluminal, which caused some consternation in early quasar work where researchers would claim these observations implied quasar redshift measurements were erroneous. This superluminal motion is caused simply by light lag and a relativistic beam moving toward the observer.

are spirals. While there is a link between galactic and AGN luminosity, it is not obvious what effect morphology would have. Another proposal is that radio-loud/quiet has something to do with spin of the black hole – since accretion jets are created by the winding of magnetic fields, this is an understandable relation (Charles).

**QUESTION 16**

**What are galaxy clusters? What are their basic properties (e.g., mass, size). List and explain three ways they can be detected.**

### QUESTION 16

**What are galaxy clusters? What are their basic properties (e.g., mass, size). List and explain three ways they can be detected.**

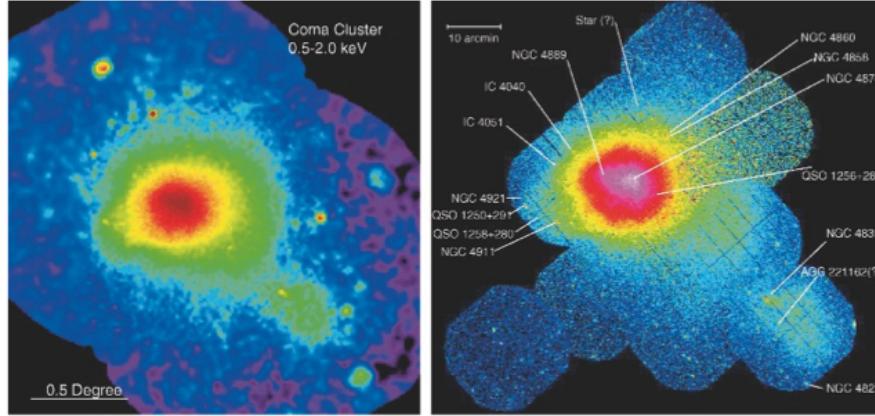


FIG. 49.— X-ray images of the Coma cluster, taken with the ROSAT-PSPC (left) and XMM-EPIC (right). The image size in the left panel is  $2.7^\circ \times 2.5^\circ$ . A remarkable feature is the secondary maximum in the X-ray emission at the lower right of the cluster center which shows that even Coma, long considered to be a regular cluster, is not completely in an equilibrium state, but is dynamically evolving, presumably by the accretion of a galaxy group. Image taken from Schneider (2002).

Galaxies are not uniformly distributed in space, but instead show a tendency to gather together in **galaxy groups** and **galaxy clusters**. The transition between groups and clusters of galaxies is smooth. The distinction is made by the number of their member galaxies. Roughly speaking, an accumulation of galaxies is called a group if it consists of  $N \lesssim 50$  members within a sphere of diameter  $D \lesssim 2$  Mpc. Clusters have  $N \gtrsim 50$  members and diameters  $D \gtrsim 2$  Mpc. Typical values for the mass of a cluster are  $M \sim 10^{14} M_\odot$  for massive clusters, whereas for groups  $M \sim 10^{13} M_\odot$  is characteristic, with the total mass range of groups and clusters extending over  $10^{12} M_\odot \lesssim M \lesssim 10^{15} M_\odot$  (Schneider 2002, pg. 223).

#### ABELL'S CRITERIA (OPTICAL SEARCH)

George Abell compiled a catalog of galaxy clusters, published in 1958, in which he identified regions in the sky that show an overdensity of galaxies. The criteria Abell applied for the identification of clusters refer to an overdensity of galaxies within a specified solid angle. According to these criteria, a cluster contains  $N \geq 50$  galaxies in a magnitude interval  $m_3 \leq m \leq m_3 + 2$ , where  $m_3$  is the apparent magnitude of the third brightest galaxy in the cluster<sup>22</sup>. These galaxies must be located within a circle of angular radius

$$\theta_A = \frac{1.7 \text{ arcmin}}{z}, \quad (198)$$

where  $z$  is the estimated redshift. The latter is determined by the assumption that the luminosity of the tenth brightest galaxy in a cluster is the same for all clusters. The so-determined redshift should be within the range  $0.02 \leq z \leq 0.2$  for the selection of Abell clusters. The lower limit is chosen such that a cluster can be found on a single POSS photoplate ( $\sim 6^\circ \times 6^\circ$ ) and does not extend over several plates, which would make the search more difficult (e.g., because the photographic sensitivity may differ for individual plates). The upper redshift bound is chosen due to the sensitivity limit of the photoplates (Schneider 2002, pg. 229).

The selection of galaxy clusters from an overdensity of galaxies on the sphere is not without problems. In particular, the Abell catalogue, which is based on this selection criteria, is neither complete (in the sense that all objects which fulfill the selection criteria are *not* in the catalogue) nor reliable (in the sense that *false positives* exists). A galaxy cluster is a three-dimensional object, whereas galaxy counts on images are necessarily based on the projection of galaxy positions onto the sky. Therefore, projection effects are inevitable. Random overdensities on the sphere caused by line-of-sight projection may easily be classified as clusters. The reverse effect is likewise possible: due to fluctuations in the number density of foreground galaxies, a cluster at high redshift may be classified as an insignificant fluctuation – and thus remain undiscovered. Of course, not all members of a cluster classified as such are in fact galaxies in the cluster, as here projection effects also play an important role. Furthermore, the redshift estimate is relatively coarse (Schneider 2002, pg. 229).

#### X-RAY EMISSION

Clusters of galaxies are the brightest extragalactic X-ray sources besides AGNs. Their characteristic luminosity is  $L_X \sim 10^{43}$  up to  $\sim 10^{45}$  erg s<sup>-1</sup> for the most massive clusters. This X-ray emission from clusters is spatially extended, so it does not originate in individual galaxies. The spatial region from which we can detect this radiation can have a size of 1 Mpc or even larger.

<sup>22</sup> The reason for choosing the third brightest galaxy is that the luminosity of the brightest galaxy may vary considerably among clusters. Even more important is the fact that there is a finite probability for the brightest galaxy in a sky region under consideration to not belong to the cluster, but to be located at some smaller distance from us.

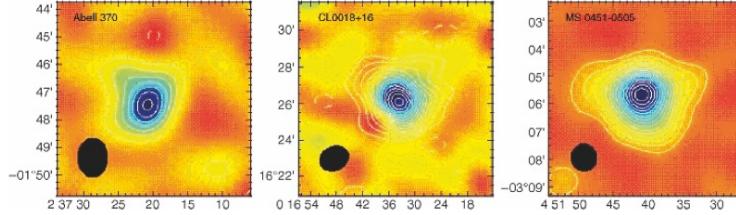


FIG. 50.— Sunyaev-Zeldovich maps of three clusters of galaxies at  $0.37 < z < 0.55$ . Plotted is the temperature difference of the measured CMB relative to the average CMB temperature (or, at fixed frequency, the difference in radiation intensities). The black ellipse in each image specifies the instrument’s beam size. For each of the clusters shown here, the spatial dependence of the SZ effect is clearly visible. Since the SZ effect is proportional to the electron density, the mass fraction of baryons in clusters can be measured if one additionally knows the total mass of the cluster from dynamical methods or from the X-ray temperature. The analysis of the clusters shown here yields for the mass fraction of the intergalactic gas  $f_g \approx 0.1$ . Image taken from Schneider (2002).

Furthermore, the X-ray radiation from clusters does not vary on timescales over which it has been observed ( $\lesssim 30$  yr). Variations would also not be expected if the radiation originates from an extended region. The spectral energy distribution of the X-rays leads to the conclusion that the emission process is optically thin thermal bremsstrahlung (free-free radiation) from a hot gas (Schneider 2002, pg. 242). See Figure 49 for X-ray images of the Coma cluster.

Searching for galaxy clusters via X-ray emission is a more reliable method than searching for overdensities of galaxies on the sphere using optical methods. Since the hot X-ray gas signifies a deep potential well, thus a real three-dimensional overdensity of matter, projection effects become virtually negligible. The X-ray emission is  $\propto n_e^2$ , which again renders projection effects improbable. In addition, the X-ray emission, its temperature in particular, seems to be a very good measure for the cluster mass, as we have discussed in a previous question (Schneider 2002, pg. 255).

The first cosmologically interesting X-ray catalog of galaxy clusters was the EMSS (Extended Medium Sensitivity Survey) catalog. It was constructed from archival images taken by the Einstein observatory which were scrutinized for X-ray sources other than the primary target in the field-of-view of the respective observation. These were compiled and then further investigated using optical methods (i.e., photometry and spectroscopy). The EMSS catalog contains 835 sources, most of them AGNs, but it also contains 104 clusters of galaxies. Since the Einstein images all have different exposure times, the EMSS is not a strictly flux-limited catalog. But with the flux limit known for each exposure, the luminosity function of clusters can be derived from this (Schneider 2002, pg. 255).

#### SUNYAEV-ZELODOVICH EFFECT

Electrons in the hot gas of the ICM can scatter photons of the CMB. The optical depth and thus the scattering probability for this Compton scattering is relatively low, but the effect is nevertheless observable and, in addition, is of great importance for the analysis of clusters. If one can spatially resolve the SZ effect, which is possible today with interferometry, one obtains information about the spatial density and temperature distribution; see Figure 50 (Etsuko). The SZ effect can be identified by noting that the CMB spectrum, measured in the direction of a galaxy cluster, deviates from a Planck spectrum; the degree of this deviation depends on the temperature of the cluster gas and on its density (Schneider 2002, pg. 253). This effect will be explored in more detail in another question.

#### WEAK GRAVITATIONAL LENSING

The weak lensing effect can not only be used to map the matter distribution of known clusters, but it can also be used to search for clusters. Mass concentrations generate a tangential shear field in their vicinity, which can specifically be searched for; see Figure 51. The advantage of this method is that it detects clusters based solely on their mass properties, in contrast to all other methods which rely on the emission of electromagnetic radiation, whether in the form of optical light from cluster galaxies or as X-ray emission from a hot ICM. In particular, if clusters with atypically low gas or galaxy content exist, they could be detected in this way (Schneider 2002, pg. 269).

With this method, quite a number of galaxy clusters have been detected already. Further candidates exist, in that from the shear signal a significant mass concentration is indicated but it cannot be identified with any concentration of galaxies on optical images. The clarification of the nature of these lens signals is of great importance: if in fact matter concentrations do exist which correspond to the mass of a cluster but which do not contain luminous galaxies, then our understanding of galaxy evolution needs to be revised. However, we cannot exclude the possibility that these statistically significant signals are statistical outliers, or result from projection effects – remember, lensing probes the line-of-sight integrated matter density. Together with the search for galaxy clusters by means of the SZ effect, the weak lensing effect provides an interesting alternative for the detection of mass concentrations compared to traditional methods (Schneider 2002, pg. 269).

#### COLOR-MAGNITUDE DIAGRAM

Plotting the colour of cluster galaxies versus their magnitude, one finds a very well-defined, nearly horizontal sequence; see Figure 51. This red cluster sequence (RCS) is populated by the early-type galaxies in the cluster. The scatter of early-type galaxies around this sequence is very small, which suggests that all early-type galaxies in a cluster have nearly the same colour, only weakly depending on luminosity. Even more surprising is the fact that the colour-magnitude diagrams of different clusters at the same redshift define a very similar red cluster sequence: cluster galaxies with the same redshift and luminosity have virtually the same colour. Comparing the red sequences of clusters at different redshifts, one finds that the sequence of cluster galaxies is

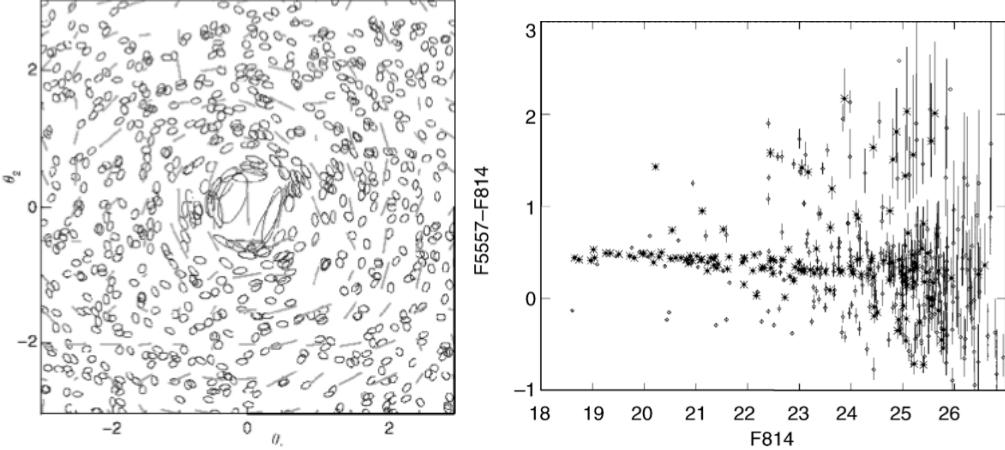


FIG. 51.— (left) The principle of the weak gravitational lensing effect is illustrated here with a simulation. Due to the tidal component of the gravitational field in a cluster, the shape of the images (ellipses) of background galaxies get distorted and, as for arcs, the galaxy images will be aligned, on average, tangentially to the cluster centre. By local averaging over the ellipticities of galaxy images, a local estimate of the tidal gravitational field can be obtained (the direction of the sticks indicates the orientation of the tidal field, and their length is proportional to its strength). From this estimated tidal field, the projected mass distribution can then be reconstructed. (right) Colour-magnitude diagram of the cluster of galaxies Abell 2390, observed with the HST. Star symbols represent early-type galaxies, identified by their morphology, while diamonds denote other galaxies in the field. The red cluster sequence is clearly visible. Images taken from Schneider (2002).

redder the higher the redshift is. In fact, the red cluster sequence is so precisely characterized that, from the colour-magnitude diagram of a cluster alone, its redshift can be estimated, whereby a typical accuracy of  $\Delta z \sim 0.1$  is achieved. The accuracy of this estimated redshift strongly depends on the choice of the colour filters. Since the most prominent spectral feature of early-type galaxies is the 4000-Å break, the redshift is estimated best if this break is located right between two of the colour bands used (Schneider 2002, pg. 271).

Not only can this method be used to determine the distance to a cluster, but the well-defined RCS is of crucial importance for our understanding of the evolution of galaxies. We know that the composition of a stellar population depends on the IMF and on its age: the older a population is, the redder it becomes. The fact that cluster galaxies at the same redshift all have roughly the same colour indicates that their stellar populations have very similar ages. In fact, the colour of cluster galaxies is compatible with their stellar populations being roughly the same age as the Universe at that particular redshift. This also provides an explanation for why the RCS is shifted towards bluer colours at higher redshifts – there, the age of the Universe was smaller, and thus the stellar population was younger (Schneider 2002, pg. 273).

#### CLUSTERS AT HIGH REDSHIFT

The search for clusters in the optical (thus, by galaxy overdensities) becomes increasingly difficult at high  $z$  because of projection effects. Nevertheless, several groups have managed to detect clusters at  $z \sim 1$  with this technique. In particular, the overdensity of galaxies in three-dimensional space can be analyzed if, besides the angular coordinates on the sphere, the galaxy colours are also taken into account. Because of the RCS, the overdensity is much more prominent in this space than in the sky projection alone (Schneider 2002, pg. 274).

Through optical methods, it is also possible to identify galaxy concentrations at very high redshift. One approach is to assume that luminous AGNs at high redshift are found preferentially in regions of high overdensity, which is also expected from models of galaxy formation. With the redshift of the AGN known, the redshift at which one should search for an overdensity of galaxies near the AGN is defined. Those searches have proven to be quite successful; for instance, they are performed using narrow-band filter photometry, with the filter centred on the redshifted Ly $\alpha$  line, tuned to the redshift of the AGN. Candidates need to be verified spectroscopically afterwards. Remember that the identification of a strong spatial concentration of galaxies at high  $z \sim 4$  using this method is not sufficient to have identified a cluster of galaxies though, because it is by no means clear whether one has found a gravitationally bound system of galaxies (and the corresponding DM). Rather, such galaxy concentrations are considered to be the predecessors of galaxy clusters which will only evolve into bound systems during later cosmological evolution (Schneider 2002, pg. 275).

**QUESTION 17**

**Describe and give results from simulations of large scale structure in the universe. What role do they have in understanding the formation of large scale structure and Galaxy formation? What are their limitations?**

## QUESTION 17

**Describe and give results from simulations of large scale structure in the universe. What role do they have in understanding the formation of large scale structure and Galaxy formation? What are their limitations?**

### NUMERICAL SIMULATIONS

Analytical models of structure formation (i.e., linear perturbation theory or the spherical collapse model) are only capable of describing limiting situations. This results from the fact that gravitational dynamics is far too complicated to be examined in close detail. Instead, we tend to rely on the results provided by numerical simulations coupled to observations in order to describe the large-scale structure of the universe. It is of course important to take note of the limitations of such a strategy (Schneider 2002, pg. 293).

We will begin by describing the structure of large-scale numerical simulations. Simulations are usually parameterized by the number of particles  $N^3$  they trace and by the size of the box  $L$  in which they reside. For many applications it is sufficient to only use DM particles since they dominate the matter density of the universe, though many simulations also incorporate baryons in order to model the hydrodynamical aspects of structure formation. Due to the limited power and memory of computers we do not trace individual particles, but rather clump them together into bodies of total mass  $M$ . The value of  $M$  is determined by the total mass in the box, as dictated by the mean matter density of the universe, and by the total number of particles  $N^3$ . The size of the box is chosen such that it be larger than the scale of the structures we are hoping to observe, but yet still small enough that a reasonable mass resolution is achieved (Schneider 2002, pg. 294).

Periodic boundary conditions are then imposed in order to model the particles at the edge of the box. That is, we assume that the universe is homogeneous on scales larger than  $L$  so that we can stack the box periodically on its sides. It is because of this assumption of periodicity that the quantitative results of these simulations should only be confined to scales  $\lesssim L/2$ . In addition, the spatial resolution of the simulation is smoothed out on small scales in order to avoid unphysical interactions that result from clumping the particles into macroscopic bodies of mass  $M$ . More specifically, at small separations – within the so-called *softening length* – the force is modified from the  $1/r^2$  law so that artificial strong interactions do not occur. This then also defines a limit for the spatial resolution in the simulation: scales below or comparable to the softening length are not resolved, and the behavior on these small scales is affected by numerical artefacts (Schneider 2002, pg. 294).

Forces are computed using Newton's formulation of gravity ([check Hockney why](#)) so that the force on the  $i^{\text{th}}$  particle is

$$F_i = \sum_{j \neq i} \frac{M^2(\mathbf{r}_j - \mathbf{r}_i)}{|\mathbf{r}_j - \mathbf{r}_i|^3}, \quad (199)$$

thus the sum of forces exerted by all the other particles, where these are periodically extended. However, the computation of the force acting on individual particles by summation, as in equation (199), is not feasible in practice, due to the large number of particles being tracked. To handle this problem, one evaluates the force in an approximate way. One first notes that the force experienced by the  $i^{\text{th}}$  particle, exerted by the  $j^{\text{th}}$  particle, is not very sensitive to small variations in the separation vector  $\mathbf{r}_i - \mathbf{r}_j$ , as long as these variations are much smaller than the separation itself. Except for the nearest particles, the force on the  $i^{\text{th}}$  particle can then be computed by introducing a grid into the cube and shifting the particles in the simulation to the nearest grid point<sup>23</sup>. With this, a discrete mass distribution on a regular grid is obtained. The force field of this mass distribution can then be computed by means of a FFT. However, the introduction of the grid establishes a lower limit to the spatial force resolution; this is often chosen such that it agrees with the softening length. Because the size of the grid cells also defines the spatial resolution of the force field, it is chosen to be roughly the mean separation between two particles, so that the number of grid points is typically of the same order as the number of particles. This is called the PM (particle-mesh) method. To achieve better spatial resolution, the interaction of closely neighbouring particles is considered separately. Of course, this force component first needs to be removed from the force field as computed by FFT. This kind of calculation of the force is called the P<sup>3</sup>M (particle-particle particle-mesh) method (Schneider 2002, pg. 295).

The initial conditions of the simulations are set at high redshift ( $z \sim 200$ ) with the particle positions chosen such that the resulting mass distribution resembles a Gaussian random field with the theoretical (linear) power spectrum  $P(k, z)$  of the chosen cosmological model. The particle positions are then traced with a time step that is either chosen such that it is short enough to monitor the interactions of the most densely packed particles or a variable time step is given individually to the particles based on their proximity to neighbours (the latter is more efficient) (Schneider 2002, pg. 295).

Through all of this, it is clear that the main limitations of large-scale simulations are their limited temporal, spatial, and mass resolutions.

### USE IN ASTROPHYSICS

An important contribution that numerical simulations have made is through their use in establishing the standard model of cosmology. Figure 52 shows the evolution of structure formation for different cosmological models. The parameters for these simulations and their initial conditions were chosen such that the resulting density distributions at  $z=0$  were as similar as possible. From this it is easy to see the redshift evolution of the density field for the different cosmologies by analyzing the data at  $z=1, 3$ . For example, we can see that considerably less structure has formed at high redshift in the SCDM model compared to the others. Through subsequent comparisons to high redshift surveys of galaxy clustering we were able to determine that the matter density

<sup>23</sup> In practice, the mass of a particle is distributed to all 8 neighbouring grid points, with the relative proportion of the mass depending on the distance of the particle to each of these grid points (i.e., CIC interpolation).

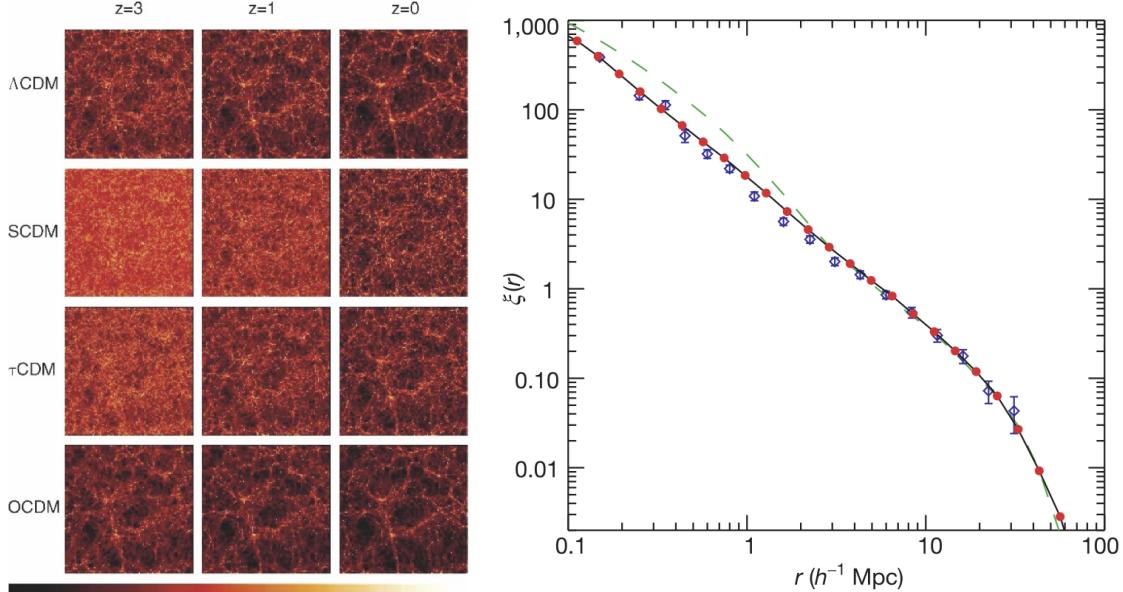


FIG. 52.—(left) Simulation by the VIRGO Consortium using  $256^3$  particles in box of side length  $240 \text{ Mpc}/h$  for different cosmological models:  $\Omega_{\text{mat}} = 0.3$ ,  $\Omega_\Lambda = 0.7$  ( $\Lambda\text{CDM}$ ),  $\Omega_{\text{mat}} = 1.0$ ;  $\Omega_\Lambda = 0.0$  (SCDM and  $\tau\text{CDM}$ );  $\Omega_{\text{mat}} = 0.3$ ,  $\Omega_\Lambda = 0.0$  (OCDM). The two Einstein-de Sitter models differ in the shape of their power spectra. Image taken from Schneider (2002). (right) Correlation function of galaxies at  $z = 0$  (filled circles connected by the solid curve), computed from the Millennium simulation in combination with semi-analytic models of galaxy evolution. This is compared to the observed galaxy correlation function as derived from the 2dFGRS (diamonds with error bars). The dashed green curve shows the correlation function of dark matter. Image taken from (Springel et al. 2005).

in the universe is considerably smaller than the critical density, thus favouring the  $\Lambda\text{CDM}$  model. This is confirmed by performing statistical tests on the distribution of structure within the volume. Examples include binning DM halos and comparing to Press-Schechter model, comparing the distribution and size of voids, comparing to observational galaxy surveys, etc. Another important contribution from these simulations is to use the computed spatial distribution of particles to construct the non-linear region of the power spectrum  $P(k, z)$ , which is only accessible through direct simulation (Schneider 2002, pg. 297).

Probably the most famous numerical simulation of large-scale structure is the Millennium Simulation (MS). This used a box of side length  $500 \text{ Mpc}/h$  with  $2160^3 \sim 10^{10}$  particles. With this choice of parameters it was possible to spatially resolve the halos of galaxies (the spatial resolution was  $\sim 5 \text{ kpc}/h$ ) at the same time as having a large enough volume to trace the evolution of large clusters (Schneider 2002, pg. 297).

A main aspect of the MS involved large-scale structure formation. The first thing to do was to construct the distribution of collapsed dark matter halos and compare this to the Press-Schechter model. This was found to agree to quite well except that it underestimated the abundance of very massive halos and overestimated the abundance of lower-mass halos. By tracing the evolution of the largest halos in the box the MS was able to shed some light on galaxy clusters and quasars. They found that the first halos to have formed in the box continue to grow into what we now observe as the centres of very massive galaxy clusters (Springel et al. 2005). Assuming that one can identify the largest halos as potential quasars suggests that quasar host galaxies may be identified as the central galaxies in clusters. This may provide an explanation as to why so many central, dominating cluster galaxies show AGN activity, though with smaller luminosities due to smaller accretion rates (Schneider 2002, pg. 400). This result from the MS helped to support the claim that such extremely rare objects like quasars can actually form in a  $\Lambda\text{CDM}$  cosmology.

The MS also had a large focus on galaxy formation and evolution. The two-point correlation function computed at  $z = 0$  is shown in Figure 52. There is good agreement between the simulated function and that which is observed; both show a nearly perfect power law. This is in contrast to the more complex behaviour exhibited by the DM correlation function. In particular, the correlation function of galaxies deviates from that of DM on small scales, implying a scale-dependent bias factor (Schneider 2002, pg. 398). The simple power law shape of the galaxy correlation function is actually a coincidence; if you look at its form for galaxy subsamples (i.e., separated galaxies by colour) it does not show a power-law form. See Schneider (2002, Figure 9.40, pg. 399) for plots of the matter distribution in the Millennium Simulation overlaid with the distribution of galaxies as determined by the semi-analytic model of galaxy formation that they incorporate. This plot clearly shows that at high redshifts there are more blue galaxies because their stellar populations are young, while at lower redshift there are many more red galaxies with old stellar populations. The MS also verified that the baryonic acoustic oscillations present in the matter power spectrum leave a signature on the galaxy distribution (Springel et al. 2005, Figure 6).

**QUESTION 18**

**What is the Sunyaev-Zeldovich effect and where is it detected? What are some challenges in using SZ measurements to determine the values of the cosmological parameters?**

### QUESTION 18

**What is the Sunyaev-Zel'dovich effect and where is it detected? What are some challenges in using SZ measurements to determine the values of the cosmological parameters?**

The **Sunyaev-Zel'dovich (SZ) effect** is an example of inverse Compton scattering, applied to the CMB. Reverse Compton scattering happens when low energy photons interact with extremely high energy particles (mostly electrons), where the end result is the particles transferring energy and momentum to the photons. This is in contrast to the “regular” Compton scattering where a high energy photon collides with a stationary electron, and an elastic collision occurs transferring energy to the electron (Yevgeni).

The basic physics of the SZ effect is simple. Since the thermal temperature of electrons is strongly coupled to the density of their environment, the dominant contribution to the SZ effect comes from high-density gas found in the ICM. Clusters of galaxies have masses that often exceed  $10^{14} M_{\odot}$ , with effective gravitational radii,  $R_{\text{eff}}$ , of order Mpc. Any gas in hydrostatic equilibrium within a cluster’s gravitational potential well must have electron temperature  $T_e$  given by

$$E_e = kT_e \approx \frac{GMm_p}{2R_{\text{eff}}} \sim 1000 \text{ eV}. \quad (200)$$

At this temperature, thermal emission from the gas appears in the X-ray part of the spectrum, and is composed of thermal bremsstrahlung and line radiation (Birkinshaw 1999). On the other hand, we know that CMB photons are much cooler; even at the time of decoupling their wavelength peaked at  $\lambda \sim 1 \mu\text{m}$  with a corresponding energy of  $E_{\gamma} \sim 1 \text{ eV}$ . Since at  $z \lesssim 10$  the universe is mostly ionized, and from equation (200)  $E_e \gg E_{\gamma}$  in the ICM, CMB photons passing through a galaxy cluster will be inverse-Compton scattered by free electrons (Yevgeni). Hence, scattered CMB photons will receive a boost of energy on average, and therefore an increase in frequency. As a consequence, this scattering leads to a reduced number of photons at lower energies, relative to the Planck spectrum, and higher energy photons being added; see Figure 53.

We have so far concluded that the SZ effect arises from passage of CMB photons through the gaseous atmospheres of galaxy clusters. Of course, passage of radiation through any electron population with significant energy content will produce a distortion of the radiation’s spectrum. Hence, other candidate locations that it can occur are through the bulk ionized content of the IGM as a whole and through ionized gas close to the MW (Birkinshaw 1999).

#### COSMOLOGICAL PARAMETERS

The simplest cosmological use of the SZ effect is to prove that the CMB is genuinely a cosmological phenomenon: the appearance of an effect from a cluster of galaxies at high redshift proves that the CMB originates at from some more distant epoch in the universe. However, it is as a probe of cosmological parameters, and as a distance-independent probe of earlier phases of the universe that the SZ effect has attracted most interest, and such uses of the effect are the focus of this section (Birkinshaw 1999).

Its use as a distance measure follows the same idea as other distance-measuring techniques that depend on a comparison of the emission and absorption of radiation from gas: the surface brightness of the gas in emission is proportional to the line-of-sight integral of some density squared,

$$E \propto \int n_e^2 dl, \quad (201)$$

while the absorption of some background source of radiation is proportional to the optical depth

$$A \propto \int n_e dl. \quad (202)$$

Thus if both the emission from the gas,  $E$ , and its absorption,  $A$ , can be measured, the quantity  $A^2/E$  is a density-weighted measure of the path-length through the gas. If the structure of the gas is known, and its angular size,  $\theta$ , can be measured, then the angular diameter distance of the gas can be estimated from  $A^2/(E\theta)$  (Birkinshaw 1999).

More specifically, the magnitude of the SZ effect is direction proportional to the temperature of the gas (which determines the particle energies) and the size of the cluster along the line of sight, and is inversely proportional to the mean free path of photons through the cloud:

$$\frac{\Delta I_{\nu}}{I_{\nu}} \propto \frac{LT_{\text{gas}}}{\lambda_{\text{mfp}}} \propto n_e LT_{\text{gas}}. \quad (203)$$

This is basically the absorption term in equation (202) integrated over path. On the other hand, the surface brightness of the X-ray radiation,  $I_X$ , behaves as

$$I_X \propto L n_e^2, \quad (204)$$

and represents the emission term in equation (201). Since  $T_{\text{gas}}$  can be measured from the X-ray spectrum, we can combine equations (203) and (204) to

$$\frac{\Delta I_{\nu}}{I_{\nu}} \propto \sqrt{L I_X}. \quad (205)$$

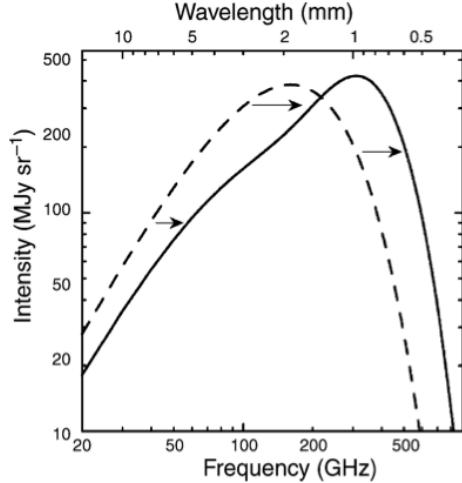


FIG. 53.— The influence of the SZ effect on the CMB. The dashed curve represents the Planck distribution of the unperturbed CMB spectrum, the solid curve shows the spectrum after the radiation has passed through a cloud of hot electrons. The magnitude of this effect, for clarity, has been very much exaggerated in this sketch. In general, a distortion of the CMB is produced where photons  $\nu \gtrsim 217$  GHz are boosted and lower frequencies are suppressed. Image taken from Schneider (2002).

By assuming that the cluster, or more importantly the gas within it, is spherical the angular diameter distance to the cluster is

$$d_A = \frac{R}{\theta} \sim \frac{L}{\theta} \propto \left( \frac{\Delta I_\nu}{I_\nu} \right)^2 \frac{1}{I_X}. \quad (206)$$

Hence, the angular-diameter distance can be determined from the measured SZ effect, the X-ray temperature of the ICM, and the surface brightness in the X-ray domain (Schneider 2002, pg. 254).

Measurements of the angular-diameter distance,  $d_A$ , at the redshift of the cluster using a combination of SZE and X-ray observations allow us to construct “ $d_A$  vs.  $z$ ” plots that can be used to infer the expansion history of the universe (in analogy to  $d_L(z)$  information obtained from SNe observations). At small redshift  $d_A(z)$  is a function of  $q_0$  and therefore this method allows us to constrain the values of  $\Omega_{\text{mat}}$  and  $\Omega_\Lambda$  (Khedekar & Majumdar 2010). Furthermore, since  $d_A$  depends on  $H_0$ , this method can be used to estimate the Hubble constant. This is a direct method of measuring the distance of a cluster of galaxies and the value of the Hubble constant: it can be applied at large cosmological distances without any intervening chain of distance estimators (as in the usual distance ladder). The distance estimate relies on simple physics – the properties of a fully ionized gas held nearly in hydrostatic equilibrium in the gravitational potential well of a cluster of galaxies (Birkinshaw 1999).

Since the SZ effect is proportional to the electron number density, knowledge of a cluster’s total mass, from dynamic methods or X-ray observations, can provide a measure of the baryon fraction in that cluster. Therefore, a large enough survey can yield global estimates of the baryon fraction,  $\Omega_{\text{bary}}/\Omega_{\text{mat}}$ , in the universe (Yevgeni).

#### OBSERVATIONAL DIFFICULTIES

First of all, the SZ effect is a relatively weak signal ( $\Delta T \sim 1$  mK) that dominates the CMB only at small angular scales associated with secondary anisotropies. Moreover, since the SZ effect is distinguished from the CMB by a change in spectral shape, observations must be made in several wavebands. In addition, X-ray measurements of the perturbing cluster are required for follow-up observations and for determination of  $d_A$  in equation (206). Such observations are generally costly since they demand detailed targeted observations of a single cluster. This is burdensome, since in order to provide statistically significant results, a large number of samples must be used (Yevgeni). Systematic errors also arise from the assumptions made of the physical form of cluster atmospheres that went into the derivation of equation (206); we considered a simple model, but more complicated models still assume some sort of profile for the ICM, usually an isothermal sphere (Birkinshaw 1999).

A number of observational techniques have been employed for measuring the SZ effect of galaxy clusters. The most primitive were implemented in the 70’s when single-dish radio observations were made. With relatively large beam sizes on the order of a few arcminutes at microwave frequencies, this approach could only be used for relatively nearby clusters. Another method is to use bolometric techniques. The principal advantage of a bolometric system is the high sensitivity that is achieved, but these devices are also of interest because of their frequency range: at present they provide the best sensitivity for observing the microwave background outside the Rayleigh-Jeans part of the spectrum, and hence for separating the thermal and kinematic components of the SZ effect using their different spectral shapes. Furthermore, the best systems consist of several detectors arranged in an array, and some provide simultaneous operation in several bands. A suitable choice of differencing between elements of the array reproduces many of the sky-noise subtraction properties of radiometric observing, and the multiband capability holds out the hope of rapid spectral measurements. Until recently, radiometry and bolometry have been the primary methods. Currently, interferometry techniques are becoming more common because of the increased angular resolution on the sky that they achieve (Birkinshaw 1999).

The **kinetic Sunyaev-Zel'dovich (kSZ) effect** is caused by Doppler shifting of CMB photons that Thomson scatter off of ionized gas clouds possessing some non-zero bulk velocity relative to the CMB rest frame. This leads to temperature fluctuations corresponding to hot (cold) spots in the CMB if the ionized gas is moving toward (away from) the observer. In contrast to the tSZ effect, this process maintains the blackbody shape of the CMB spectrum. Moreover, the amplitude of the kSZ effect is proportional to the bulk peculiar velocity of electrons and is not weighted by temperature like the tSZ effect is. Consequently, the kSZ effect can be used to trace the overall structure of the IGM, down to small mass scales. Unfortunately, the combination of the kSZ effect being much weaker than the tSZ effect and not producing spectral features that can be easily extracted from the CMB has made it relatively difficult to detect (from my thesis proposal).

Thermal fluctuations in the CMB from the kSZ effect result from the projection of electron motion along the line of sight meaning that a 2D sky map entangles information from all redshifts contributing to the effect. Hence, a redshift-sensitive interpretation of the observed signal requires a theoretical understanding of the physics governing the kSZ effect. Namely, since the kSZ effect is produced by Doppler shifts from electrons in bulk motion, the observed signal depends on the details of structure formation and evolution of the IGM during reionization. In fact, the literature often refers to two separate components of the kSZ signal: the “homogeneous” signal from the post-reionization epoch and the “inhomogenous” signal arising from the patchy reionization model. Measurements of the kSZ signal through ground-based observatories such as ACT and SPT, will allow us to learn more about the epoch of reionization (from my thesis proposal).

**QUESTION 1**

**What is a stellar Initial Mass Function (IMF)? Sketch it. Give a couple of examples of simple parametric forms used to describe the IMF.**

## QUESTION 1

**What is a stellar Initial Mass Function (IMF)? Sketch it. Give a couple of examples of simple parametric forms used to describe the IMF.**

### STAR FORMATION

We begin by describing the basic scenario of star formation from the collapse of a molecular cloud using a simplified model based on the virial theorem, as outlined in Carroll & Ostlie (2007, pg. 412). The virial theorem,  $2K + U = 0$ , describes the condition of equilibrium for a stable, gravitationally bound system. In particular, if twice the total internal kinetic energy of a molecular cloud,  $2K$ , exceeds the absolute value of the gravitational potential energy,  $U$ , the force due to the gas pressure will dominate the force of gravity and the cloud will expand. On the other hand, if the internal kinetic energy is too low, the cloud will collapse. The boundary between these two cases describes the critical condition for stability when rotation, turbulence, and magnetic fields are neglected.

Assuming a spherical cloud of constant density  $\rho_0$ , the gravitational potential energy is roughly

$$U \sim -\frac{3}{5} \frac{GM^2}{R}, \quad (207)$$

where  $M$  and  $R$  are the mass and radius of the cloud, respectively. We may also estimate the total internal kinetic energy of the cloud as

$$K = \frac{3}{2} Nkt = \frac{3}{2} \frac{MkT}{\mu m_H}, \quad (208)$$

where  $N$  is the total number of particles in the cloud and  $\mu$  is their mean molecular mass. With equations (207) and (208), the condition that  $2K = |U|$  leads to the definition of a critical mass,

$$M_J = \left( \frac{5kT}{G\mu m_H} \right)^{3/2} \left( \frac{3}{4\pi\rho_0} \right)^{1/2}, \quad (209)$$

often referred to as the **Jeans mass**. If the mass of the cloud exceeds the Jeans mass, the cloud collapses under its gravity. Note that this derivation neglected the important fact that there must exist an external pressure on the cloud due to the surrounding ISM. The critical mass derived when this factor is considered is called the **Bonnor-Ebert mass**, which has a slightly smaller value than the Jeans mass, since an external pressure will add to the compression forces on the cloud (Carroll & Ostlie 2007, pg. 414).

Once the cloud has reached the Jeans mass it will begin to collapse on itself. If the collapse is **isothermal** so that the temperature remains constant then equation (209) dictates that  $M_J$  will decrease as the density of the cloud naturally increases. An isothermal collapse will occur so long that the cloud remains optically thin so that it can efficiently radiate away the release of gravitational potential energy. After collapse has begun, any initial inhomogeneities in density will cause individual sections of the cloud to reach the local Jeans mass independently and begin to collapse, proceeding smaller features within the original cloud. Such a cascading collapse appears to lead to the formation of large numbers of smaller objects. However, the fragmentation process eventually ceases during the transition to an **adiabatic** collapse, where no energy is able to escape the cloud. This occurs when the cloud becomes dense enough to reabsorb its own radiation, so that the temperature must increase. This has the effect of imposing a minimum mass of objects that can form from this process (Carroll & Ostlie 2007, pg. 418).

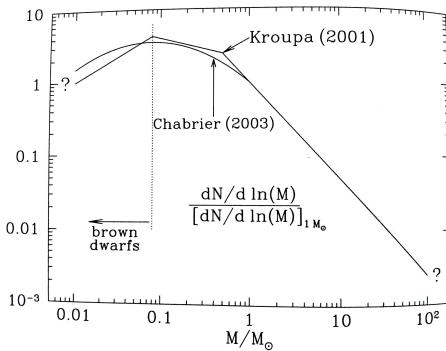


FIG. 54.— Shown here is  $\xi(\log m)$ , the number of stars formed per logarithmic interval in stellar mass  $m$ , normalized to the value at  $m = 1 M_\odot$ . Both the Kroupa (2001) and Chabrier (2003) approximations to the stellar IMF are plotted. Image taken from Draine (2011).

off of  $m_L \sim 0.1 M_\odot$  since less massive stars do not ignite hydrogen burning, and an upper cut-off of  $m_U \sim 100 M_\odot$  since more massive stars have not been observed. Such massive stars would be difficult to observe due to their short lifetime; furthermore the theory of stellar structure suggests that more massive stars probably cannot form a stable configuration due to excessive radiation pressure (Schneider 2002, pg. 132).

The shape of the IMF is also subject to uncertainties, with the most common usage being a power law of the form:

$$\xi(m) = \frac{dn}{dm} \propto m^{-\alpha} \rightarrow \xi(\log m) \propto m^{-\alpha+1}, \quad (210)$$

where  $n$  is the number density of stars. Salpeter (1955) take  $\alpha = 2.35$  while Kroupa (2001) adopt a broken power law with  $\alpha = 1.3$  for  $m \leq 0.5 M_\odot$  and  $\alpha = 2.3$  for  $m \geq 0.5 M_\odot$ . Another common form is referred to as the generalized Rosin-Rammler function:

$$\xi(\log m) \propto m^{-\alpha+1} \exp \left[ - \left( \frac{B}{m} \right)^\beta \right], \quad (211)$$

which asymptotically recovers the power law behaviour at large  $m$  and produces a lognormal form in the opposite limit. Chabrier (2003) construct a lognormal form of the IMF valid in the low mass regime, with a model calibrated to observations of star clusters. Figure 54 compares the Chabrier (2003) and Kroupa (2001) IMFs.

Since stars evolve off the MS after a certain age, the PDMF, which is determined from the observed present-day luminosity function, differs from the IMF, defined as the number of stars that were originally created per mass interval in the MW. The determination of the IMF from the PDMF involves coupling the SFR (i.e. the number of stars formed per time interval along galactic evolution) to knowledge of the turnoff mass. The turnoff mass is defined as the mass for which the age at which the star starts evolving off the MS equals the age of the MW. Together, these quantities make up the so-called stellar creation function which can be used to determine the IMF from the PDMF (Chabrier 2003).

**QUESTION 2**

**Describe the orbits of stars in a galactic disk and in galactic spheroid.**

## QUESTION 2

**Describe the orbits of stars in a galactic disk and in galactic spheroid.**

The short answer to this question is that stars in the galactic disk move in nearly circular orbits with small perturbations resulting in minor oscillations within the radial and vertical directions. The orbits of stars in the Galactic halo (i.e. spheroid) are randomly oriented and often highly elliptical with paths that send them plunging through the plane of MW. Orbit in the Galactic bulge are similar to those of the halo. Figure 55 illustrates the typical orbits within the different components of the MW. The determination of allowed stellar orbits involves considering the motion of individual stars under the influence of the large-scale potential caused by the overall mass distribution of the MW. Below we attempt to do this for the bulge and disk components of the MW. We then explore related topics including Galactic disk formation scenarios and the stellar orbits associated with elliptical galaxies.

### GALACTIC BULGE

We begin by considering the case of a star moving in a static, spherically symmetric potential. This potential is appropriate for globular clusters and can also be used to model the Galactic bulge. The motion of a star in a centrally-directed, spherically symmetric potential is simplified by conservation of angular momentum. In particular, since the acceleration is directed inward as  $\vec{g} = g(r)\hat{r}$ , we naturally find that  $d\vec{L}/dt = d(\vec{r} \times d\vec{r}/dt)/dt = 0$ . Geometrically,  $\vec{L}$  is a vector perpendicular to the plane defined by the instantaneous position and velocity vectors of the star in question. Since this vector is constant, we conclude that the stellar orbit must be confined to a plane.

In general, either bound or unbound orbits can arise in this configuration. Of course, further details on the exact shape of the orbit depend on the details of the potential landscape and usually require numerical integration of the equation of motion. Typically, however, the shape takes the form of a rosette, as depicted in Figure 56. A rosette is an example of an open orbit – one in which the orbital period is not an integer multiple of the radial period (Binney & Tremaine 1994, pg. 107).

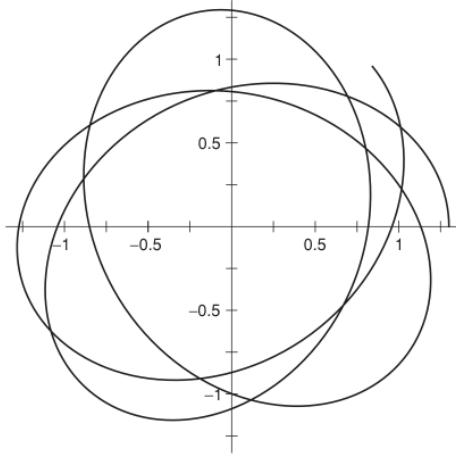


FIG. 56.— The typical path of a star in a spherically symmetric potential landscape traces out the shape of a rosette. Image taken from Sparke & Gallagher (2007).

complex motion. In general, we find that the motion of the star will trace out a rosette augmented with perturbations perpendicular to the plane of the disk, causing oscillatory motion in the  $z$  direction. Such orbits are often referred to as loop orbits. The physical phenomena responsible for these oscillations is simply the flattened potential of the Galactic disk. The orbits can achieve a significant range of vertical heights but – due to conservation of energy and angular momentum – remain bounded by some minimum and maximum values in both the radial and vertical directions. An example of this type of motion is shown in Binney & Tremaine (1994, Figure 3.3), obtained through numerical means.

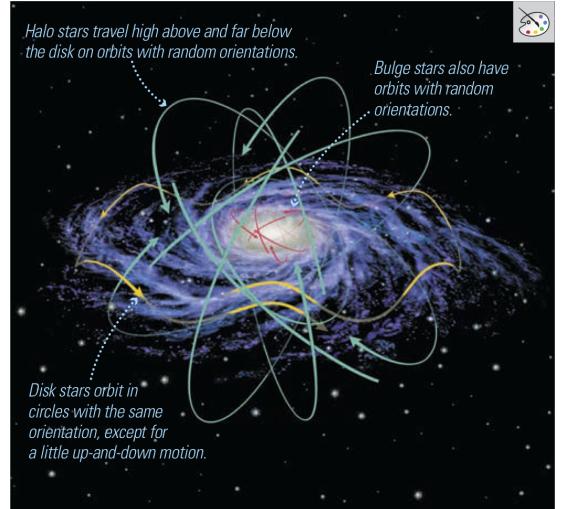


FIG. 55.— Characteristic orbits of disk stars (yellow), bulge stars (red), and halo stars (green). The oscillatory motion exhibited in the direction orthogonal to the disk is largely exaggerated in the yellow curves. Image taken from Bennett et al. (2012).

Two familiar examples of spherically symmetric potentials are the spherical harmonic oscillator potential and the Kepler potential. The former is generated by a homogenous sphere while the later arises from the gravitation interaction of a point mass. It turns out that orbits in a harmonic potential complete a radial oscillation in the time the azimuthal angle  $\phi$  has increased by  $\Delta\phi = \pi$ . On the other hand, Keplerian orbits complete a radial oscillation by the time  $\phi$  has increased by  $2\pi$ . Both types of orbits are therefore closed and do not trace out rosettes. Since galaxies are more extended than point sources and less extended than homogeneous spheres, a typical star in a spherical galaxy completes a radial oscillation after  $\phi$  has increased by an amount that lies between the two extremes:  $\pi < \Delta\phi < 2\pi$ . Thus, we expect a star to oscillate from its apocenter through its pericenter and back in a shorter time than is required for one complete azimuthal cycle about the galactic centre (Binney & Tremaine 1994, pg. 109).

### GALACTIC DISK

To understand the allowed orbits in the galactic disk it is useful to consider an axisymmetric potential. For this purpose it is useful to adopt a cylindrical coordinate system  $(r, \phi, z)$  with an origin at the Galactic centre and is symmetric about the plane  $z = 0$ . We begin by noting that stars that are confined to the equatorial plane of an axisymmetric galaxy have no way of perceiving that the potential in which they move is not spherically symmetric. Hence, their orbits will be identical to what we considered previously and will therefore resemble rosettes (Binney & Tremaine 1994, pg. 114).

Obviously, stars that are not confined to the equatorial disk will exhibit more complex motion. In general, we find that the motion of the star will trace out a rosette augmented with perturbations perpendicular to the plane of the disk, causing oscillatory motion in the  $z$  direction. Such orbits are often referred to as loop orbits. The physical phenomena responsible for these oscillations is simply the flattened potential of the Galactic disk. The orbits can achieve a significant range of vertical heights but – due to conservation of energy and angular momentum – remain bounded by some minimum and maximum values in both the radial and vertical directions. An example of this type of motion is shown in Binney & Tremaine (1994, Figure 3.3), obtained through numerical means.

### GALACTIC DISK FORMATION

The basis picture of galactic disk formation involves the gravitational collapse of a dark matter halo and its subsequent relaxation to virial equilibrium. Baryons begin to collect in the potential well of the halo and become shock-heated to the halo virial temperature through the release of gravitational potential energy. The gas cools radiatively (i.e. through bremsstrahlung) from inside out, gradually building up the disk and forming stars quiescently (van der Kruit & Freeman 2011). Due to conservation of angular momentum, a centrifugally thin disk forms through this process, with the coolest portions of the cloud being the most likely to form stars. In general, the star formation rate will vary between galaxies, but for the MW, the circular motion of stars in the thin disk seems to suggest that they were born after the thin disk formed. Star formation within the thin disk began roughly 10 Gyr ago and, given the wide range of stellar ages observed, it appears as though disk formation has been an extended process continuing to the present time. Thick disk formation is not well understood; one proposed scenarios argues that the thick disk arose from puffing of the thin disk through minor mergers and tidal interactions (van der Kruit & Freeman 2011).

### STELLAR ORBITS IN ELLIPTICAL GALAXIES

Elliptical galaxies are in general triaxial, with a potential landscape that can be thought of as a spherically symmetric potential flattened along two axes. Accordingly, they admit a wider range of orbits than the approximately planar, precessing, rosette orbits we have already seen. Stellar orbits in elliptical galaxies come in two types: the so-called box and tube orbits. Box orbits do not exhibit any particular sense of rotation about the centre and trace out three-dimensional shapes resembling a distorted box. On the other hand, tube orbits do possess an ordered sense of rotation about the centre and consequently contain an axis of rotation that they will never approach (i.e. tube orbits trace out a three-dimensional shape that has a hole in the middle). The most prominent tube orbits are those for which the axis of rotation is aligned with the shortest axis of the potential landscape; these are slightly distorted versions of the loop orbits considered previously. Another type of tube orbit exists for which the axis of rotation is aligned with the longest axis of the potential landscape. The population of these two types of tube orbits will in general be different, explaining why velocity maps of ellipticals do not resemble crosses. Orbits for which the axis of rotation is oriented along the intermediate axis of the potential are unstable. Binney & Tremaine (1994, Figure 3.20) provides a representation of the different box and tube orbits for a typical elliptical galaxy.

Triaxial potentials can also give rise to chaotic, or irregular orbits. These orbits can be created by adding a cusp or core to the potential landscape, physically motivated by central black holes. These features tend to deflect orbits thereby adding random motions to their trajectories.

**QUESTION 3**

**Every now and then a supernova explosion occurs within 3 pc of the Earth. Estimate how long one typically has to wait for this to happen. Why are newborn stars likely to experience this even when they are much younger than the waiting time you have just estimated?**

### QUESTION 3

**Every now and then a supernova explosion occurs within 3 pc of the Earth. Estimate how long one typically has to wait for this to happen. Why are newborn stars likely to experience this even when they are much younger than the waiting time you have just estimated?**

To estimate the supernova rate within the solar neighbourhood we must determine the number of stars that are capable of going supernova and their associated destruction rate. For the former we can use the IMF with the usual caveat that this will not faithfully trace the PDMF since stellar lifetime is a function of mass. However, for the massive stars considered here we may suppose that their lifetimes are short compared to the age of the galactic disk, implying that the region of the PDMF we are interested in takes the same shape as the IMF. To determine their destruction rate we will simply take the galactic SFR and assume equilibrium between formation and destruction.

We have previously seen that the IMF, denoted  $\xi(m) dm$ , specifies the number density of stars in the mass interval of width  $dm$  around  $m$ , at the time of stellar birth. We can impose a lower limit on the IMF of  $m_L = 0.1 M_\odot$  since below this mass the pressure and temperature within the core are insufficient to ignite hydrogen fusion (Schneider 2002, pg. 427).

Our first step is to calculate the number of stars capable of going supernova within the solar neighbourhood. By integrating over the IMF we can determine the number,  $N_*$ , of supernova progenitors that form per unit mass of stellar material:

$$N_* = \frac{\int_{m_{\min}}^{\infty} \xi(m) dm}{\int_{m_L}^{\infty} m \xi(m) dm}, \quad (212)$$

where  $m_{\min}$  is the minimum mass above which stars will be identified as supernova progenitors; we take  $m_{\min} = 8 M_\odot$  from Schneider (2002, pg. 48). To evaluate equation (212) we must assign some form to  $\xi(m)$ ; a good choice is the Salpeter (1955) power law  $\xi(m) \propto m^{-2.35}$ . Using this we evaluate equation (212) to obtain

$$N_* = \frac{\int_{m_{\min}}^{\infty} m^{-2.35} dm}{\int_{m_L}^{\infty} m^{-1.35} dm} \approx \frac{1}{143} \text{ stars } M_\odot^{-1}. \quad (213)$$

Hence, roughly one supernova progenitor forms for every 143 solar masses of stellar material to accrete within the MW. With the assumption of equilibrium between stellar formation and destruction, we can turn this statement around to say that we expect a supernova explosion for every 143 solar masses of stellar material to accrete within the solar neighbourhood.

Obviously, the next step is to calculate how long it takes to form 143 solar masses worth of stars within 3 pc of the Sun. This is a simple calculation that requires knowledge of the SFR within the solar neighbourhood. Note that we can approximate the total SFR of the MW by dividing its stellar mass by its age:  $SFR \approx 10^{10} M_\odot / 10^{10} \text{ yr} = M_\odot \text{ yr}^{-1}$ . More direct measures of the SFR can be obtained through measuring (1) FIR luminosity of warm dust heated by hot young stars, (2) H $\alpha$  emission from the H II regions associated with young stars, or (3) UV blackbody radiation from hot young stars (Schneider 2002, pg. 387). If we assume that star formation is constant throughout the Galactic disk, then the SFR within the solar neighbourhood is simply

$$SFR_{ngbd} = SFR_{MW} \frac{4R_{ngbd}^3}{3R_{disk}^2 h_{disk}} \approx 5 \times 10^{-10} M_\odot \text{ yr}^{-1}, \quad (214)$$

where  $R_{ngbd} = 3 \text{ pc}$  and we take  $R_{disk} = 15 \text{ kpc}$  and  $h_{disk} = 300 \text{ pc}$  as the radius and height of the Galactic disk.

The average time between supernova explosions is therefore

$$\Delta t = \frac{143 M_\odot}{5 \times 10^{-10} M_\odot \text{ yr}^{-1}} \approx 300 \text{ Gyr}. \quad (215)$$

Though this value is much larger than the current age of the universe, newborn stars are likely to experience this since they typically form in active star forming regions. In these regions we expect a significant fraction of high mass stars to form around the same time. They will exhaust their nuclear fuel and explode as supernova within the lifetime of the less massive stars.

**QUESTION 4**

**Galactic stars are described as a collision-less system. Why?**

#### QUESTION 4

**Galactic stars are described as a collision-less system. Why?**

We will begin by showing that a direct collision between galactic stars is highly unlikely. To do this we must first discuss some properties of the stellar distribution within the MW. The MW contains about  $10^{11}$  stars, where most of these travel on nearly circular orbits in a thin disk with a radius of  $\sim 10$  kpc and thickness of  $\sim 1$  kpc. The typical circular speed of a star in the disk is  $\sim 200$  km/s, so that the time required to complete an orbit at a distance of 10 kpc from the centre is  $3 \times 10^8$  years. The dispersion in the velocities of the stars at a given position is roughly 40 km/s (Binney & Tremaine 1994, pg. 4).

With this information we can now calculate the typical mean free path of a star before it collides with another star. For an assembly of particles moving on straight-line orbits, the mean free path is  $\lambda = 1/(n\sigma)$ , where  $n$  is the number density of stars and  $\sigma$  is their cross-section. We will make the crude assumption that all stars are like the Sun so that the cross-section for a collision is  $2\pi R_\odot^2$ , where  $R_\odot \sim 7 \times 10^{10}$  cm. If we spread the  $10^{11}$  stars uniformly over the think disk, then the stellar number density is  $n \sim 0.3 \text{ pc}^{-3}$ . We thus have that

$$\lambda = \frac{1}{n\sigma} \sim 10^{15} \text{ pc}, \quad (216)$$

which is  $\sim 10^4$  times larger than the observable universe! The interval between collisions is roughly  $\lambda/v$  where  $v$  is the random velocity of stars at a given location. With  $v = 40$  km/s we find the collision interval to be  $\sim 10^{19}$  years, much larger than the age of the Universe! For this reason we can completely ignore collisions between galactic stars and model them as a collision-less system of point particles. Note that we have ignored the increase in the collisional cross-section due to the gravitational attraction between stars. In general, this only enhances the collision rate by a factor of about 100 and therefore does not affect the conclusions we have just made (Binney & Tremaine 1994, pg. 4).

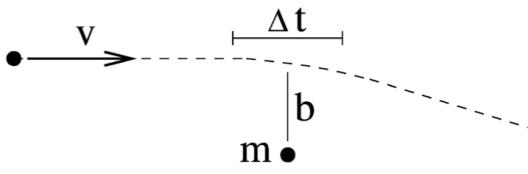


FIG. 57.— A star with velocity  $v$  passes by another star with impact parameter  $b$ . The star will be deflected with a velocity component perpendicular to its incoming motion. This component is roughly proportional to the gravitational acceleration at closest approach multiplied by the time  $\Delta t \approx 2b/v$  for which it spends interacting with the nearby star.

We have shown that direct collisions between galactic stars is incredibly unlikely. However, the term “collision-less system” is commonly used to refer to a system in which stellar encounters causing perturbations to stars’ orbits do not play a significant role (Binney & Tremaine 1994, pg. 489). So we will now shift gears and investigate this issue.

For this purpose we will define the **relaxation time**,  $t_{\text{relax}}$ , as the characteristic time in which a star changes its velocity distribution by  $\sim 90^\circ$  due to pair collisions with other stars (not physical collisions, but close gravitational encounters). We will assume we have a spherical ensemble of  $N$  stars, each with mass  $m$  and radius  $R$ , so that the total mass of the system is  $M = Nm$  and the mean stellar density is  $n = 3N/(4\pi R^3)$ .

We consider a star passing by another with an impact parameter  $b$ , as shown in Figure (57). Through gravitational deflection, the star obtains a velocity component perpendicular to the incoming direction of

$$v_\perp^{(1)} \approx a\Delta t \approx \left(\frac{Gm}{b^2}\right) \left(\frac{2b}{v}\right) = \frac{2Gm}{bv}, \quad (217)$$

where  $a$  is the acceleration at closest separation and  $\Delta t$  is the “duration of the collision”, which we approximate as  $\Delta t = 2b/v$  (Schneider 2002, pg. 95). This result can be derived more rigorously by integrating the perpendicular acceleration along the orbit (see, for example, Binney & Tremaine 1994, pg. 188).

As a star undergoes a variety of these pair collisions, it will acquire an accumulation of perpendicular velocity components. After many such collisions the expectation value of the sum of the individual components will tend to zero since the stars are oriented randomly. However, the mean square velocity perpendicular to the incoming directions does not vanish. To determine this value we integrate over all collision parameters  $b$ .

Skipping all of the details (for the gory details, see Schneider 2002, pg. 95) after a time  $t$  has elapsed, we arrive at

$$\langle |v_\perp|^2(t) \rangle = 2\pi \left(\frac{2Gm}{v}\right)^2 v t n \ln N. \quad (218)$$

We define the relaxation time by  $\langle |v_\perp|^2(t_{\text{relax}}) \rangle = v^2$ , so that

$$t_{\text{relax}} = \frac{1}{2\pi nv} \left(\frac{M}{2Rm}\right)^2 \frac{1}{\ln N} \approx \frac{R}{v} \frac{N}{\ln N} = t_{\text{cross}} \frac{N}{\ln N}, \quad (219)$$

where  $t_{\text{cross}} \equiv R/v$  is the crossing time-scale at which it takes a star to traverse the system.

If we consider a typical galaxy with  $t_{\text{cross}} \sim 10^8$  yr and  $N \sim 10^{11}$ , we find that the relaxation time is much larger than the age of the Universe. Hence, pair collisions do not play any role in the evolution of stellar orbits. The dynamics of the orbits are determined solely by the large-scale gravitational field of the system (Schneider 2002, pg. 94). However, other systems have much shorter relaxation times and are therefore likely to be influenced by stellar encounters. For a list of such systems see Table 3 which was taken from Binney & Tremaine (1994, pg. 489).

TABLE 3  
SYSTEMS FOR WHICH STELLAR ENCOUNTERS MAY BE IMPORTANT.

System	$N$	$t_{\text{cross}}$ [yr]	$t_{\text{relax}}$ [yr]	$t_{\text{age}}$ [yr]
Globular clusters	$\sim 10^5$	$\sim 10^5$	$\sim 10^9$	$\sim 10^{10}$
Galactic or open clusters	$\sim 10^2$	$\sim 10^6$	$\sim 10^7$	$\sim 10^8$
Galactic nuclei	$\sim 10^8$	$\sim 10^4$	$\sim 10^{10}$	$\sim 10^{10}$
Galaxy clusters	$\sim 10^3$	$\sim 10^9$	$\sim 10^{11}$	$\sim 10^{10}$

### QUESTIONS 5 AND 6

Given that only a tiny fraction of the mass of the ISM consists of dust, why is dust important to the process of star formation? The ISM mainly consists of hydrogen and helium which are very poor coolants. How, then, do molecular cloud cores ever manage to lose enough heat to collapse and form stars?

## QUESTIONS 5 AND 6

**Given that only a tiny fraction of the mass of the ISM consists of dust, why is dust important to the process of star formation? The ISM mainly consists of hydrogen and helium which are very poor coolants. How, then, do molecular cloud cores ever manage to lose enough heat to collapse and form stars?**

As opposed to collisionless DM, baryons have the property that they are dissipative, in the sense that they can lose energy through a myriad of radiative processes. Radiative release from a gas cloud results in a loss of thermal energy from the system with a corresponding decrease in internal pressure. As a result, the region will contract until pressure equilibrium is reestablished.

In the MW stars are formed within cool, dusty regions, and primarily in GMCs that pervade the Galactic disk. The likely sequence of events is that a region within a cool dust cloud becomes Jeans unstable and begins to collapse. The collapse may also be stimulated by external influences, such as the passage of a gas cloud through a spiral arm, through compression by the blast wave of a SNR, or through some other large-scale dynamical perturbation, including strong gravitational interactions or galactic collisions. A star can only form from the collapsing cloud if it can expel its gravitational binding energy. This is mostly accomplished through radiative dissipation which occurs until the cloud becomes optically thick to its own radiation. After this point the loss of binding energy is mediated by dust grains that are heated to  $\sim 100$  K and release their thermal energy as FIR blackbody radiation (Longair 2008, pg. 478). Obviously, there are a variety of cooling mechanisms exploited by the gas cloud; we explore the most important below.

### MOLECULAR COOLING

Since the first excited state of hydrogen has a high energy (that of the Ly $\alpha$  transition, thus  $E \sim 10.2$  eV), it can only be an efficient coolant at temperatures  $T \gtrsim 10^4$  K. Helium is, of course, even worse since its excitation temperature is higher than that of hydrogen (Schneider 2002, pg. 384). The next natural coolant to consider are molecules with two important examples being H<sub>2</sub> and CO. The former is thought to be particularly important in primordial metal-poor gas where it dominates cooling below  $\sim 10^4$  K (Bromm et al. 2002). The latter is often used as a tracer of H<sub>2</sub> and molecular clouds due to its relatively high abundance and easily observable emission lines (Carroll & Ostlie 2007, pg. 407). Molecular cooling takes place through rovibrational emission of thermally excited molecules (Kitayama et al. 2001).

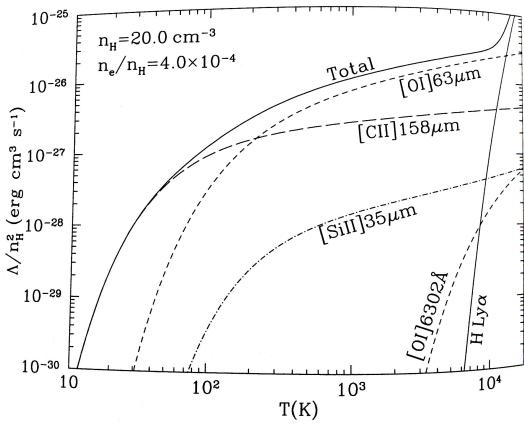


FIG. 58.— Cooling function for predominantly neutral gas (ionized fraction of  $x = 4 \times 10^{-4}$ ) versus temperature. Gas density is chosen to be suitable for conditions expected in the diffuse ISM of the MW, not for a collapsing gas cloud. Image taken from Draine (2011, pg. 340).

densities appropriate for those found in the diffuse ISM of the MW. For  $100 \lesssim T \lesssim 10^4$  K the fine structure lines of [C II] (158  $\mu\text{m}$ ) and [O I] (63  $\mu\text{m}$ ) are major coolants. As discussed previously, Ly $\alpha$  cooling is only important for temperatures above  $10^4$  K. Of course, as the collapsing gas cloud increases in density forbidden transitions will be suppressed by collisional deexcitation, and other metal lines will become dominant.

Metals also play an important role in enhancing the degree of fragmentation within the gas cloud. Fragmentation ultimately ceases when the heat released through gravitational collapse can no longer be radiated away efficiently. This implies that a greater number of small objects can form out of a more metal-rich gas cloud. Detailed calculations show that the lowest mass that is possible to form by successive fragmentation is about  $0.01 M_\odot$  (Dyson & Williams 1997, Section 8.1.2.).

### FIR DUST EMISSION

To understand the role of FIR grain emission in molecular cloud cooling we must first discuss dust extinction. This is heavily based on Mie Theory which states that the scattering cross section,  $\sigma_\lambda$ , of radiation with wavelength  $\lambda$  to a spherical grain of

Dust grains are important for molecular hydrogen cooling since they provide the dominant process of H<sub>2</sub> formation. The idea is that a hydrogen atom arrives at a grain and becomes bound to its surface. Initially, the binding energy may be weak enough that the H atom is able to diffuse (i.e. random walk) some distance on the grain surface, until it happens to arrive at a site where it is bound strongly enough that it becomes trapped in the sense that thermal fluctuations at the low temperature ( $\sim 20$  K) of the grain are unable to knock it loose. Subsequent H atoms arrive at random locations on the grain surface though occasionally two will meet up at one of these trapping locations. When this occurs they will react to form H<sub>2</sub> in an exothermic reaction that releases enough energy to eject the molecule from the grain surface. Gas phase formation of H<sub>2</sub> has substantially lower rates due to the complicated reaction channels required for its formation (Draine 2011, pg. 345). Note that the formation and destruction of H<sub>2</sub> are very sensitive to the presence of a radiation field. For instance, H<sub>2</sub> molecules are easily dissociated by Lyman-Werner photons (11.2 - 13.6 eV) while its gas phase formation is enhanced in a weakly ionized medium (Kitayama et al. 2001).

### METAL COOLING

Once the first generation of massive stars formed the fraction of heavy elements in the ISM could build up rapidly. The presence of metals in a collapsing gas cloud offers another channel for radiative deexcitation to remove thermal energy from the cloud. Figure (58) shows the cooling function for predominantly neutral gas as a function of temperature with

radius  $a$  goes like

$$\sigma_\lambda \propto \begin{cases} 0 & \lambda \gg a \\ a^3/\lambda & \lambda \gtrsim a \\ a^2 & \lambda \ll a \end{cases} \quad (220)$$

Consequently, long wavelength red light is not scattered as strongly as blue light, so starlight passing through intervening dust clouds becomes reddened as the blue light is removed; this process is known as **interstellar reddening** (Carroll & Ostlie 2007, pg. 401). Based on observations of UV extinction, scattering of visible light, and polarization of starlight, the interstellar grain population must have a broad size distribution, extending from sizes as small as  $a \approx 0.01 \mu\text{m}$  (or even smaller) up to  $0.2 \mu\text{m}$  (or even larger) (Draine 2011, pg. 243). Hence, from equation (220) we see that FIR and submm continuum emission from dust grains will generally be optically thin, allowing efficient energy transport out of the cloud. This is why continuum emission from dust plays an important role in dense regions that are otherwise optically thick to metal line and molecular radiation (Draine 2011, pg. 235).

We can understand the limiting behaviour of equation (220) through analogy to waves on the surface of a lake. If the wavelength of waves is much larger than an object in their way, such as a grain of sand, the waves pass by almost completely unaffected. On the other hand, if the waves are much smaller than the obstructing object, such as an island, they are simply blocked; the only waves that continue on are those that miss the island altogether. Similarly, at sufficiently short wavelengths, the only light we detect passing through the dust cloud is the light that travels between the particles. Note that Mie scattering is similar to Rayleigh scattering except that in the latter the sizes of the scattering molecules are much smaller than the wavelength of visible light, leading to  $\sigma_\lambda \propto \lambda^{-4}$  (Carroll & Ostlie 2007, pg. 401).

**QUESTION 7**

**What's the difference between a globular cluster and a dwarf spheroidal galaxy?**

## QUESTION 7

### What's the difference between a globular cluster and a dwarf spheroidal galaxy?

Observations show that the MW is embedded in a halo of globular clusters and an extended corona of dSph's. Globular clusters are compact, high-surface brightness objects, that consist of a collection of old stars believed to have formed from the collapse of a single molecular cloud. As such, they contain a single population of stars with uniform age and chemical composition. On the other hand, dSph's are diffuse and extended with low surface brightnesses and luminosities, and display complex star formation and chemical enrichment histories (Lotz et al. 2004). Globular clusters tend to be smaller and less massive with  $R_h \sim 5$  pc and  $M \sim 10^5 - 10^6 M_\odot$  (CITEME) while dSph's generally have  $R_h \sim 500$  pc and  $M \sim 10^7 - 10^8 M_\odot$  (Lotz et al. 2004; Carroll & Ostlie 2007, pg. 984). Here we are using  $R_h$  to denote the half-light radius as the radius within which half of the luminosity is contained<sup>24</sup>.

It has become customary to regard the presence of DM as the touchstone that allows us to unambiguously distinguish between galaxies and star clusters (van den Bergh 2008a). This is the main difference between globular clusters and dSph's: dSph's sit inside the gravitational potential wells of DM halos and the kinematics of their stars are increasingly dominated by this underlying DM halo, whereas globular clusters do not have associated DM halos. Although the smallest dSph's have masses and luminosities comparable to massive globular clusters, it is this feature that separates the two as physically different entities. dSph's are likely survivors from the hierarchical growth of structure formation spawned by small fluctuations in the DM density distribution of the early universe while globular clusters are more likely associated with star formation activity within their parent galaxy (Lotz et al. 2004).

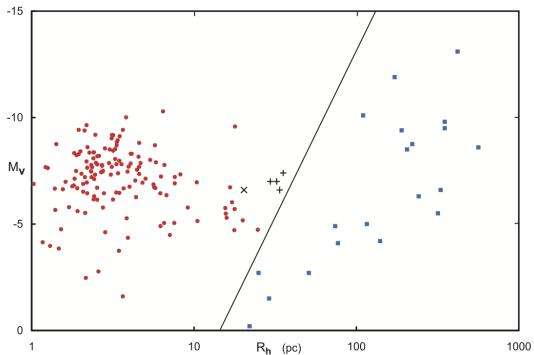


FIG. 59.— Absolute magnitudes versus half-light radii for globular clusters (red circles) and dSph's (blue squares) that orbit the MW. The solid black line shows the striking difference between the luminosity distributions of the separate components. Black plus signs denote the locations of four luminous and extended globular clusters that have been identified in the outskirts of M31. Image taken from van den Bergh (2008b).

impact on compact globular clusters than on extended dSph's. We note that the conclusion that typical dSph's are quite faint, and therefore difficult to observe, conforms to the framework of a hierarchical clustering scenario in which massive galaxies, such as the MW, should be surrounded by large numbers of satellite DM-dominated halos (van den Bergh 2008b).

Interestingly, many globular clusters appear to have been associated with past dwarf satellite galaxies. For example, an over-density of stars has been identified in the constellation Canis Major near the plane of the MW. A group of globular clusters and open clusters are associated with their overdensity in both position and radial velocity. This feature strongly suggests that a dwarf galaxy was integrated into the MW in the past and may now be a part of the thick disk. Another good example involves the Sagittarius dSph which is the closest galaxy to Earth. At just 24 kpc from Earth and 16 kpc from the centre of the MW, the Sagittarius dSph has suffered strong tidal interactions from the MW, resulting in an elongation pointed toward the centre of the MW (Carroll & Ostlie 2007, pg. 891). Over time, tidal forces will remove much of the stellar population of the Sagittarius dSph, leaving only behind its nucleus, the Messier object M54, which has mistakenly been classified as a luminous globular cluster (van den Bergh 2008a). Similarly, the unusual globular cluster  $\omega$  Centauri appears to be the stripped core of a dwarf galaxy that has been subsumed by the MW. It is the largest and brightest globular cluster visible from Earth and has an unusually high surface brightness, supported this hypothesis (Carroll & Ostlie 2007, pg. 891).

Since it is often difficult to confirm the presence of DM by measuring velocity dispersions of faint stars in distance objects, it is useful to have additional criteria to help discriminate between globular clusters and dSph's. Generally, the two can be distinguished using some of the following criteria: (1) mass, (2) luminosity, (3) size, (4) mass-to-light ratio, (5) spread in metallicity, and (6) ellipticity. As we have previously mentioned, globular clusters tend to be less massive, more luminous, smaller in size, have lower mass-to-light ratios, and exhibit a smaller range in metallicity. van den Bergh (2008b) have shown that dSph's are, on average, much more flattened than are globular clusters. The reason for this is not well understood though it may be due to the fact that they are embedded in triaxial DM mini halos.

Figure (59) shows a plot of the distribution of the absolute magnitudes  $M_v$  of Galactic globular clusters and of the presently known dwarf spheroidal companions of the MW as a function of their half-life radii  $R_h$ . Excluded from the set is the Sagittarius dwarf galaxy because both its present luminosity and half-life radius have probably been affected by tidal interactions with the MW. We see that dSph's are spread over a wider range of luminosities and have a fainter magnitude of  $M_v \sim -6.5$  than that of  $M_v \sim -7.5$  for the average globular cluster. This difference is likely to increase as more faint very faint and very low surface brightness dSph's are discovered around the MW. Furthermore, at least part of this difference is due to the fact that many low-mass globular clusters were destroyed by stellar-dynamical evaporation, which has a stronger

<sup>24</sup> <http://astronomy.swin.edu.au/cms/astro/cosmos/h/Half-light+Radius>

**QUESTION 8**

The stars in the solar neighbourhood, roughly the 300 pc around us, have a range of ages, metallicities and orbital properties. How are those properties related?

### QUESTION 8

The stars in the solar neighbourhood, roughly the 300 pc around us, have a range of ages, metallicities and orbital properties. How are those properties related?

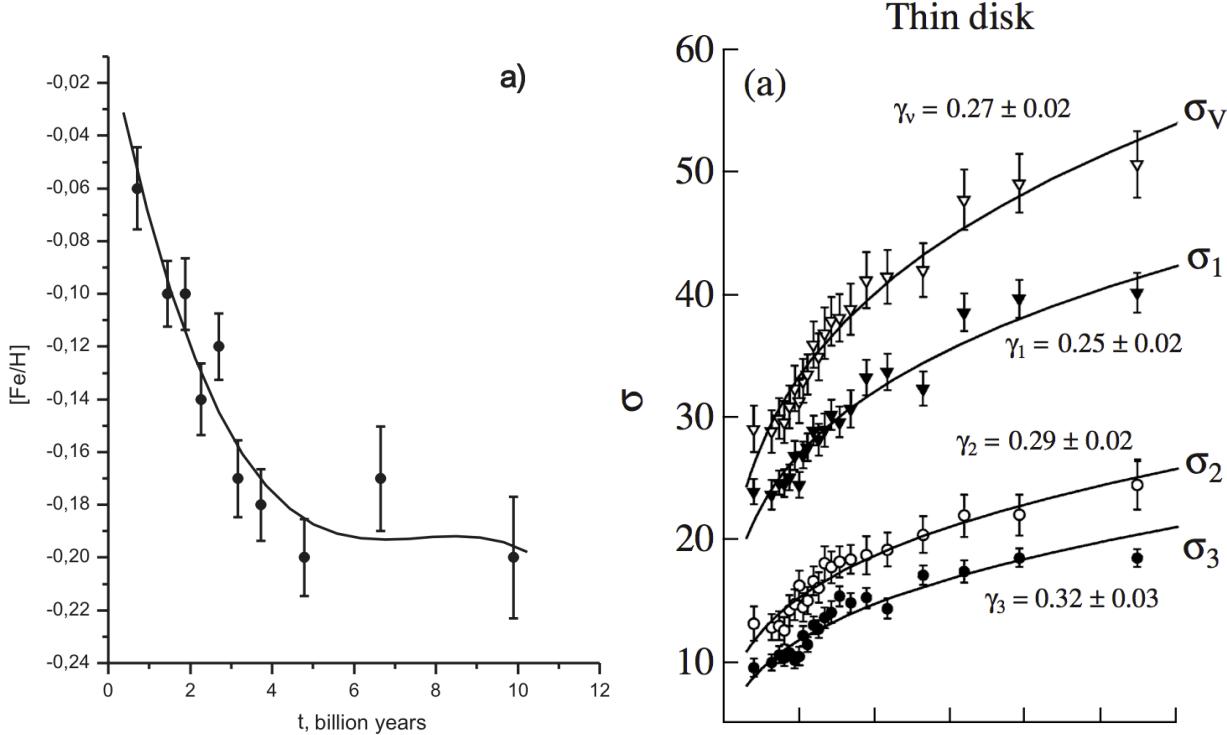


FIG. 60.— (left) Histogram of the mean metallicity versus stellar age as computed from thin-disk star residing with 70 pc of the Sun. Image taken from Marsakov et al. (2011) (right) Histogram of velocity dispersion versus age for thin-disk stars residing within 60 pc of the Sun. In this plot  $\sigma_U$ ,  $\sigma_V$ , and  $\sigma_W$  are denoted as  $\sigma_1$ ,  $\sigma_2$ , and  $\sigma_3$ , respectively, while total dispersion is  $\sigma_v$ . The bottom axis is stellar age from 0 to 12 Gyr, in 2 Gyr steps. Image taken from Koval' et al. (2009).

#### AGE-METALLICITY RELATION

The Big Bang created a universe consisting primarily of hydrogen, helium, and trace amounts of lithium. Hence, the first generation of Population III stars to form would have been depleted of heavier elements. Throughout the evolution of subsequent stellar generations a variety of methods would have enriched the ISM of metals. The primary driving force behind this are SNe though additional phenomena contribute including stellar winds or during phases in which stars eject part of their envelope, as in a PN. The latter requires that the matter within the star had previously been mixed by convection so that the metal-rich core material had been transported toward the surface (Schneider 2002, pg. 49).

Assuming that at the beginning of its evolution the MW had a chemical composition with only low metal content, the metallicity should be strongly related to the age of the stellar population. With each new generation of stars, more metals are produced and ejected into the ISM, partially by stellar winds, but mainly by SN explosions. Stars that are formed later should therefore have a higher metal content than those that were formed in the early phase of the Galaxy. One would therefore expect a relation to exist between the age of a star and its metallicity.

For instance, under this assumption  $[Fe/H]$  can be used as an age indicator for a stellar population, with the iron predominantly being produced and ejected in Type 1a SNe. Therefore, newly formed stars have a higher fraction of iron when they are born than their predecessors, and the youngest stars would have the highest iron abundance. Indeed, one finds  $[Fe/H] = -4.5$  for the oldest stars, whereas very young stars can have  $[Fe/H] = 1$  (Schneider 2002, pg. 50).

The apparent correlation between age and composition is referred to as the **age-metallicity relation (AMR)**. Its validity has been identified through studies of the MS turn-off points in both globular and galactic clusters which confirm that metal-rich stars tend to be younger than metal-poor stars of similar spectral type (Carroll & Ostlie 2007, pg. 885). Figure (60) shows the AMR for thin-disk stars within 70 pc of the Sun. We see that the metallicity first decreases appreciably with increasing age, but then remains constant within the uncertainties after 5 Gyr, with a limiting value of  $[Fe/H] \approx -0.2$ . This implies that within the first several Gyr of formation of the thin-disk system, the ISM was, on average, fairly rich in heavy elements. Marsakov et al. (2011) argue that the reason for this is a low SFR coupled to a continuous raining of metal-poor interstellar material from outer parts of the Galaxy onto the disk. However, approximately 5 Gyr ago, the mean metallicity began to rise due to a sudden increase in the SFR in the thin disk, an idea that is separately confirmed by the substantial increase in the number of stars beginning approximately from

this age. They further suggest that the burst in star formation 5 Gyr may have been caused by tidal interactions within the MW as it began its transformation into becoming one of the central galaxies in the Local Group. If the MW had an outer gaseous halo at that time, it probably would have fallen toward the disk, possibility triggering the burst in star formation.

We note that in many situations the correlation between age and [Fe/H] may not be as reliable as first believed. For example, significant numbers of Type Ia SNe do not appear until roughly 1 Gyr after star formation begins, and since Type 1a SNe are responsible for most of the iron production, iron is not available in large quantities to enrich the ISM until then. Furthermore, mixing of the ISM after a SN Ia event occurs only locally, so that large inhomogeneities in the [Fe/H] ratio may be present in the ISM, and thus even for stars of the same age (Carroll & Ostlie 2007, pg. 886). An alternative measure for metallicity is [O/H] because oxygen, which is an  $\alpha$ -element<sup>25</sup>, is mainly produced and ejected during core-collapse SN explosions of massive stars. These begin only  $\sim 0.01$  Gyr after the formation of a stellar population and therefore occur virtually instantaneously (Schneider 2002, pg. 50).

#### AGE-VELOCITY-DISPERSION RELATION

Observations show that the velocity dispersion of the molecular clouds in which star formation is taking place is much less, by about a factor of 3, than the dispersion of old stars near the Sun. We refer to this apparent correlation between stellar age and velocity dispersion as the **age-velocity-dispersion relation (AVR)** and it is thought to arise from some set of physical processes that are dynamically linked to the formation history of the disk (Carlberg et al. 1985). Such processes evidently transform slow-moving young stellar orbits into “hotter”, higher dispersion orbits, and are thus called “heating” mechanisms (Seabroke & Gilmore 2007).

Recall that velocity dispersion simply measures the standard deviation in the velocity distribution of a collection of stars (Carroll & Ostlie 2007, pg. 906). For stars in the solar neighbourhood we can measure a variety of one-dimensional velocity dispersions: denoted  $\sigma_U$ ,  $\sigma_V$ , and  $\sigma_W$  for radial motion towards the Galactic centre, azimuthal motion in the direction of Galactic rotation, and vertical motion perpendicular to the disk, respectively. The increase in these quantities with age is usually described by a power law of the form  $\sigma \propto t^\gamma$ , where  $t$  is stellar age (Koval' et al. 2009). Figure (60) shows the observed power-law behaviour in each velocity component for thin-disk stars within 60 pc of the Sun.

Of course, it is natural to wonder what physical processes are responsible for the observed heating of stellar orbits over time. A variety of in-plane heating mechanisms have been proposed, with the most dominant likely being transient spiral waves that generate gravitational potential fluctuations capable of exciting random motions in disk stars. Similar gravitational scattering by GMCs is also thought to play a minor in-plane heating role. However, neither of these mechanisms will strongly affect  $\sigma_W$  since both spiral waves and GMCs are generally confined to the thin disk. Since stars spend the majority of their time near the apocentre of their vertical orbits (i.e. furthest from the Galactic plane), there must be some sort of heating process operating away from the plane, in order to generate the  $\sigma_W$  behaviour seen in Figure (??). Seabroke & Gilmore (2007) argue that, in the context of the hierarchical galaxy formation paradigm, minor mergers with dwarf galaxies can act as a vertical heating mechanism without destroying the disk. The observed age- $\sigma_W$  may then be the combination of dwarf galaxy accretion for  $t \gtrsim 3$  Gyr and scattering by GMCs for  $t \lesssim 3$  Gyr. This may explain the apparent levelling off in the age- $\sigma_W$  displayed above (Seabroke & Gilmore 2007; Koval' et al. 2009).

#### VELOCITY-DISPERSION-METALLICITY RELATION

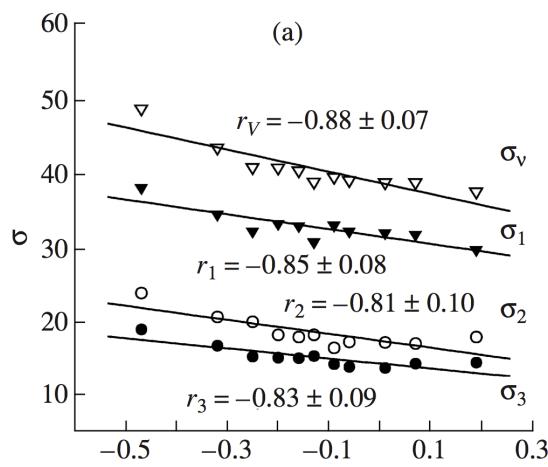


FIG. 61.— Histogram of velocity dispersion versus [Fe/H] for stars residing within 60 pc of the Sun. In this plot  $\sigma_U$ ,  $\sigma_V$ , and  $\sigma_W$  are denoted as  $\sigma_1$ ,  $\sigma_2$ , and  $\sigma_3$ , respectively, while total dispersion is  $\sigma_v$ . Image taken from Koval' et al. (2009).

<sup>25</sup>  $\alpha$ -elements are those which can be built up successively through the addition of an  $\alpha$ -particle ( ${}^4\text{He}$  nucleus). These elements dominate the abundance of ISM species and are a clear indication of nuclear fusion occurring in the He-rich zones of stars where hydrogen has been exhausted (Schneider 2002, pg. 49).

<sup>26</sup> The LSR, or local standard of rest, is defined to be a point that is instantaneously centred on the Sun and moving in a perfectly circular orbit along the solar circle about the Galactic centre (Carroll & Ostlie 2007, pg. 903).

The age-metallicity and age-velocity-dispersion relations can be combined as the **velocity-dispersion-metallicity relation (VMR)**. The VMR suggests that the oldest stars in the MW have the widest range of peculiar velocities, a trend that is evident in all three  $U$ ,  $V$ , and  $W$  coordinates; as displayed in Figure (61). Because stars with the smallest peculiar velocities do not drift away from the LSR<sup>26</sup> as quickly, they must occupy orbits that are similar to that of the LSR, implying that these young stars are members of the thin disk. On the other hand, the stars with the largest peculiar velocities follow very different paths about the centre of galaxy. In particular, stars with large  $\sigma_W$  must be passing through the solar neighbourhood on trajectories that will carry them to great distances above and below the disk. These old, metal-poor stars are therefore members of the thick disk or stellar halo (Carroll & Ostlie 2007, pg. 906).

Since much of the work introduced in this question has centred around stellar ages, it will be useful as an aside to summarize how this quantity is observationally determined for field stars. Typically, determining age requires knowledge of the photometrically measured effective temperature,  $T_{\text{eff}}$ , V-magnitude, and metallicity, [Fe/H]. With these values, a star can be placed on a three-dimensional HR diagram and its age can be estimated by applying Bayesian statistics to a set of theoretical isochrone models. In practice, isochrone ages can only be determined for stars that have evolved significantly off the ZAMS and are therefore not very

reliable for young stars (Seabroke & Gilmore 2007). Furthermore, the determination of ages for individual field stars is ripe with large systematic and random errors. Some of these arise from the calibrations of the observational data while others are due to the various approximations used in stellar model calculations (Holmberg et al. 2007).

**QUESTION 9**

**What are the main sources of heat in the interstellar medium?**

## QUESTION 9

### What are the main sources of heat in the interstellar medium?

There are several sources of energy for the gas making up the ISM. These include (in no particular order) photoionization and photodissociation induced by UV starlight, cosmic ray and X-ray excitation and ionization, photoelectric emission from dust and PAHs, collisional and frictional heating from shock waves and MHD waves, and transient events such as stellar winds and supernovae; we discuss each of these in detail below.

#### HEATING BY STARLIGHT

The photoionization of an atom A with ionization potential I by a photon with energy  $h\nu$  releases an electron with kinetic energy ( $h\nu - I$ ):



The released photoelectron adds thermal energy to the gas through collisions with other atoms. Of course, not all of the kinetic energy of the photoelectron is imparted to the gas; its collisions may induce transitions in other atoms, allowing for possible reradiation to escape the system. Despite this competing effect, there is an overall contribution to the thermal reservoir of cool gas clouds by the photoionization of atoms such as C, Si, and Fe. In H II regions photoionization of H and He are the dominant heat sources, consuming all photons with energies above 13.6 eV. H I regions, on the other hand, have photoionization contributions from atoms with lower ionization thresholds, such as carbon, for which photons with energies greater than 11.3 eV deposit roughly 2.1 eV of heat per C ionization (Dyson & Williams 1997, Section 3.2.2).

A similar heating source occurs through the photodissociation of H<sub>2</sub> via photons with 11.2 eV  $\leq h\nu \leq 13.6$  eV:



This process brings H<sub>2</sub> from some initial level X(v, J) of the ground electronic state to a level B(v', J') or C(v', J') of the first or second electronic excited states (v and J denote vibrational and rotational substructure within the given electronic state). From here the molecule will relax either by falling to a bound state X(v'', J'') of the electronic ground state (happens  $\sim 85\%$  of the time) or by falling to the vibrational continuum (other  $\sim 15\%$  of the time) (Draine 2011, pg. 346). For the latter, the molecule will become unstable and dissociate, imparting the kinetic energy of the resultant hydrogen nuclei into heating the gas. The average energy released per photodissociation is 0.4 eV, providing a significant heat source in regions where molecules are formed and destroyed rapidly (Dyson & Williams 1997, Section 3.2.2). For the former process, the molecule will further relax by cascading to the ground rovibrational state of the ground electronic state. For a high density gas, collisional deexcitation can be an important relaxation mechanism and will distribute the  $\sim 2$  eV of vibrational energy to the rest of the gas (Hollenbach & Tielens 1999).

#### HEATING BY COSMIC RAYS AND X-RAYS

Another important heating source for the ISM involves cosmic ray excitation and ionization. Most electrons and ions in interstellar space have speeds drawn from the local thermal distribution, but a small fraction of the particles have energies that are much larger than thermal – these “nonthermal” electrons and ions are referred to as cosmic rays (Draine 2011, pg. 134). Cosmic rays can heat gas through two processes: firstly, through interactions with bound electrons resulting in the ejection of an energetic secondary electron, and secondly, through the transfer of kinetic energy to free electrons by elastic scattering (Draine 2011, pg. 318). Cosmic ray ionization proceeds through the process



where p is a cosmic ray particle. Electrons released in this way will have a wide distribution of energy, with a mean of  $\sim 35$  eV. Some of this kinetic energy will go into secondary ionizations and excitations of bound states of H, H<sub>2</sub>, and He that will then deexcite radiatively and possibly escape the system. The remaining energy will eventually wind up as thermal energy, with the heating efficiency depending upon the fractional ionization of the gas. If the ionization is high, then the primary cosmic ray particle will lose most of its energy through long-range Coulomb scattering off free electrons, and  $\sim 100\%$  of its initial kinetic energy will be converted to heat. However, when the gas is neutral, a fraction of the cosmic ray energy is lost through secondary ionizations and excitations of bound states. In a purely neutral medium the fraction of energy going into heat is  $\sim 0.2$  the initial kinetic energy of the cosmic ray particle (Draine 2011, pg. 337).

X-rays emitted by compact objects or hot interstellar plasma can impinge on neutral regions. Photoelectrons produced by X-ray ionization of hydrogen will have substantial energy and, per primary ionization, the secondary ionizations and heating are greater than for cosmic ray ionization. The X-ray photoabsorption cross section for hydrogen is so small that X-rays can penetrate great hydrogen columns, allowing for heating of the interiors of thick clouds. X-ray ionization of helium can also be important since the photoabsorption cross section for this process is much larger than that of hydrogen. X-rays will be an important source of heating only in clouds that happen to be close to strong X-ray sources (Draine 2011, pg. 338).

On a side note, cosmic rays are somewhat of a misnomer since they actually consist of charged particles including electrons, positrons, and protons as well as more exotic species including muons and a host of atomic nuclei of metals. Particularly striking are the wide range of energies associated with cosmic rays, spanning less than  $10^7$  to  $10^{20}$  eV. The most abundant are those with low energies and are likely associated with solar winds, flares, and coronal mass ejections from the Sun. The sources of higher-energy cosmic rays ( $\lesssim 10^{16}$  eV) have been identified with supernovae. It has long been suggested that shock waves from supernovae could be sites of acceleration for cosmic ray particles. The charged particles are at first bound to the magnetic field lines of the SNR, but can be accelerated to very high energies through successive collisions with the advancing shock wave. After

absorbing energy from the shock, a particle is accelerated forward in the direction of the shock motion. However, being tied to the magnetic field in the shock's vicinity, the particle is forced to return, only to collide with the shock again, receiving additional energy. This process repeats many times until the particle possesses sufficient energy to escape the magnetic field. It remains unclear where the highest-energy and incredibly rare cosmic rays come from. It may be possible that they are extragalactic, being produced through collisions involving intergalactic shocks, or perhaps from AGN (Carroll & Ostlie 2007, pg. 550).

#### HEATING BY DUST

Another important heating mechanism is through the photoelectric emission from dust. When an energetic photon is absorbed on a dust grain, it may excite an electron to a sufficiently high energy that it can overcome the work function on the grain and escape from the surface. This process requires photons with  $h\nu \gtrsim 8$  eV and the effects of thermal heating are dominated by very small grains, including PAHs. Small grains dominate since they are the most abundant and because their photoelectric yields are higher. This process is the dominant heating mechanism in the diffuse neutral ISM (Draine 2011, pg. 339).

The photons responsible for ejecting electrons in this case come from the FUV part of the spectrum, below the hydrogen ionization threshold. This mechanism becomes inefficient if regions where FUV photons from the ISRF cannot penetrate, such as the cores of molecular clouds. In warmer regions or high ambient FUV fields, grains can become charged of PAHs ionized, which means the escaping photoelectrons must also break free of Coulomb interactions as well as the material work functions, suppressing the heating rate (Pogge 2011 lecture notes - see Galactic Question 16).

#### OTHER HEATING MECHANISMS

There are a variety of more exotic phenomena that play important roles in the heating of the ISM. One such form involves collisional and frictional heating through the damping of MHD waves. In clumpy mediums, pressure gradients will drive gas flows, which may excite MHD waves, in some cases even shock waves. Shock waves will directly heat the gas to large temperatures, and other MHD waves will in general be damped, with conversion of the wave energy into heat. (Draine 2011, pg. 318). Transient events including stellar winds and supernovae can also add substantial thermal energy to their local surroundings (Dyson & Williams 1997, Section 3.3.1).

#### COOLING MECHANISMS

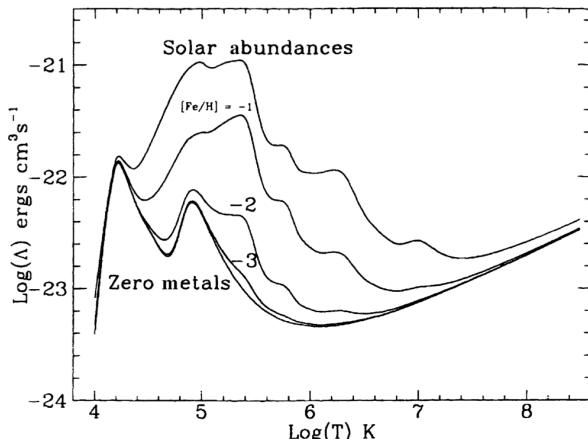
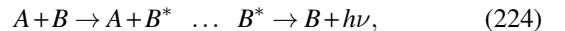


FIG. 62.— The cooling rate per unit volume of a typical astrophysical plasma for different cosmic abundances of the heavy elements, ranging from zero metals to the solar metallicity. In the zero metal case, the two maxima in the cooling curve are associated with the recombination of hydrogen ions and double ionized helium. At high temperatures thermal bremsstrahlung dominates and the cooling rate is proportional to  $T^{1/2}$ . Image taken from Longair (2008).

with kinetic energies of just a few eV. If the collisional excitation is followed by radiative deexcitation thermal energy will be removed from the cloud (Dyson & Williams 1997, Section 3.2.2).

Molecular hydrogen can be very abundant in interstellar gas, and may therefore also be an important coolant. However, electric dipole transitions are forbidden for  $H_2$  because the molecule has no dipole moment. As a result, it must radiatively deexcite through quadrupole transitions, which will only be effective at low densities. After  $H_2$  the next most abundant interstellar molecule is CO. CO is the most important coolant in dense clouds because it possesses a dipole moment implying that its rotational transitions are permitted. The CO molecule therefore relaxes very quickly, cooling the gas and being then ready for another excitation. When the column density of CO becomes too large, it may no longer remain optically thin, and so other less abundant molecules (i.e. OH and  $H_2O$ ) may contribute significantly to cooling (Dyson & Williams 1997, Section 3.2.3).

It is now useful to consider processes that remove thermal kinetic energy from gas. In general, interstellar clouds cool by emitting radiation; they do not usually conduct or convect heat very quickly. The mechanism by which this radiation occurs is usually initiated by an excitation of an atomic, ionic, or molecular transition during a collision. In this excitation the atom, ion, or molecule gains its energy from the kinetic energy of the colliding partner. After a time, the excited system radiates this energy away in a photon which may escape from the cloud. Thus, the gas loses kinetic energy, so it cools. We can summarize this process as



where the kinetic energy is removed from  $A$  in exciting  $B$  to  $B^*$ . The most efficient cooling processes are likely to be those in which the following criteria are satisfied: (1) frequent collisions, (2) excitation energy comparable or less than the thermal kinetic energy, (3) a high probability of excitation during the collision, (4) that the photon is normally emitted before a second collision occurs on the excited partner, (5) the photons are emitted and are not reabsorbed (Dyson & Williams 1997, Section 3.2.1).

Since hydrogen is the most abundant element, excitation of transitions in atomic hydrogen are likely to be an important cooling mechanism. However, the transitions are so energetic that only above  $\sim 10^4$  K does this mechanism play a role. At smaller temperatures other atoms and their ions – such as C, O, Fe, and Si – have energy levels that can be collisionally excited by electrons

TABLE 4  
MAIN COOLING MECHANISMS AT DIFFERENT TEMPERATURES.

Temperature [K]	Main Coolant
10	CO
$10^2$	$H_2, C\alpha$
$10^3$	Metastable ions
$10^4$	$H, H^+ + e$
$10^5$	$He, He^{2+} + e$
$> 10^6$	Thermal Bremsstrahlung

At  $T \gtrsim 10^4$  K, the medium will begin to become ionized and other cooling mechanisms will take over. Of primary importance are radiative deexcitation and radiative recombination of hydrogen and singly ionized helium. At the highest temperatures thermal bremsstrahlung will act as the primary agent for the removal of thermal energy from the gas (Longair 2008, pg. 479). This is the process through which an electron emits a photon as it passes near an ion, slowing the electron down (Carroll & Ostlie 2007, pg. 246). Figure (62) shows the cooling rate of gas with different metallicities as a function of temperature and Table 4 summarizes the main gas coolants within each temperature range.

**QUESTION 10**

**Draw an interstellar extinction curve (i.e. opacity), from the X-ray to the infrared. What are the physical processes responsible?**

## QUESTION 10

**Draw an interstellar extinction curve (i.e. opacity), from the X-ray to the infrared. What are the physical processes responsible?**

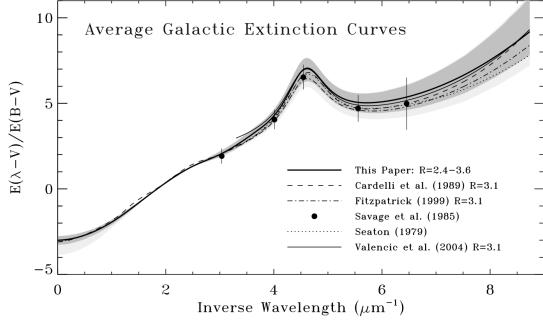


FIG. 63.— The thick solid line shows the average IR-through-UV extinction curve determined by analyzing extinction of roughly 300 O and B stars. The dark shaded region shows the sample variance about the mean while the other lines compare to similar studies. Image taken from Fitzpatrick & Massa (2007).

from the beam of starlight. If we define the extinction present along the line of sight, then from equation (227) and the relation that  $S_\lambda \propto 10^{-2/5m_\lambda}$ , we find that

$$A_\lambda \equiv m_\lambda - m_{\lambda,0} = 1.086\tau_\lambda, \quad (228)$$

where  $m_\lambda$  and  $m_{\lambda,0}$  are the attenuated and unattenuated apparent magnitudes of the source, respectively. Hence, the change in magnitude due to interstellar extinction is approximately equal to the optical depth along the line of sight (Carroll & Ostlie 2007, pg. 400).

We can visualize this process through the use of an interstellar extinction curve plotting the strength of extinction against  $\lambda$ . Usually the **relative colour excess**, defined as

$$K(\lambda-V) \equiv \frac{E(\lambda-V)}{E(B-V)}, \quad (229)$$

is shown, where  $E(X-Y) \equiv A_X - A_Y$ . An example of this is plotted from the IR to UV wavelengths in Figure 63. This extinction curve was generated by Fitzpatrick & Massa (2007) through the modelling of the observed SEDs of early-type O and B stars that have suffered interstellar extinction. Their method involves a  $\chi^2$ -minimization procedure to determine simultaneously the basic properties of a reddened star (i.e.  $T_{\text{eff}}$  and [Fe/H]) and the shape of its extinction curve, utilizing grids of stellar atmosphere models to represent intrinsic SEDs and an analytic form of the extinction curve, whose shape is determined by a set of adjustable parameters. The chief limitation of this method is that the intrinsic SEDs must be well represented by model atmosphere calculations. Figure 64 provides an extended view of the extinction curve from the IR to X-ray regime. This curve was constructed by Ryter (1996) and is built upon a variety observational studies coupled to analytic models.

A variety of interesting features are immediately seen in Figures 63 and 64. Each features arises from a specific physical process, which we can roughly summarize by stating that extinction from the IR to UV is due to the scattering of radiation of dust, whereas from the UV to X-rays, the extinction is almost exclusively dominated by photoionization of H, He, and metals (Ryter 1996).

The IR to visible portion of the extinction curve agrees well with the Mie theory prediction that  $A_\lambda \propto \lambda^{-1}$  (see equation (220)) due to scattering off of dust grains. For wavelengths shorter than the blue wavelength band (B), the curve begins to deviate significantly from this relation, and its use is completely untrusted at even shorter wavelengths (Carroll & Ostlie 2007, pg. 402). There are two prominent features within the IR to UV portion of the extinction curve: the 10 micron feature and the 2175 Å bump. The 10 micron feature is believed to be due to silicon absorption, specifically in relation to the stretching of the Si-O molecular bond. A similar feature at  $\lambda \approx 20 \mu\text{m}$  is thought to be related to the bending of Si-O-Si bonds in silicates. The existence of these features suggests that silicate grains are present in large abundance within dust clouds and the diffuse dust of the ISM (Carroll & Ostlie 2007, pg. 404). The origin of the 2175 Å bump is believed to be the result of molecular electronic transitions within aromatic PAHs, though this is still largely up for debate. PAHs are thought to occur ubiquitously in the ISM as revealed by their strong “UIR” emission bands spanning 3.3 to 12  $\mu\text{m}$  that appear to be due to vibrations in their C-C and C-H bonds. Xiang et al. (2011) perform an extensive literature search to compare the strength of the 2175 Å bump to the EWs of the UIR absorption features of extinction curves for various lines of sights, to see if a possible correlation exists between the two. Although no correlation is found, they propose a possible explanation may be that UIRs are produced by small free gas-phase PAH molecules and ions, whereas the 2175 Å bump is mainly from large PAHs or PAH clusters in condensed phase. Another candidate for the

The passage of starlight through intervening ISM material is subject to various scattering and absorption effects; a process collectively known as **interstellar extinction**. We know that for radiative transfer with pure absorption a beam of light will be attenuated according to

$$\frac{dI_\lambda}{ds} = -\kappa_\lambda \rho I_\lambda = -n\sigma_\lambda I_\lambda, \quad (225)$$

where  $I_\lambda$  is the specific intensity of the light,  $\kappa_\lambda$  is the absorption coefficient or opacity of the medium, and  $\sigma_\lambda$  is the absorption cross section, while  $\rho$  and  $n$  denote the density and number density of absorbing material, respectively. Equation (225) has the trivial solution:

$$I_\lambda = I_{\lambda,0} e^{-\tau_\lambda} \quad \text{where} \quad \tau_\lambda \equiv \int_0^s n\sigma_\lambda ds. \quad (226)$$

From this we can deduce an equivalent expression,

$$S_\lambda = S_{\lambda,0} e^{-\tau_\lambda}, \quad (227)$$

relating the measured attenuated flux,  $S_\lambda$ , to the intrinsic flux,  $S_{\lambda,0}$ , that has travelled through a medium with optical depth  $\tau_\lambda$ . Obviously, the flux weakens as scattering and absorption processes remove photons coefficient,  $A_\lambda$ , as the number of magnitudes of interstellar extinction

2175 Å bump is graphite, although it is unclear how carbon can organize into large graphite particles in the ISM (Carroll & Ostlie 2007, pg. 403).

The remainder of the extinction curve from the EUV to X-rays is due to the photoionization of hydrogen, helium, and metals. In particular, the sharp increase in the extinction curve at the Lyman limit of 91.2 nm corresponds to the ionization of hydrogen and portrays its ability to absorb photons with energies greater than this threshold. The subsequent spikes in the extinction curve correspond to K shell (i.e. at the absorption edge) ionization from progressively larger atomic species. Hence, the first three spikes correspond to the ionization of hydrogen, the first ionization of helium, and the second ionization of helium, respectively. After each spike the extinction curve falls off with a characteristic slope that arises from the quantum-mechanical result that the photoionization cross section decreases roughly as  $\lambda^{-3}$  above the absorption edge (Ryter 1996). Despite the fact that heavy elements have only  $\sim 10^{-3}$  the abundance of hydrogen, they dominate extinction at high energies where they provide a photoionization cross section  $\sim 10^4$  times larger than that of hydrogen (Draine 2011, pg. 129). For energies above the oxygen absorption edge at  $\sim 0.5$  keV the X-ray opacity is dominated by metals and H and He are relatively unimportant. Below 1 keV the dominant metal absorbers are C, N, O, and Ne, while above 1 keV the dominant absorbers are Si, S, and Fe (Wilms et al. 2000).

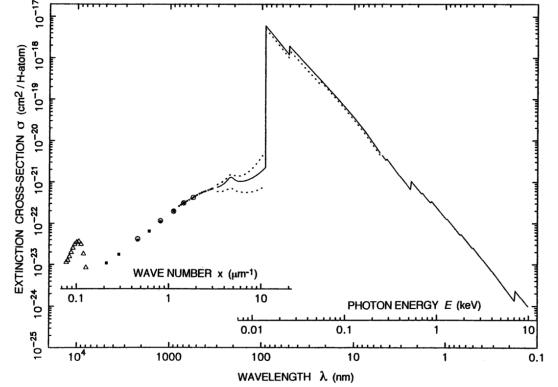


FIG. 64.— The extinction cross section normalized per H atom of the diffuse natural ISM from the FIR to X-rays. The thin solid line traces the curve for the rightmost part of the image while the individual data points show its structure for the leftmost part of the image. Image taken from Ryter (1996).

**QUESTION 11**

**What is dynamical friction? Explain how this operates in the merger of a small galaxy into a large one.**

## QUESTION 11

**What is dynamical friction? Explain how this operates in the merger of a small galaxy into a large one.**

A characteristic feature of collisions of stellar systems is the systematic transfer of energy from their relative orbital motion into random motions of their constituent particles. This process is simplest to understand in the case of **minor mergers** in which one system is much smaller than the other (Binney & Tremaine 1994, pg. 643). In this case we will consider the motion of a test body of mass  $M$  through an infinite and homogeneous collection of stars, gas clouds, and DM with constant mass density  $\rho$  and individual masses much less than  $M$  so that the test body moves in a straight line instead of being deflected. In the absence of collisions, it might be thought that  $M$  would move unimpeded. However, as  $M$  moves forward, the other objects are gravitationally pulled toward its path, with the closest ones feeling the largest force. This produces a region of enhanced density along the path, with a high-density wake trailing the test body. The result, known as **dynamical friction**, is a net gravitational force on  $M$  that opposes its motion. Kinetic energy is transferred from  $M$  to the surrounding material as its speed is reduced (Carroll & Ostlie 2007, pg. 1001).

Dimensional analysis can be used to show that the dynamical friction force take the form

$$f_d \simeq C \frac{G^2 M^2 \rho}{v_M^2}, \quad (230)$$

where  $C$  is a dimensionless constant whose specific value depends on how the speed,  $v_M$ , of the massive object compares with the velocity dispersion,  $\sigma$ , of the surrounding medium. Careful examination of equation (230) tells us why the terms enter as they do. Clearly, the dynamical friction must be proportional to the mass density of stars. Assuming that the relative numbers of objects of various masses do not change, doubling  $\rho$  means doubling the total number of objects, which would in turn double the gravitational force on  $M$ . The mass  $M$  itself is squared; one power comes from its role in producing the high-density wake and the other from the gravitational force on  $M$  produced by the enhanced density. Finally, we see that dynamical friction reduces with the square of  $v_M$ . If  $M$  moves twice as fast, it will spend only half as much time near a given object, and so the impulse given to that object is only half as great. As a result, the wake develops only half as rapidly, and  $M$  will be twice as far away by the time the enhancement arises. The presence of the  $v_M^2$  term in the denominator thus arises from the inverse-square law of gravity. This means that slow encounters are much more effective at decreasing the speed of an intruding mass (Carroll & Ostlie 2007, pg. 1003).

We can easily estimate the timescale associated with dynamical friction acting on the test body  $M$ . We will assume that the host galaxy has a density distribution of a singular isothermal sphere:

$$\rho(r) = \frac{v_M^2}{4\pi G r^2}, \quad (231)$$

where we assume the object moves at the circular speed of the rotation curve in the outer part of the host galaxy. If the test body's orbit is circular and of radius  $r$ , its orbital angular momentum is simply  $L = M v_M r$ . Since dynamical friction acts tangentially to the orbit and opposes the cluster's motion, a torque of magnitude  $\tau = r f_d$  will be exerted on  $M$ . The object's angular momentum therefore reduces according to  $dL/dt = \tau$ . If we assume that  $v_M$  remains constant then we can plug equation (231) into equation (230) and solve the differential equation in  $dL/dt$  to yield

$$t_{\text{fric}} = \frac{2\pi v_M r^2}{C G M} = \frac{\mathcal{M}(r)}{M} t_{\text{cross}}, \quad (232)$$

as the time required for  $M$  to spiral into the centre of the host galaxy. Here  $\mathcal{M}(r)$  is the total mass of the host galaxy contained within the initial radius  $r$  and  $t_{\text{cross}} = r/v_M$  is time required for  $M$  to cross the galaxy (pg. 1004 of Carroll & Ostlie 2007; Binney & Tremaine 1994, pg. 648).

### MINOR MERGERS

Most large galaxies are accompanied by several satellite galaxies, small companion galaxies that travel on bound orbits in the gravitational potential of the larger host. Satellites orbiting within the extended DM halo of their host galaxy experience dynamical friction, leading to orbital decay. As the satellite orbit decays, tidal forces from the host galaxy strip stars from the outer parts of the satellite, until eventually the entire satellite galaxy is disrupted. Through the modified form of equation (232) that includes the mass loss due to tidal stripping as the satellite spirals inward, we find that typical satellite galaxies merge from 30 kpc within  $\sim 10$  Gyr (Binney & Tremaine 1994, pg. 649). This is a fate that has already befallen the Sagittarius dSph, the remnant of the dwarf galaxy in Canis Major, and possibly the progenitor of  $\omega$  Centauri. It is also likely that dynamical friction will cause the LMC and SMC to merge with the MW some 14 Gyr in the future (Carroll & Ostlie 2007, pg. 1004).

Analysis of the dynamics of the Magellanic Stream<sup>27</sup> have found that the LMC and SMC orbit nearly perpendicular to the Galactic plane with the Magellanic Stream trailing behind. Figure 65 shows the predicted orbital decay of the LMC and SMC through dynamical friction as they pass through the halo of the MW. The ongoing mass loss that generates the Magellanic Stream provides circumstantial evidence that their orbits are continuing to shrink. (Binney & Tremaine 1994, pg. 651).

We can also apply dynamical friction to massive black holes and globular clusters. The centres of galaxies often contain SMBHs with masses from  $10^6 - 10^9 M_\odot$ . It is natural to ask whether such objects could also be present at other locations within

<sup>27</sup> The Magellanic Stream is a narrow band of natural hydrogen gas that extends over  $120^\circ$  in the sky and is believed to have been torn off the SMC by the gravitational field of the MW  $\sim 1$  Gyr ago

the galaxy, where they would be even harder to find. Using equation (232) we find that the typical inspiral time is only  $t_{\text{fric}} \sim 3$  Gyr. Black holes on eccentric orbits have even shorter inspiral times than those on circular orbits with the same mass, since the eccentric orbit passes through regions of higher density where the drag force is stronger. Hence, any SMBH that is formed within  $\sim 10$  kpc of the centre of a typical galaxy will spiral inwards within the age of the universe and we should therefore only expect to observe them in the centre, unless they formed far out in the galactic halo (Binney & Tremaine 1994, pg. 649). Similar calculations involving globular clusters show that any  $10^6 M_\odot$  clusters that form within roughly 4 kpc of the centre of a typical galaxy will have spiralled inwards by now (Carroll & Ostlie 2007, pg. 1004).

#### MAJOR MERGERS

Observational evidence appears to suggest that **major mergers** between galaxies of similar size play an important role in galactic evolution. Nearly all galaxies belong to clusters where the galaxies occupy a significant fraction of the total volume and the spacing between galaxies is only  $\sim 100$  times larger than their individual sizes. It therefore seems reasonable to suppose that strong galactic interactions occur frequently. Indeed, observations of densely populated rich clusters show that they have a higher proportion of elliptical galaxies in their centres than they do in their outer, less dense regions, with this trend being less prominent for more diffuse clusters. This strengthens the idea that interactions should increase the stellar velocity dispersions within galaxies, possibly transforming the disks of spiral galaxies into a more elliptical structure. It is also clear through X-ray observations that hot gas occupies much of the space between galaxies in rich clusters, with a total mass possibly exceeding the total stellar mass of the cluster. Evidently, gravitational influences remove gas from individual galaxies, trapping it within the potential well of the cluster. Although only a small fraction of gas is removed from each galaxy during a direct collision, mergers may initiate bursts of star formation that produce stellar mass loss and SNe, generating galactic winds capable of liberating large amounts of gas (Carroll & Ostlie 2007, pg. 999).

The collision rate for galaxies within a cluster is actually much larger than would be expected based on an analysis of their number density, luminous extent, and peculiar velocities. Firstly, the stars within a galaxy are embedded in a DM halo which can extend to  $\sim 100$  kpc, dramatically increasing the effective cross section for collision. Plus, once two DM halos start to merge, their high-density centres, which contain the stars and other baryonic matter, experience a drag force from dynamical friction that increases the likelihood of the two forming a merged product.

Second, the peculiar velocities of galaxies relative to the Hubble flow are caused by gravitational forces from nearby galaxies. Consequently, the peculiar velocities of nearby galaxies are correlated in that they are falling towards one another (i.e. MW and M31) so that the collision rate is much higher than if the peculiar velocities were randomly oriented. An involved analysis estimates that roughly 1 in every 10 MW-like galaxies suffer a collision during the age of the universe (Binney & Tremaine 1994, pg. 642).

Of course, not every encounter between galaxies leads to a merger. In particular, if galaxies approach each other with a relative speed greater than some critical value, then the two will have sufficient orbital energy to escape each other once their interaction is complete. This explains why most galaxies in rich clusters have not merged: although the density of galaxies in clusters is high, so collisions are frequent, the random velocities of most cluster members is large enough that their loss in orbital energy during a collision is negligible; the galaxies simply pass through one another (Binney & Tremaine 1994, pg. 640).

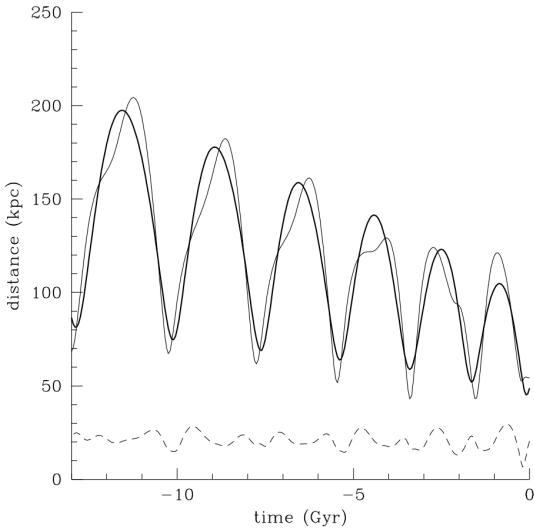


FIG. 65.— The predicted decay of the orbits of the LMC (thick line) and SMC (thin line) around the MW. The upper curves show the orbital distance from the Galactic centre while the lower, dashed curve shows the distance between the two clouds. Image taken from Binney & Tremaine (1994).

friction that increases the likelihood of the two forming a merged product. Second, the peculiar velocities of galaxies relative to the Hubble flow are caused by gravitational forces from nearby galaxies. Consequently, the peculiar velocities of nearby galaxies are correlated in that they are falling towards one another (i.e. MW and M31) so that the collision rate is much higher than if the peculiar velocities were randomly oriented. An involved analysis estimates that roughly 1 in every 10 MW-like galaxies suffer a collision during the age of the universe (Binney & Tremaine 1994, pg. 642).

Of course, not every encounter between galaxies leads to a merger. In particular, if galaxies approach each other with a relative speed greater than some critical value, then the two will have sufficient orbital energy to escape each other once their interaction is complete. This explains why most galaxies in rich clusters have not merged: although the density of galaxies in clusters is high, so collisions are frequent, the random velocities of most cluster members is large enough that their loss in orbital energy during a collision is negligible; the galaxies simply pass through one another (Binney & Tremaine 1994, pg. 640).

**QUESTION 12**

**Sketch the SED, from the radio to Gamma, of a spiral galaxy like the Milky Way. Describe the source and radiative mechanism of each feature.**

## QUESTION 12

**Sketch the SED, from the radio to Gamma, of a spiral galaxy like the Milky Way. Describe the source and radiative mechanism of each feature.**

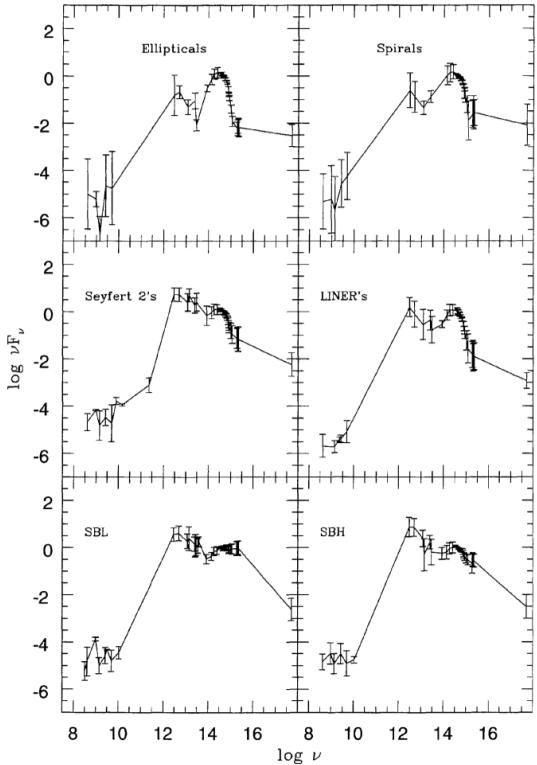


FIG. 66.— Plots of average SEDs generated by analyzing a collection of normal ellipticals (top left), normal spirals (top right), Seyfert 2 galaxies (middle left), LINERs (middle right), low reddening Starbursts (bottom left), and high reddening Starbursts (bottom right). Data points from different photometric surveys are shown, with straight line approximations between each survey. Image taken from Schmitt et al. (1997).

However, as hydrogen absorption is limited by the recombination rate, dust absorption becomes relatively more important as the strength of the ionizing radiation field increases, notably in compact H II regions and AGN. Dust attenuation within the optical-UV regime is characterized by its reddening, produced by the fact that shorter wavelength photons are more readily scattered and absorbed by dust. Dust emission is not important within this regime since the maximum temperature of a dust grain is constrained by sublimation to  $\sim 2000$  K. Conversely, as dust opacity strongly decreases with increasing wavelength, only dust emission is important in the FIR. Dust emission in the FIR and sub-mm is commonly modelled by a single blackbody of the dust temperature. In the MIR range simple blackbodies are not sufficient due to the stochastic heating processes that become important for small dust grains as well as strong PAH features present here. The former arises from the variation in temperature achieved by small grains because the impingement of photons onto the dust grain surface become less frequent and more random as the size of the dust grain decreases. The latter arises from the complex bending and stretching modes of PAHs (Walcher et al. 2011).

The radio portion of the SED of a typical spiral galaxy is dominated by synchrotron emission produced by relativistic electrons tracing interstellar magnetic field lines and by thermal bremsstrahlung emission of free electrons scattering off of ionized hydrogen atoms in H II regions. The synchrotron emissivity is spatially variable with enhancements near SNRs due in part to electron acceleration associated with the SN blast wave and in part to increased magnetic field strengths in the shocked gas (Draine 2011, pg. 120). As we have already mentioned, the FIR is dominated by thermal blackbody emission of interstellar dust warmed by absorbed starlight while the MIR is dominated by PAH emission. Since dust does not strongly obscure emission in the NIR, this region is dominated by blackbody emission of relatively cool, giant K stars. The optical and UV are dominated by stellar blackbody emission from massive stars, with scattering and absorption due to interstellar gas and dust described above. X-ray emission is dominated by thermal emission from hot, shocked plasma associated with SN explosion that inject larger amounts of kinetic energy into surrounding gas. Most of this energy is radiated as X-rays and EUV light, where the lower energy (EUV and soft X-rays) photons can be absorbed by small amounts of neutral gas (Draine 2011, pg. 125). Finally, gamma ray emission results from a combination of collisions between cosmic rays and atomic nuclei in interstellar clouds and high-energy events like

The **spectral energy distribution (SED)** of a source shows its wavelength dependency on brightness or flux, usually plotted in terms of  $\nu F_\nu$  versus  $\nu$ . Since different physical processes occurring within a source will leave their imprint on the global and detailed shape of the spectrum, detailed analysis of the SED allows us to study the properties of a source. For normal galaxies (excluding AGN) the majority of their light originates from the stars they contain, predominantly in the IR to UV range where the radiation is either direct starlight or reprocessed by the gas and dust in the surrounding ISM. If the distribution of the number density of stars is known as a function of their mass, chemical composition, and evolutionary stage, we can compute the light emitted by them through theoretical stellar atmosphere models. The method of creating galactic spectra through the superposition of stellar spectra is known as **stellar population synthesis**. The SED will naturally change in time since the stellar distribution changes in time as stars are continuously evolving, dying, and being reborn. The SED of a galaxy therefore reflects its history of star formation and stellar evolution (Walcher et al. 2011; Schneider 2002, pg. 132).

Figure 66 shows the SED for a regular spiral galaxy from radio ( $\nu \sim 10^8$  Hz) to soft X-rays ( $\nu \sim 10^{18}$  Hz). As we have briefly mentioned, the details of the IR-to-UV portion of the spectrum are dominated by blackbody emission from the stellar population reprocessed through scattering and absorption due to gas and dust within the ISM. Theoretical SEDs are generally constructed within this range by taking the stellar blackbody energy absorbed in the optical to UV regime and redistributing it across the MIR and FIR, assuming simple absorption and emission models for the gas and dust. Atomic gas is the predominant opacity source in the EUV ( $h\nu \gtrsim 13.6$  eV) where it reprocesses this light into strong emission lines in the UV, optical, and IR. It is thus especially important for young, actively star-forming galaxies. Molecular gas is not considered a significant contributor to the overall SED since it generally has a small volume filling factor, though some molecular emission lines can be observed in the radio-NIR originating in PDRs heated by the diffuse interstellar radiation field of the galaxy. Molecular gas is only a noticeable opacity source in galaxies dominated by nuclear or heavily obscured sources, such as AGN or ULIRGs. Since both the number density and absorption cross-section of dust is low relative to hydrogen in the EUV, dust is often ignored as an opacity source there.

synchrotron emission from pulsars (i.e as observed with the Crab, Geminga, and Vela pulsars in the MW)<sup>28</sup>.

In conclusion, we can split the spiral SED in Figure 66 into several components: synchrotron and bremsstrahlung radio ( $\nu \lesssim 10^{11}$  Hz) emission; sub-mm ( $10^{11}$  Hz  $\lesssim \nu \lesssim 10^{12}$  Hz) and FIR ( $10^{12}$  Hz  $\lesssim \nu \lesssim 10^{13}$  Hz) thermal dust emission; MIR ( $10^{13}$  Hz  $\lesssim \nu \lesssim 10^{14}$  Hz) PAH emission; NIR ( $10^{14}$  Hz  $\lesssim \nu \lesssim 4 \times 10^{14}$  Hz) blackbody emission from cool, red giants; optical ( $10^{14}$  Hz  $\lesssim \nu \lesssim 10^{15}$  Hz) and UV ( $10^{15}$  Hz  $\lesssim \nu \lesssim 10^{16}$  Hz) stellar blackbody radiation scattered and absorbed by gas and dust; thermal plasma X-ray ( $10^{16}$  Hz  $\lesssim \nu \lesssim 10^{18}$  Hz) emission; cosmic ray and synchrotron gamma ray ( $\nu \gtrsim 10^{18}$  Hz) emission. See also Draine (2011, Fig. 12.1) for a similar decomposition, minus the CMB contribution and plus the X-ray and gamma ray contributions.

#### OTHER GALAXIES

Figure 66 allows us to compare the average SEDs between different galaxy types. Schmitt et al. (1997) constructed each plot by averaging over a collection of galaxies obtained through a literature search; error bars on the data points are the standard deviation about the average. The ellipticals have small spread from the NIR-to-UV representing an old, red stellar population. However, they exhibit large variance in the FIR and radio regimes, the former attributed to different amounts of dust and the latter to the existence of a radio loud nucleus. When compared to ellipticals, the spirals show a large spread in the NIR-to-UV due to the presence of H II regions in the disk (Schmitt et al. 1997).

When comparing ellipticals to spirals we see that the two groups are very similar from the radio to the optical band. The most significant difference is in the UV portion where the spirals have an increasing contribution from H II regions. When comparing Seyfert 2's and LINERs, we see that Seyfert 2 galaxies have more UV and X-ray emission, consistent with a larger contribution from the active nucleus. The Seyfert 2's are also brighter in the IR where most of this emission is probably due to reradiation of the nuclear emission by a circumnuclear torus, which is possibly not present in LINERs. Finally, comparison between SBs and SBLs show that they are very similar over the entire spectrum, except in the UV where SBLs are brighter due to the lower reddening, and in the MIR/FIR where the SBs are brighter from the reradiation of the UV light (Schmitt et al. 1997).

Figure 67 plots a comparison between the visual SEDs of galaxies with differing Hubble type. It is easy to recognize the general trends in these spectra: the later the Hubble type, (1) the bluer the overall spectral distribution, (2) the stronger the emission lines, (3) the weaker the absorption lines, and (4) the smaller the 4000 Å break. Elliptical and S0 galaxies essentially have no star formation activity, which renders their SED dominated by red stars. Furthermore, in these galaxies there are no H II regions where emission lines could be generated. The old stellar population produces a pronounced 4000 Å break, which corresponds to a jump by a factor of  $\sim 2$  in the spectra of early-type galaxies. By contrast, Sc spirals and irregulars have a spectrum which is dominated by emission lines, where the Balmer lines of hydrogen as well as nitrogen and oxygen lines are most pronounced. The relative strength of these emission lines is characteristic for H II regions, implying that most of this line emission is produced in the ionized regions surrounding young stars. For irregular galaxies, the spectrum is nearly totally dominated by the stellar continuum light of hot stars and the emission lines from H II regions, whereas clear contributions from cooler stars can be identified in Sc spectra. The spectra of Sa and Sb galaxies from a kind of transition between those of early-type galaxies and Sc galaxies. Their spectra can be described as the superposition of an old stellar population generating a red continuum and a young population with its blue continuum and its emission lines. This can be seen in connection with the decreasing contribution of the bulge to the galaxy luminosity towards later type spirals (Schneider 2002, pg. 138).

#### TIME EVOLUTION

In the beginning, the spectrum and luminosity of a stellar population are dominated by the most massive stars, which emit intense UV radiation. But after  $\sim 10$  Myr, the flux below 1000 Å is diminished significantly and after 100 Myr, it hardly exists anymore. At the same time, the flux in the NIR increases because the massive stars evolve into red supergiants. The emission in the NIR remains high while the short-wavelength radiation is continuously diminished until the the population reaches an age of  $\sim 1$  Gyr. After this point red giant stars account for most of the NIR production. Then, after  $\sim 3$  Gyr, the UV radiation increases again from blue stars on the horizontal branch into which stars evolve after the AGB phase and also from white dwarfs which are hot when they are born. Between an age of 4 and 13 Gyr, the spectrum of a stellar population evolves fairly little (Schneider 2002, pg. 133). THE REMAINDER OF THIS SUBSECTION SHOULD BE ADOPTED INTO EXTRAGALACTIC Q12.

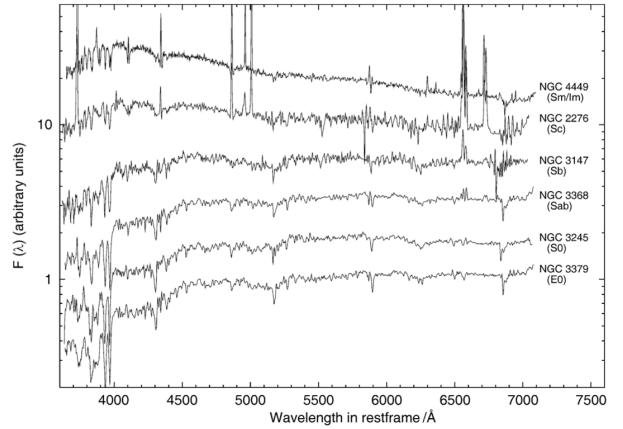


FIG. 67.— Spectra of galaxies of different types, where the spectral flux is plotted logarithmically in arbitrary units. The spectra are ordered according to the Hubble sequence, with early types at the bottom and late-type spectra at the top. Image taken from Schneider (2002).

<sup>28</sup> Unless otherwise cited, the information in this paragraph was obtained from the NASA website [http://mwmw.gsfc.nasa.gov/mmw\\_sci.html](http://mwmw.gsfc.nasa.gov/mmw_sci.html) that describes emission observed in different wavelengths for the MW.

**QUESTION 13**

**How many stars does one expect to find within 100 pc of the Sun? If all stars are distributed evenly across the galaxy, how many of these will be B spectral type or earlier? How many of these are younger than 100 Myrs?**

### QUESTION 13

**How many stars does one expect to find within 100 pc of the Sun? If all stars are distributed evenly across the galaxy, how many of these will be B spectral type or earlier? How many of these are younger than 100 Myrs?**

The MW contains some  $N_{\text{total}} \sim 10^{11}$  stars that are roughly distributed in a disk of radius  $R_{\text{disk}} \sim 10$  kpc and thickness  $L_{\text{disk}} \sim 1$  kpc (Binney & Tremaine 1994, pg. 3). The Sun resides within this disk at approximately 8 kpc from its centre. If we assume that the stars are uniformly distributed throughout this disk then we expect to find

$$N = N_{\text{total}} \frac{4r^3}{3R_{\text{disk}}^2 L_{\text{disk}}}, \quad (233)$$

stars within a distance  $r$  from the Sun (assuming  $r$  is small compared to  $R_{\text{disk}}$  and  $L_{\text{disk}}$ ). Using this result, we would expect to find  $N \sim 10^6$  stars within 100 pc of the Sun.

Computing the relative abundances of stellar types we expect to find within the solar neighbourhood requires knowledge of the PDMF, defined to be the stellar distribution in mass at the present time. We cannot expect the PDMF to have the same shape as the IMF since the MS lifetime  $\tau_{\text{MS}}$  of a star is a strong function of mass. In particular, for stars with MS lifetimes less than the current age of the MW,  $\tau_{\text{MW}}$ , we observe only those stars born within the last  $\tau_{\text{MS}}$  years because all stars born between  $t = 0$  and  $t = \tau_{\text{MW}} - \tau_{\text{MS}}$  have since evolved off the MS (Miller & Scalo 1979). Hence, it is instructive to determine the **turnoff mass**,  $m_{\text{crit}}$ , defined to be the maximum mass for which a star that formed along with the Galaxy would still exist on the MS. Stars with masses above  $m_{\text{crit}}$  will have evolved as red giants and white dwarfs or neutron stars and black holes, depending on their initial MS mass (Chabrier 2003).

For this task we must formulate a model for the MS lifetime as a function of stellar mass. Following Schneider (2002, pg. 429) we can estimate  $\tau_{\text{MS}}$  by dividing the star's total nuclear fusion energy production by its luminosity. Roughly speaking, the total energy produced throughout a star's MS lifetime is  $E_{\text{MS}} \approx (0.1)(0.007)mc^2$ , where we assume that fusion ceases after 10% of hydrogen is consumed and the factor of 0.007 arises from the efficiency of the fusion process. In addition, observations show that luminosity scales like  $L_{\text{MS}} \propto m^{3.5}$  for MS stars. With these two pieces of information we find that

$$\tau_{\text{MS}}(m) \approx 10 \left( \frac{m}{M_{\odot}} \right)^{-2.5} \text{ Gyr.} \quad (234)$$

If we suppose that  $\tau_{\text{MW}} \approx 10$  Gyr old then equation (234) yields a turnoff mass of  $m_{\text{crit}} = 1 M_{\odot}$ .

With an IMF  $\xi(m)$  and a Galactic SFR  $\psi$ , the number density  $dn$  of stars in the mass interval of width  $dm$  around  $m$  that we expect to observe today is

$$dn = \begin{cases} \psi \tau_{\text{MW}} \xi(m) dm & m \leq m_{\text{crit}} \\ \psi \tau_{\text{MS}}(m) \xi(m) dm & m > m_{\text{crit}}. \end{cases} \quad (235)$$

Note that a couple of assumptions went into the construction of equation (235). Specifically, we assumed that both the shape of  $\xi(m)$  and the value for  $\psi$  have remained unchanged in time. From a theoretical point of view it would be naive to expect a time-independent IMF because the process of star formation depends on a number of variables – specifically metal content; through it is unclear to what degree this affects the shape of  $\xi(m)$  (Miller & Scalo 1979). One argument in favour of a dynamic IMF is the so-called G-dwarf problem based on the lack of metal-depleted F and G MS stars that implies relatively few low mass stars were formed when metallicity was low at early times. This may suggest a top-heavy IMF at early times with a minimum low-mass cutoff  $\gtrsim 1 M_{\odot}$  (Chabrier 2003). Alternative models for the SFR are to assume that  $\psi$  exponentially decreases with time or that it first rises until it reaches a maximum value and then falls as the available gas and dust in the ISM is gradually consumed (Carroll & Ostlie 2007, pg. 1019).

All stars of B spectral type or earlier reside above the turnoff point since they have a minimum mass of  $m_B \approx 3 M_{\odot} > m_{\text{crit}}$  (Carroll & Ostlie 2007, Appendix G). From the use of equation (235) we see that the fraction of stars of B spectral type or earlier are

$$\chi_B = \frac{\int_{m_B}^{\infty} \tau_{\text{MS}}(m) \xi(m) dm}{\tau_{\text{MW}} \int_{m_L}^{m_{\text{crit}}} \xi(m) dm + \int_{m_{\text{crit}}}^{\infty} \tau_{\text{MS}}(m) \xi(m) dm}, \quad (236)$$

where we take a minimum stellar mass of  $m_L = 0.1 M_{\odot}$ . Note that  $\psi$  cancelled with itself in equation (236) under the assumption that it is constant in both mass and time. Using equation (234) and the Salpeter (1955) IMF  $\xi(m) \propto m^{-2.35}$ , equation (236) can be solved to obtain

$$\chi_B = \frac{\frac{1}{3.85} m_B^{-3.85}}{\frac{1}{1.35} (m_L^{-1.35} - m_{\text{crit}}^{-1.35}) + \frac{1}{3.85} m_{\text{crit}}^{-3.85}} \approx 2 \times 10^{-4}. \quad (237)$$

We therefore expect to find roughly 200 stars of B spectral type or earlier within 100 pc of the Sun.

Through the use of equation (234) we find that stars with lifetimes shorter than 100 Myr have masses  $\gtrsim 6 M_{\odot}$ . If we wish to determine the fraction of 200 stars that are younger than 100 Myr we proceed in a similar manner as before, modifying equation (236) to

$$\chi_{100} = \frac{\int_6^{\infty} \tau_{\text{MS}}(m) \xi(m) dm}{\tau_{100} \int_{m_B}^6 \xi(m) dm + \int_6^{\infty} \tau_{\text{MS}}(m) \xi(m) dm}, \quad (238)$$

where  $\tau_{100} = 100$  Myr. Plugging in the appropriate numbers yields

$$\chi_{100} = \frac{\frac{100}{3.85} 6^{-3.85}}{\frac{1}{1.35} (m_B^{-1.35} - 6^{-1.35}) + \frac{100}{3.85} 6^{-3.85}} \approx 0.2. \quad (239)$$

Hence, of the 200 stars with B spectral type or earlier, roughly 40 of them formed within the past 100 Myr.

**QUESTION 14**

**Describe what happens as a cloud starts to collapse and form a star. What is the difference between the collapse and contraction stages? What happens to the internal temperature in both? When does this phase end, and why does the end point depend on the mass of the object?**

### QUESTION 14

**Describe what happens as a cloud starts to collapse and form a star. What is the difference between the collapse and contraction stages? What happens to the internal temperature in both? When does this phase end, and why does the end point depend on the mass of the object?**

Before discussing this question in detail it is useful to briefly summarize the application of thermodynamics to fluid mechanics. In the study of heat transport, conservation of energy is expressed by the **first law of thermodynamics**,

$$dU = dQ - dW, \quad (240)$$

where the change in internal energy,  $dU$ , of a mass element is given by the amount of heat added to that element,  $dQ$ , minus the work done by that element on its surroundings,  $dW$ . The  $d$  here denotes an inexact differential which reflects the fact that both the amount of heat added to a system or the amount of work done by a system depends on the explicit ways in which the processes are carried out. Throughout our discussion we will assume that all energy changes are measured per unit mass. For an ideal gas it is useful to define the parameter  $\gamma$  to be the ratio of specific heats:

$$\gamma \equiv \frac{C_P}{C_V}, \quad (241)$$

where  $C_P$  and  $C_V$  are the amount of heat required to raise the temperature of a unit mass of material by a unit temperature interval at constant pressure and volume, respectively; ideal gases obey  $C_P = C_V + nk$ . If ionization occurs, some of the heat that would normally go into increasing the average kinetic energy of the particles must go into ionizing the atoms instead. Therefore, the temperature of the gas will not rise as rapidly, implying larger values for the specific heats. As both  $C_P$  and  $C_V$  increase,  $\gamma$  approaches unity. Since an ideal gas performs a work  $dW = PdV$  when expanding into its surroundings, we can rewrite equation (240) as

$$dU = dQ - PdV. \quad (242)$$

In the special case of an **adiabatic** process for which no heat flows into or out of the mass element, equation (242) is simplified further by setting  $dQ = 0$ . When coupled to the ideal gas law,  $PV = nkT$ , it is easy to find the adiabatic gas law:

$$P \propto V^{-\gamma} \propto \rho^{\gamma} \propto T^{\gamma/(\gamma-1)}. \quad (243)$$

On the other extreme is an **isothermal** process for which the temperature of the gas remains constant. In this case, the ideal gas law states that  $P \propto V^{-1}$ , so that the isothermal gas law has the form of equation (243) with  $\gamma = 1$  (Carroll & Ostlie 2007, pgs. 318-320).

For a star to be stable against a density perturbation (i.e. upon compression), pressure has to increase faster than gravitational self-attraction. With ideal gases we have that  $P \propto \rho^{\gamma} \propto R^{-3\gamma}$ , with  $\gamma = 5/3$  and  $4/3$  for the simplest non-relativistic and relativistic cases, respectively. For a star in hydrostatic equilibrium,  $P \propto R^{-4}$ , suggesting that stability thus requires  $\gamma \geq 4/3$ <sup>29</sup>.

#### COLLAPSE STAGE

In the case that the criterion for gravitational collapse has been satisfied in the absence of rotation, turbulence, and magnetic fields, the molecular cloud will collapse. If we make the simplifying assumption that any existing pressure gradients are too small to influence the motion appreciably, then the cloud is essentially in free-fall during the **collapse stage** of its evolution. Throughout this stage the collapse is isothermal as long as the cloud remains optically thin so that any heat generated from the release of gravitational potential energy can be efficiently radiated away. For this process the spherically symmetric hydrodynamic equation can be used to describe the collapse of the cloud if we assume that  $|dP/dr| \ll GM\rho/r^2$  (i.e. pressure support is negligible). In this case we find that the cloud collapses in the characteristic **free-fall timescale**

$$t_{\text{ff}} = \left( \frac{3\pi}{32} \frac{1}{G\rho_0} \right)^{1/2}. \quad (244)$$

Note that equation (244) is actually independent of the initial radius of the cloud. Consequently, as long as the original density of the molecular cloud was uniform, all parts will take the same amount of time to collapse, and the density will increase at the same rate everywhere. This behaviour is known as a **homologous collapse**. However, if the cloud is somewhat centrally condensed when the collapse begins,  $t_{\text{ff}}$  will be shorter for material near the centre meaning density increases fastest there; this is referred to as an **inside-out collapse** (Carroll & Ostlie 2007, pg. 417).

**Kramer's opacity law** states that the opacity of a molecular cloud goes like  $\kappa \propto \rho T^{-7/2}$  (Carroll & Ostlie 2007, pg. 250). Hence, as the cloud continues to collapse isothermally it will become increasingly opaque since  $\kappa \propto \rho$ . Eventually, the cloud will start absorbing a significant fraction of its own radiation and the isothermal assumption will break down; this marks the end of the collapse stage. The opacity of the cloud at this point is primarily due to the presence of dust (Carroll & Ostlie 2007, pg. 422).

#### CONTRACTION PHASE

<sup>29</sup> Much of the information presented throughout this question is obtained in Marten's lecture notes from his transients class.

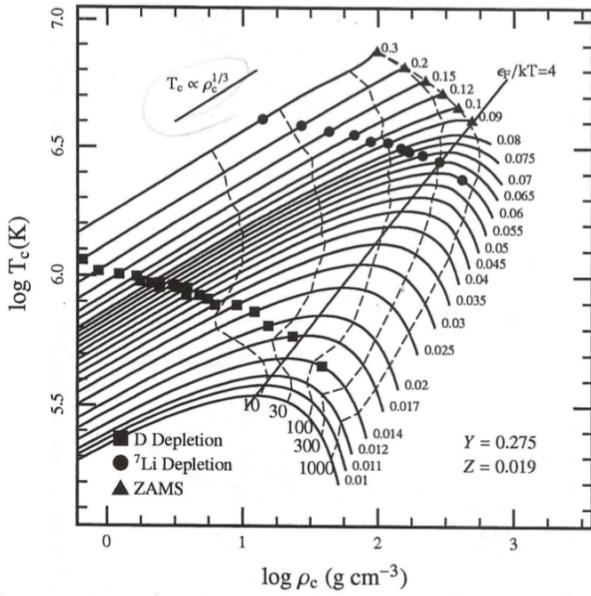


FIG. 68.— Evolution of the central temperate and density for low-mass ( $M < 0.3 M_{\odot}$ ) stars as they approach the MS or become brown dwarfs. These fully convective stars have internal structure that follows the expected  $T \propto \rho^{1/3}$  relation. Tracks turn over when the conditions become fairly degenerate as they pass through the  $T_D \propto \rho^{2/3}$  line. Note that massive radiation-dominated stars for which  $P_T \propto T^4$  do not approach degeneracy since in the relativistic regime  $P_D \propto \rho^{4/3}$  and therefore lines of constant degeneracy have  $T_D \propto \rho^{1/3}$ . Image taken from Marten's transient notes.

pressure which is independent of temperature and, in the non-relativistic regime, goes like  $P_D \propto \rho^{5/3}$ . If we equate these two pressure contributions then we find that the temperature at which the cloud becomes degenerate scales with density as  $T_D \propto \rho^{2/3}$ . Hence, depending on the density of the cloud, the temperature can either become large enough to ignite hydrogen fusion ( $T \sim 10^7$  K) in which case a star is born and enters the MS, or else it will become degenerate first and halt further contraction. It turns out that the critical mass at which this occurs is  $M_{\text{crit}} \approx 0.08 M_{\odot}$ , with all objects with masses  $M < M_{\text{crit}}$  becoming **brown dwarfs** that eventually cool through blackbody radiation. Figure 68 plots the numerical trajectories of stars as they approach the MS and shows the transition to brown dwarfs where the core becomes degenerate and fail to ignite hydrogen. Note that brown dwarfs may still be able to achieve deuterium burning through the second step of the PP chain, though this does not last long due to the limited supply of deuterium, and therefore does not slow the contraction significantly<sup>30</sup>. Lithium burning may also be achieved at higher temperatures.

#### HAYASHI TRACK

As we have just seen, once the collapse of a molecular cloud has begun, it is characterized by the free-fall timescale given by equation (244). With the formation of a protostar, the rate of evolution becomes controlled by the rate at which the star can thermally adjust to the collapse. This is just the Kelvin-Helmholtz timescale, and since  $t_{\text{KH}} \gg t_{\text{ff}}$ , protostellar evolution proceeds at a much slower rate than free-fall collapse (Carroll & Ostlie 2007, pg. 425).

With the steadily increasing temperature of the protostar, the opacity of the outer layers become dominated by the H<sup>-</sup> ion, the extra electrons coming from the partial ionization of metals that have lower ionization potentials. As with the envelope of the Sun, this large opacity contribution causes the envelope of a contracting protostar to become convective. It turns out that the constraints that convection places on the structure of a star produces an evolutionary path that traces a nearly vertical line on an HR diagram, known as the **Hayashi track**. The Hayashi track actually represents a boundary between “allowed” hydrostatic stellar models and those that are “forbidden”. To the right of the Hayashi track, there is no mechanism that can adequately transport the luminosity out of the star at those low effective temperatures; hence no stable stars can exist there. To the left of the Hayashi track, convection and/or radiation is responsible for the necessary energy transport<sup>31</sup> (Carroll & Ostlie 2007, pg. 425).

Figure 12.11 of Carroll & Ostlie (2007, pg. 426) shows the pre-MS evolution for stars of various masses started on the Hayashi track. We will consider the evolution of a  $1 M_{\odot}$  star that takes roughly 40 Myr to complete contraction. With the high H<sup>-</sup> opacity

<sup>30</sup> This reaction is favoured over the first step in the PP chain because it has a fairly large cross-section at low energies (Carroll & Ostlie 2007, pg. 423).

<sup>31</sup> This distinction between allowed and forbidden models is not in conflict with the free-fall evolution of collapsing gas clouds to the right of the Hayashi track since those objects are far from being in hydrostatic equilibrium.

Once the molecular clouds becomes optically thick the collapse will become adiabatic and we enter the **contraction phase**. In this case, the further contraction of the cloud progresses at a much slower pace due to increased thermal pressure. According to the virial theorem, energy must be liberated in order for the cloud to shrink in size. Also note that the virial theorem states that the total energy of a system of particles in equilibrium is one-half of the system's total potential energy. Therefore, only one-half of the change in gravitational potential energy of a star is actually available to be radiated away; the remaining potential energy supplies thermal energy (Carroll & Ostlie 2007, pg. 296). From this we can approximate the **Kelvin-Helmholtz timescale** over which the contraction phase proceeds:

$$t_{\text{KH}} \approx \frac{GM_J}{2R_J^2} \frac{1}{L}, \quad (245)$$

where  $L$  is the luminosity released as blackbody radiation during the contraction phase (Carroll & Ostlie 2007, pg. 418).

We can also use the virial theorem to determine how the temperature changes during the contraction phase. In particular, we know that  $MkT \approx GM^2/2R$  so that  $T \propto R^{-1} \propto \rho^{1/3}$  and is therefore a polytrope with  $\gamma = 4/3$ . Obviously, the cloud heats up as it contracts with this process maintaining hydrostatic equilibrium and continuing until a star is eventually born. Certain dynamical instabilities will occur during this contraction phase at times when  $\gamma$  momentarily falls below 4/3. Specifically, this happens when the hydrogen in the cloud is first dissociated and then ionized as the temperature rises. During these stages the heat gained from contraction is lost to dissociations/ionizations and the lack of thermal support causes temporary periods of isothermal collapse.

It is important to note that thermal pressure support,  $P_T \propto \rho T$ , is not the only factor slowing the contraction of the molecular cloud. Another potential contribution comes from electron degeneracy

which is independent of temperature and, in the non-relativistic regime, goes like  $P_D \propto \rho^{5/3}$ . If we equate these two pressure contributions then we find that the temperature at which the cloud becomes degenerate scales with density as  $T_D \propto \rho^{2/3}$ . Hence, depending on the density of the cloud, the temperature can either become large enough to ignite hydrogen fusion ( $T \sim 10^7$  K) in which case a star is born and enters the MS, or else it will become degenerate first and halt further contraction. It turns out that the critical mass at which this occurs is  $M_{\text{crit}} \approx 0.08 M_{\odot}$ , with all objects with masses  $M < M_{\text{crit}}$  becoming **brown dwarfs** that eventually cool through blackbody radiation. Figure 68 plots the numerical trajectories of stars as they approach the MS and shows the transition to brown dwarfs where the core becomes degenerate and fail to ignite hydrogen. Note that brown dwarfs may still be able to achieve deuterium burning through the second step of the PP chain, though this does not last long due to the limited supply of deuterium, and therefore does not slow the contraction significantly<sup>30</sup>. Lithium burning may also be achieved at higher temperatures.

#### HAYASHI TRACK

As we have just seen, once the collapse of a molecular cloud has begun, it is characterized by the free-fall timescale given by equation (244). With the formation of a protostar, the rate of evolution becomes controlled by the rate at which the star can thermally adjust to the collapse. This is just the Kelvin-Helmholtz timescale, and since  $t_{\text{KH}} \gg t_{\text{ff}}$ , protostellar evolution proceeds at a much slower rate than free-fall collapse (Carroll & Ostlie 2007, pg. 425).

With the steadily increasing temperature of the protostar, the opacity of the outer layers become dominated by the H<sup>-</sup> ion, the extra electrons coming from the partial ionization of metals that have lower ionization potentials. As with the envelope of the Sun, this large opacity contribution causes the envelope of a contracting protostar to become convective. It turns out that the constraints that convection places on the structure of a star produces an evolutionary path that traces a nearly vertical line on an HR diagram, known as the **Hayashi track**. The Hayashi track actually represents a boundary between “allowed” hydrostatic stellar models and those that are “forbidden”. To the right of the Hayashi track, there is no mechanism that can adequately transport the luminosity out of the star at those low effective temperatures; hence no stable stars can exist there. To the left of the Hayashi track, convection and/or radiation is responsible for the necessary energy transport<sup>31</sup> (Carroll & Ostlie 2007, pg. 425).

Figure 12.11 of Carroll & Ostlie (2007, pg. 426) shows the pre-MS evolution for stars of various masses started on the Hayashi track. We will consider the evolution of a  $1 M_{\odot}$  star that takes roughly 40 Myr to complete contraction. With the high H<sup>-</sup> opacity

<sup>30</sup> This reaction is favoured over the first step in the PP chain because it has a fairly large cross-section at low energies (Carroll & Ostlie 2007, pg. 423).

<sup>31</sup> This distinction between allowed and forbidden models is not in conflict with the free-fall evolution of collapsing gas clouds to the right of the Hayashi track since those objects are far from being in hydrostatic equilibrium.

near the surface, the star is completely convective during the first  $\sim 1$  Myr of the collapse. As the central temperature continues to rise, increasing levels of ionization decrease the opacity in that region and a radiative core develops, progressively encompassing more and more of the star's mass. At the point of minimum luminosity in the tracks following the descent along the Hayashi track, the existence of the radiative core allows energy to escape into the convective envelope more readily, causing the luminosity of the star to increase again. In addition, the effective temperature continues to increase since the star is still shrinking (Carroll & Ostlie 2007, pg. 427).

At about the same time that the luminosity begins to increase again, the temperature near the centre has become high enough for nuclear reactions to begin in earnest, although not yet at their equilibrium rates. Initially, the first two steps of the PP chain and the CNO reactions that turn  $^{12}\text{C}$  into  $^{14}\text{N}$  dominate the nuclear energy production. With time, these reactions provide an increasingly larger fraction of the luminosity, while the energy production due to gravitational collapse makes less of a contribution. Due to the onset of the highly temperature-dependent CNO reactions, a steep temperature gradient is established in the core, and some convection again develops in that regions. At the local maximum in the luminosity on the HR diagram near the shorted dashed line, the rate of nuclear energy production has become so great that the central core is forced to expand somewhat. This effect is apparent at the surface as the total luminosity decreases toward its MS value, accompanied by a decrease in the effective temperature. When the  $^{12}\text{C}$  is finally exhausted, the core completes its readjustment to nuclear burning, reaching a sufficiently high temperature for the remainder of the PP chain to become important. The star finally settles onto the ZAMS with the establishment of this stable energy source (Carroll & Ostlie 2007, pg. 427).

For lower-mass stars ( $M \lesssim 0.5 M_{\odot}$ ) the upward branch is missing just before the MS. This happens because the central temperature never gets hot enough to burn  $^{12}\text{C}$  efficiently. In addition, temperatures remain cool enough and the opacity stays sufficiently high in low-mass stars that a radiative core never develops. Consequently, these stars remain fully convective all the way to the MS. For massive stars, the central temperature quickly becomes hot enough to burn  $^{12}\text{C}$  as well as convert  $^1\text{H}$  into  $^3\text{He}$ . This means that these stars leave the Hayashi track at higher luminosities and evolve nearly horizontally across the HR diagram. Because of the much larger central temperatures, the full CNO cycle becomes the dominant mechanism for hydrogen burning in these MS stars. Since the CNO cycle is so strongly temperature dependent, the core remains convective even after the MS is reached (Carroll & Ostlie 2007, pg. 428).

**QUESTION 15**

Sketch the rotation curve for a typical spiral galaxy. Show that a flat rotation curve implies the existence of a dark matter halo with a density profile that drops off as  $1/r^2$ .

### QUESTION 15

**Sketch the rotation curve for a typical spiral galaxy. Show that a flat rotation curve implies the existence of a dark matter halo with a density profile that drops off as  $1/r^2$ .**

Observational determination of the rotation curves of spiral galaxies revolves around the Doppler shift, where the inclination of the disk must be accounted for. Usually the inclination of the disk is found from the observed axis ratio of the disk, assuming that disks are intrinsically axially symmetric. Mainly stars and H I gas in galaxies are used as luminous tracers, where the observable H I disk is in general significantly more extended than the stellar disk. Therefore, the rotation curves measured from the 21-cm line typically extend to much larger radii than those from optical stellar spectroscopy (Schneider 2002, pg. 100). Measuring the rotation curve of the MW is more complicated due to our interior vantage point. Due to geometry, 21-cm line emission can only be used within the Sun's orbital radius, so for greater distances we must rely on objects available in the Galactic plane, such as Cepheids for which we can directly measure distance (Carroll & Ostlie 2007, pg. 914). At the largest radii, luminous tracers like satellite galaxies are required to measure rotation curves. Since we cannot presume that satellite galaxies move on circular paths around their parent galaxy, conclusions can only be drawn based on a statistical sample of satellites (Schneider 2002, pg. 101).

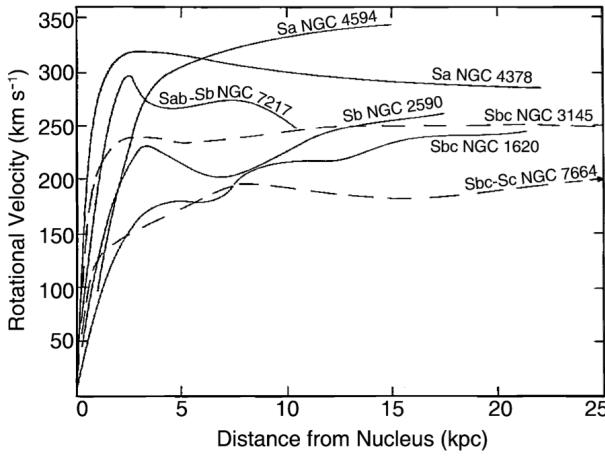


FIG. 69.— Rotation curves from a sample of spiral galaxies with their respective Hubble types labelled. Image taken from Schneider (2002).

It is remarkable that  $\rho \propto r^{-2}$  implies a mass profile  $M \propto r$  so that the mass of the halo increases linearly with radius for large  $r$ . Rotation curves produced from 21-cm observations and analysis of the relative velocities of satellite galaxies around spirals give no indication of an “edge” to the halo, leading to a lower limit for the radius of  $r_{\text{halo}} \gtrsim 100$  kpc (Schneider 2002, pg. 101).

The  $r^{-2}$  density dependence is much shallower than the number density of stars in the luminous stellar halo of the MW that are found to fall off like  $r^{-3.5}$  (Carroll & Ostlie 2007, pg. 918). It is this discrepancy that lead us to conclude that the majority of the mass within spirals is in the form of some non-luminous (dark) matter. The rotation curve for only baryonic matter can be determined by assuming a constant, plausible value for the MLR of the luminous matter. This value is obtained either from the spectral light distribution of the stars, together with knowledge of the properties of stellar populations, or by fitting the innermost part of the rotation curve (where the DM contribution can presumably be neglected), assuming that  $M/L$  is independent of radius for the stellar population. An example decomposition of a rotation curve into its bulge, disk, and DM halo components is displayed in Carroll & Ostlie (2007, Figure 24.28 - pg. 919). As also shown in Figure 69, at first rotation curves exhibit a rapid rise in rotation speed where  $v \propto r$ , out to a few kpc from the centre. This type of rotation is referred to as rigid-body rotation because when  $v \propto r$  all stars have the same orbital period regardless of distance, just as a rigid object would. This part of the rotation curve corresponds to the bulge component of spirals, where after peaking it proceeds to fall off as  $r^{-1/2}$ , as predicted from equation (246) for a matter distribution that truncates after some finite radius. The disk component displays a similar falloff above its scale length (see also Schneider 2002, Figure 3.16 - pg 101).

#### DEPENDENCE ON HUBBLE TYPE

As shown in Figure 69, the form and amplitude of the rotation curves of spirals are correlated with their luminosity and Hubble type. The larger the luminosity of a spiral, the steeper the rise of  $v(r)$  in the central region, and the larger the maximum rotation velocity  $v_{\text{max}}$ . This latter fact indicates that the mass of a galaxy increases with luminosity, as expected. For the characteristic values of the various Hubble types, we find  $v_{\text{max}} \sim 300$  km/s for Sa's,  $v_{\text{max}} \sim 175$  km/s for Sc's, and  $v_{\text{max}} \sim 70$  km/s for Irr's. For equal luminosity,  $v_{\text{max}}$  is higher for earlier types of spirals. However, the shape remains similar amongst Hubble types, despite the fact that they have different brightness profiles as seen, for example, from the varying bulge-to-disk ratio. This point is another indicator that the rotation curves cannot be explained by visible matter alone (Schneider 2002, pg. 102). Note that the small  $v_{\text{max}}$

Figure 69 shows rotation curves measured from a sample of spiral galaxies. It is clear that the rotation curves of spirals do not decrease at radii above the scale-length of their disks, as would be expected from the light distribution. Instead, the curves become roughly flat, leading us to conclude that spirals are surrounded by a halo of DM. Since rotation speed depends on the distribution of mass, a great deal can be learned about the matter in galaxies by studying these curves. For example, flat rotation curves suggest that the bulk of the mass in the outer regions of the galaxy is specially distributed with a density law that is proportional to  $r^{-2}$ . To see this, we balance the centripetal and gravitational forces acting on a star of mass  $m$  at radius  $r$  within a spherically symmetric galaxy:

$$\frac{mv^2}{r} = \frac{GM(r)m}{r^2} \rightarrow M(r) = \frac{v^2 r}{G} \rightarrow \frac{dM}{dr} = \frac{v^2}{G}, \quad (246)$$

where  $M(r)$  is the galactic mass contained within radius  $r$ . Conservation of mass within an infinitesimally thin shell states that  $dM/dr = 4\pi r^2 \rho$ . Hence, from equation (246) we see that the mass density must vary as

$$\rho(r) = \frac{v^2}{4\pi G r^2} \propto r^{-2}. \quad (247)$$

of irregular galaxies seems to suggest that a minimum rotation speed of roughly 100 km/s is required for the development of a well-organized spiral pattern. The slower rotation velocities of Irr's imply that their values of the rotational angular momentum per unit mass are only about 10% of the value found for the MW in the solar neighbourhood (Carroll & Ostlie 2007, pg. 952).

#### TULLY-FISHER RELATION

Through the use of 21-cm observations of spiral galaxies it has been found that the maximum rotation speed of spirals is closely related to their luminosity, following the so-called **Tully-Fisher relation**,

$$L \propto v_{\max}^{\alpha}, \quad (248)$$

where the slope is  $\alpha \sim 4$ . The larger the wavelength of the filter in which the luminosity is measured, the smaller the dispersion of this relation. This is to be expected because radiation at longer wavelengths is less affected by dust absorption and by the current SFR, which may vary to some extent between individual spirals. Observing in IR is particularly beneficial since this radiation comes primarily from late-type giant stars that are good tracers of the galaxy's overall luminous mass distribution (Carroll & Ostlie 2007, pg. 954). Because of the close correlation in equation (248) the luminosity of spirals can be estimated quite precisely by measuring the rotational velocity. Since  $v_{\max}$  is independent of the galaxy's distance the Tully-Fisher relation is used as an important distance measure. The measurement of  $v_{\max}$  is obtained either through a spatially resolved rotation curve which is possible for relatively nearby galaxies, or by observing an integrated spectrum of the 21-cm line (Schneider 2002, pg. 104). When the 21-cm line is sampled across the entire galaxy at one time, it typically displays a double peak, as shown in Carroll & Ostlie (2007, Figure 25.9 - pg. 953). The double peak arises because of the flat rotation curve of the galaxy, which generally has the highest rotational velocity in the flat part of the curve. Since so much H I participates in the rotation at this maximum velocity, the flux density is greatest at this value. The double peak occurs because a portion of the disk is rotating toward the observer, causing the line to be blueshifted, and a portion of the disk is rotating away from the observer, resulting in a redshifted line. The average radial velocity of the galaxy relative to the observer is the midpoint value between the two peaks.

It is easy to make a crude derivation of the Tully-Fisher relation based on the information we have seen so far. To begin, note that equation (246) implies that

$$M = \frac{v_{\max}^2 R}{G}, \quad (249)$$

where  $R$  is the distance from the centre of the galaxy to the flat part of the rotation curve; if  $v(R)$  is constant then its exact value is immaterial. We can rewrite equation (249) as

$$L = \left( \frac{M}{L} \right)^{-1} \frac{v_{\max}^2 R}{G}, \quad (250)$$

and by replacing  $R$  with the mean surface brightness  $\langle I \rangle = L/R^2$ , we obtain

$$L = \left( \frac{M}{L} \right)^{-2} \left( \frac{1}{G^2 \langle I \rangle} \right) v_{\max}^4. \quad (251)$$

This is the Tully-Fisher relation if  $M/L$  and  $\langle I \rangle$  are the same for all spirals. The latter is in fact suggested by **Freeman's Law** that states that the central surface brightness of disks has a very low spread amongst spirals. Since the shapes of the rotation curves of spirals are very similar, the radial dependence of the ratio of luminous to DM may also be quite similar among spirals. Furthermore, since the red of IR mass-to-light ratios of a stellar population do not depend strongly on age, the constancy of  $M/L$  could also be valid if DM is included (Schneider 2002, pg. 105).

The Tully-Fisher relation is slightly shifted between spirals of different Hubble type due to their intrinsic differences in  $v_{\max}$ . A good example of this is shown in Carroll & Ostlie (2007, Figure 25.10 - pg. 954).

#### ELLIPTICAL GALAXIES

For elliptical galaxies, the mass estimate and thus the detection of a possible DM component is significantly more complicated, since the orbits of stars are substantially more complex than in spirals. In particular, the mass estimate from measuring stellar velocity dispersion via line widths depends on the anisotropy of the stellar orbits, which is a priori unknown. Nevertheless, in recent years it has been unambiguously proven that DM also exists in ellipticals. First, the degeneracy between the anisotropy of the orbits and the mass determination was broken by detailed kinematic analysis. Second, in some ellipticals hot gas has been detected through X-ray emission; the temperature of this gas allows an estimate of the depth of the potential well, and therefore its mass. Both methods reveal that ellipticals are also surrounded by DM halos. This is further corroborated through weak gravitational lensing that offers another way to measure the masses of galaxies up to very large radii (Schneider 2002, pg. 101).

**QUESTION 16**

**What thermal phases are postulated to exist in the interstellar medium? Describe the dominant mechanism of cooling for each phase.**

### QUESTION 16

What thermal phases are postulated to exist in the interstellar medium? Describe the dominant mechanism of cooling for each phase.

The gas in the ISM exists in a number of thermal phases depending upon the local conditions of heating, ionization, chemical composition, etc. We discuss in detail the primary phases of the ISM below<sup>32</sup> and summarize the main points in Table 5.

#### MOLECULAR PHASE

The first phase we will discuss are **molecular clouds**. These are predominantly H<sub>2</sub> clouds with temperatures of roughly 10 K and densities of  $10^3 - 10^6 \text{ cm}^{-3}$ . Molecular clouds comprise about 30% of the mass of the ISM, but occupy only roughly 0.1% of its volume. This groups is further subdivided into being either “diffuse” or “dense” where the latter are gravitationally bound and are often dark with visual extinction greater than 3 mag through their central regions. In these dense clouds, the dust grains are often coated with mantles composed of water and other molecular ices. It is within these regions that star formation takes place. It should be noted that although they are called dense, the gas pressures within them would qualify as ultrahigh vacuum in a terrestrial laboratory. Molecular clouds are primarily heated by photoelectrons from dust and by heating and ionization from cosmic rays and starlight. They cool mostly through fine structure line emission of C I and molecular line emission of CO. The ground state of C I has a fine-structure transition with an excitation energy of 92 K that emits a FIR photon with a wavelength of 158  $\mu\text{m}$ . The excited state is populated by collisions with electrons, or if the ionized fraction is low, it can be excited by collisions with either H atoms or H<sub>2</sub> molecules. Rovibrational emission from CO is the dominant cooling source at the coldest temperatures in molecular clouds. Molecular clouds are primarily traced through 2.6-mm CO emission and FIR dust emission.

Also associated with molecular clouds are **photodissociation regions (PDRs)** which are the warm, partially ionized surfaces of molecular clouds. They get their name since UV radiation impinging on these regions – either from a discrete source like O and B stars or from the ISRF – cause H<sub>2</sub> to be dissociated into H I.

#### NEUTRAL PHASES

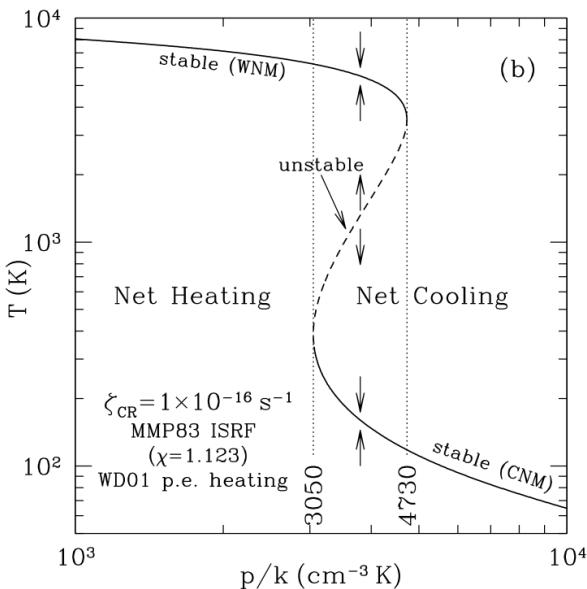


FIG. 70.— Steady state temperature as a function of thermal pressure for which the heating and cooling processes in neutral hydrogen gas balance. Image taken from Draine (2011).

340) provides a nice depiction of the cooling function for neutral atomic gas at these temperatures.

It is natural to wonder why we find atomic hydrogen to exist in only these two separate phases. This question can be answered by considering the temperature at which the heating processes (photoelectric dust emission and cosmic rays) and cooling mechanisms (fine-structure line emission) balance as a function of temperature. Figure 70 show the result of an analytical calculation using the best models for the different heating and cooling mechanisms. At low pressures, heating balances cooling at  $T \approx 6000$  K, similar to the WNM conditions. At high pressures, heating and cooling balance for  $T \approx 100$  K, which we recognize as the

The coolest phase in which neutral hydrogen exists is known as the **cold neutral medium (CNM)**. This consists of primarily atomic gas distributed in sheets and filaments occupying roughly 1% of the ISM by volume. The gas in this phase has characteristic temperatures of 100 K and densities of roughly  $10 \text{ cm}^{-3}$ . The other thermal phase in which neutral atomic hydrogen can exist is in the **warm neutral medium (WNM)**. The WNM fills roughly 40% of the volume of the ISM and is located mainly in PDRs on the boundaries of H II regions and on the outer surfaces of molecular clouds. This thermal phase has typical temperatures of 5000 K and hydrogen number densities of  $1 \text{ cm}^{-3}$ . Both the CNM and WNM are mainly traced through 21-cm line emission and absorption and optical and UV absorption along lines of sight towards bright stars or quasars. The WNM is often in pressure equilibrium with its surroundings whereas the CNM is only approximately in pressure equilibrium.

The heating and cooling mechanisms in the CNM and WNM are fairly similar. Both are predominantly heated by photoelectric emission from dust grains and interactions with cosmic rays. For both phases, the dominant cooling sources are the [C II] fine-structure line at 158  $\mu\text{m}$  that is effective from  $10 \lesssim T \lesssim 10^4$  K and the [O I] fine-structure line at 63  $\mu\text{m}$  that is important for temperatures above 100 K. The former is the dominant source for the CNM while the latter is most important for the WNM. The critical densities for [C II] and [O I] are roughly  $10^3 \text{ cm}^{-3}$  and  $10^5 \text{ cm}^{-3}$ , respectively, implying that collisional deexcitation of these levels is unimportant for both the CNM and WNM. Ly $\alpha$  only becomes an important cooling mechanism above  $10^4$  K and so does not have much of an effect here. Draine (2011, Figure 30.1, pg.

<sup>32</sup> All information in this question is obtained from Draine (2011) and the lecture notes of Prof. Richard Pogge found at <http://www.astronomy.ohio-state.edu/~pogge/Ast871/Notes/>

TABLE 5  
THERMAL PHASES IN THE ISM.

Phase	Temperature [K]	Density cm <sup>-3</sup>	$f_V$	Heating Mechanisms	Cooling Mechanisms
Molecular clouds	10	$10^3 - 10^6$	0.001	Photoelectric dust, cosmic rays, starlight	[C II], CO
CNM	100	10	0.01	Photoelectric dust, cosmic rays	[C II], [O I]
WNM	5000	1	0.4	Photoelectric dust, cosmic rays	[C II], [O I]
WIM	$10^4$	0.1	0.1	Photoionization of H and He	[O III], [N II], [S III], Ly $\alpha$ , H $\alpha$ , free-free
HIM	$10^6$	0.001	0.5	Shock-heated from SN and winds	Adiabatic expansion and X-ray emission

CNM conditions. However, there is an intermediate pressure range where, for a given pressure, there are three possible solutions. The upper and lower solutions are stable – if the gas temperature is perturbed away from equilibrium, it will return to it. However, the intermediate solution is thermally unstable – if  $T$  is perturbed upwards, the gas will warm up to the stable WNM solution, and if  $T$  is perturbed downward, it will cool to the stable CNM branch. Thus, we conclude that an ISM that is in thermal and dynamic equilibrium (uniform pressure) will have diffuse atomic gas in two distance phases, provided the pressure falls within the ranges considered here.

#### IONIZED PHASES

Ionized hydrogen begins to emerge in the **warm ionized medium (WIM)**. This consists of diffuse ionized gas occupying roughly 10% of the volume of the ISM and has temperatures and densities of  $10^4$  K and  $0.1 \text{ cm}^{-3}$  that can either be in pressure equilibrium or expanding. The WIM mainly encompasses the extended low-density photoionized regions in the ISM though ionized hydrogen can also be found in bright, high-density H II regions around O stars (exist for 10 Myr) and PN (exist for  $10^4$  yr). While primarily photoionized by UV starlight from O and B stars, there is evidence of shock or collisional ionization high above the plane of the MW. The principle source of heating in an ionized region is photoionization where the released photoelectrons gain some kinetic energy and quickly thermalize into a Maxwellian velocity distribution by electron-electron interactions. The dominant cooling mechanisms in these regions are radiative recombination, thermal bremsstrahlung continuum cooling, and line emission from collisionally excited ions. The first process contributes to cooling since every time an electron recombines with an ion, the plasma loses the kinetic energy of the recombining electron, which presumably is radiated away. Since the region is mostly ionized thermal electrons can scatter off ions in the plasma and emit free-free bremsstrahlung radiation, with a cooling rate that is proportional to  $T^{1/2}$  and  $n_e^2$ ; this is not a very efficient process in the WIM. Finally, the most dominant cooling mechanism is through collisionally-excited line emission from metals. Of course, most of the hydrogen in these regions is ionized and even if some He or He $^+$  are present, the energy of the first excited state is so far above the ground state that collisional excitation is negligible because the thermal electrons have insufficient energy. However, if heavy elements are present there may be ions like O I, O II, and N I with energy levels that can be collisionally excited by electrons with kinetic energies of just a few eV. If the collisional excitation is followed by a collisional deexcitation, the kinetic energy of the gas will be unchanged, but if it is followed by radiative deexcitation, the emitted photon will easily escape the region and contribute to cooling. The effectiveness of this process obviously depends on the density of the gas especially since the low-lying levels of many of the metal ions arise from the same electronic configuration as the grounds state and are therefore forbidden under the usual dipole selection rules. Nonetheless, electron-ion<sup>33</sup> impact excitation of metal ions followed by the radiation of line emission is the dominant cooling mechanism in the WIM with metallicities greater than a few percent of the solar value. The WIM is generally traced by the major hydrogen recombination lines (H $\alpha$  - 656.6 nm; H $\beta$  - 486.3 nm; H $\gamma$  - 434.2 nm) and free-free emission.

The final phase we consider is the **hot ionized medium (HIM)**. This consists of gas that has been shock-heated to temperatures above  $10^6$  K by SN explosions and stellar wind driven bubbles around Wolf-Rayet stars. The gas is collisionally ionized<sup>34</sup> and is often in pressure equilibrium with its surroundings implying low densities on the order of  $0.001 \text{ cm}^{-3}$ . Individual regions have characteristic scales of 20 pc and may be connected to other regions; collectively the HIM occupies approximately half of the volume of the galactic disk. The HIM is often buoyant, appearing as bubbles with a vertical scale height of 3 kpc that eventually cool and rain back down on the disk as galactic fountains. The HIM cools on Myr time scales mainly through adiabatic expansion (if it is expanding and hence not in pressure equilibrium) and soft X-ray emission from gas hotter than  $10^6$  K. The timescale for thermal bremsstrahlung continuum cooling is very long for such low-density gas and is therefore unimportant. The primary HIM tracers are absorption lines along lines of sight towards hot stars seen in nearby Local Group galaxies (i.e. LMC and SMC) or strong extragalactic UV and X-ray sources like quasars. They can also be identified through soft X-ray emission of the heated gas.

<sup>33</sup> Proton-ion and ion-ion impact excitation are inefficient because the Coulomb repulsion between the ions is so large.

<sup>34</sup> In contrast to photoionization, collisional ionization actually results in a net cooling of the plasma because energy equal to the ionization potential is removed from the electron gas.

**QUESTION 17**

Characterize the stellar populations in the following regions: i) the Galactic bulge ii) the Galactic disk, outside of star clusters iii) open star clusters iv) globular clusters v) a typical elliptical galaxy.

## QUESTION 17

**Characterize the stellar populations in the following regions: i) the Galactic bulge ii) the Galactic disk, outside of star clusters iii) open star clusters iv) globular clusters v) a typical elliptical galaxy.**

The structure of the MW is split up into a variety of components each characterized by the stellar populations they contain. Roughly speaking, stellar populations are distinguished by their age, metallicity, kinematics, and spatial distribution. For our following analysis there are a variety of concepts worth explaining, which we will do now.

The first concept we will discuss involves **metallicity**, which is a measure of the abundance of metals (all chemical elements heavier than helium) within a star. We define the metallicity  $Z$  to be the total mass fraction of all elements heavier than helium; the Sun has  $Z = 0.02$ , meaning that 98% of the Solar mass is locked up in hydrogen and helium (Schneider 2002, pg. 44). Another way to quantify metallicity is by comparing the ratios of iron to hydrogen in stars relative to the Sun, in which case we define the metallicity index to be

$$[\text{Fe}/\text{H}] \equiv \log \left[ \frac{(N_{\text{Fe}}/N_{\text{H}})_{\text{star}}}{(N_{\text{Fe}}/N_{\text{H}})_{\odot}} \right], \quad (252)$$

where  $N_{\text{Fe}}$  and  $N_{\text{H}}$  denote the number of iron and hydrogen atoms (Carroll & Ostlie 2007, pg. 852). Stars with  $[\text{Fe}/\text{H}] < 0$  are metal-poor relative to the Sun, and stars with  $[\text{Fe}/\text{H}] > 0$  are relatively metal-rich. The highest values for metal-rich stars in the MW is about 0.6 whereas the lowest values for metal-poor stars are about -5.4.

Based on metallicity, it is useful to distinguish between Population I (Pop I) stars which have solar-like metallicities ( $Z \sim 0.02$ ) and Population II (Pop II) stars that are metal poor ( $Z \sim 0.001$ ). The former are mainly located in the thin disk while the latter are predominantly found in the thick disk, galactic halo, and galactic bulge (Schneider 2002, pg. 47). The two populations also differ in age (Pop I stars are younger), velocity dispersion perpendicular to the disk (larger for Pop II stars), and scale-height (larger for Pop II due to larger velocity dispersions). The **scale-height** is defined to be the vertical distance from the galactic plane at which the density has decreased to 1/e its central value.

### GALACTIC BULGE

The galactic bulge is marked by the central thickening of the MW. The mass of the bulge is roughly  $10^{10} M_{\odot}$  and from COBE observations we know that it has the shape of a bar. The characteristic scale-length and scale-height of the bar are roughly 1 and 0.4 kpc, respectively (Schneider 2002, pg. 55).

The chemical abundance of stars in the bulge vary significantly, from quite metal-poor to very metal-rich with metallicity indices from  $-2 < [\text{Fe}/\text{H}] < 0.5$ , with a mean of about 0.3 (i.e. twice that of the Sun) (Carroll & Ostlie 2007, pg. 893). Such large metallicities hint at a rather young stellar population, in contrast to the colour of the bulge, which hints toward an old stellar population. However, it is important to note that the distinction in membership between disk and bulge is complicated by the fact that the scale-height of the (young) thin disk is comparable to that of the bulge. Nevertheless, current observations suggest the existence of a young stellar population with ongoing star formation, though the majority of the bulge is probably dominated by an old stellar population (Wyse & Gilmore 2005).

In a counterintuitive trend the oldest stars in the bulge tend to have the highest metallicities while the youngest stars show a fairly uniform distribution of metallicity indices spanning -2 to 0.5. This is likely the result of a burst of massive star formation when the MW was young. Core-collapse supernovae would have enriched the interstellar medium early in the life of the bulge, implying that subsequent generations of stars contained an enhanced abundance of heavier elements (Carroll & Ostlie 2007, pg. 893). Indeed, elemental abundances from stars in the bulge tend to show enhanced  $[\alpha/\text{Fe}]$  signatures of enrichment by predominantly Type II supernovae (Wyse & Gilmore 2005). The more uniform distribution of metallicity in recent generations of stars may be the result of fresh, in falling material.

### GALACTIC DISK (OUTSIDE OF STAR CLUSTERS)

When discussing the galactic disk it is useful to split it up into two components: the thin disk and the thick disk. The chemical composition of stars in the two disks differ in that we observe higher metallicities in the thin disk (Schneider 2002, pg. 47). In contrast, the metallicity of the stars within the galactic halo and bulge is smaller. As such, the thin disk is marked by a collection of Pop I stars whereas the thick disk contains predominantly Pop II stars.

The thin disk can be further subdivided into the young thin disk and the old thin disk (Schneider 2002, pg. 46). The young thin disk contains the largest fraction of gas and dust in the MW and is the site of active star formation. This population is concentrated very close to the plane of the galaxy with a scale-height of 100 pc. Representative objects include neutral atomic hydrogen, molecular gas, H II regions, protostars, O and B stars, supergiants, and classical cepheids<sup>35</sup>. The old thin disk is characterized by a slightly older stellar population and a correspondingly larger scale-height of roughly 325 pc. Representative objects include stars of type A or later, planetary nebulae, and white dwarfs.

In contrast to the thin disk, the thick disk contains an older stellar population with a larger scale-height of 1.5 kpc. Representative objects include Mira and RR Lyrae variables. The total mass contained in the thick disk is  $\sim 0.3 \times 10^9 M_{\odot}$  and is roughly 20 times less massive than the thin disk (Schneider 2002, pg. 46).

Characteristic metallicity values are  $-0.5 \lesssim [\text{Fe}/\text{H}] \lesssim 0.3$  and  $-1 \lesssim [\text{Fe}/\text{H}] \lesssim -0.4$  for the thin and thick disks respectively (Schneider 2002, pg. 50). From this we can deduce that stars in the thin disk are significantly younger on average than those in the thick disk. Either star formation started earlier, or ceased earlier, in the thick disk than in the thin disk, or stars that originally

<sup>35</sup> Info on representative objects taken from a graduate course description at the University of Hawaii: [http://www.ifa.hawaii.edu/~barnes/ast626\\_95/pcmw.html](http://www.ifa.hawaii.edu/~barnes/ast626_95/pcmw.html)

belonged to the thin disk have migrated into the thick disk. The latter is more appealing for a variety of reasons<sup>36</sup>. First of all, it is hard to understand why molecular gas, out of which stars form, was much more broadly distributed in earlier times than it is today, where we find it well concentrated near the Galactic plane. Moreover, the widening of an initially narrow stellar population in time is expected. Stars within the disk experience inhomogeneous gravitational interactions caused by neighbouring stars, spiral arms, and massive molecular clouds, producing random perturbations to their orbits. This results in an increase to their velocity dispersion perpendicular to the disk and therefore an increased scale-height (for details see Schneider 2002, pg. 47). In contrast to stars, the gas maintains a narrow distribution around the Galactic plane due to internal friction.

#### *OPEN STAR CLUSTERS*

During the collapse of a molecular cloud, stellar clusters can form, ranging in size from tens of stars to hundreds of thousands of stars. Since every member of the cluster formed from the same cloud, they formed with essentially identical compositions and within a relatively short period of time. Small stellar clusters that form in this way are referred to as open or galactic clusters (in contrast to globular clusters) (Carroll & Ostlie 2007, pg. 474).

Open clusters usually contain from a few tens to a few hundreds of stars. The kinetic energy of the cluster members, differential rotation of the MW, and external gravitational interactions tend to gradually disperse open clusters (Karttunen et al. 2006, pg. 339). For this reason the vast majority of open clusters persist as clusters for no more than a few 100 Myr. Janes & Phelps (1994) find that young open clusters ( $\sim 200$  Myr old) are distributed within the galactic plane with a scale-height of 55 pc. An older population of open clusters ( $\sim 1$  Gyr old) consisting of rich clusters is found in the outer regions of the Galactic disk (roughly beyond Sun's location) with a scale-height of 375 pc. The reason for the lack of old open clusters in the inner disk is attributed to the increased gravitational interactions with giant molecular clouds that are predominately located within the inner disk. Since open star clusters tend to be quite young, they are characterized by containing Pop I stars.

#### *GLOBULAR CLUSTERS*

Globular clusters differ from open clusters in that they contain roughly  $10^5$  members and are marked by older metal-poor Pop II stars. The distribution of stars within a globular cluster is spherically symmetric and the central densities are about 10 times larger than in open clusters. Only faint red stars remain on the main sequence within globular clusters. The main sequence is lower on a colour-magnitude diagram than that of open clusters because of the reduced metallicity (Karttunen et al. 2006, pg. 343).

Globular clusters are located within the Galactic halo with two distinct spatial groupings delineated by metallicity. Older ( $\sim 13$  Gyr), metal-poor clusters whose members have  $[Fe/H] < -0.8$  belong to an extended spherical halo of stars, while younger clusters ( $\sim 11$  Gyr) with  $[Fe/H] > -0.8$  form a much flatter distribution and may even be associated with the thick disk (Carroll & Ostlie 2007, pg. 894). The younger population tend to follow the general rotation of the MW whereas the older population have random orbits (Karttunen et al. 2006, pg. 344).

#### *TYPICAL ELLIPTICAL GALAXY*

Elliptical galaxies are similar to globular clusters in the sense that they contain an old stellar population. A metallicity gradient exists within typical elliptical galaxies in which the metallicity increases toward the galactic centre, as derived from colour gradients (Schneider 2002, pg. 92). Talk about motion of stars... Binney and Merrifield...

#### *RELATION TO GALACTIC FORMATION*

We will now attempt to explain the properties observed in the stellar populations of the various components of the MW based on a model of Galactic formation. This will follow Carroll & Ostlie (2007, pg. 1024-1027).

The MW is assumed to have formed through hierarchical merging of density fluctuations involving  $10^6$  to  $10^8 M_{\odot}$  proto-galactic fragments. Merging fragments would have formed a growing spheroidal mass distribution with many individual fragments evolving independently to form stars and possibly globular cluster cores. In the inner regions of the spheroid where the density was greatest, the rate of collapse and evolution would be fastest, yielding an old stellar population with enhanced chemical enrichment (i.e. the old, metal-rich central bulge). In the rarefied outer regions of the galaxy, chemical evolution and star formation would have been much slower.

Collisions and tidal interactions between the merging fragments would have disrupted the majority of fragments and left exposed the globular cluster cores of others. The disrupted systems would have led to the present distribution of field halo stars, while leaving the remaining globular clusters scattered throughout the spheroid. Those fragments that were initially moving in a retrograde direction relative to the eventual orbital motion of the galactic disk and inner halo produced the net zero rotation of the outer halo that is observed today. Certainly, the rate of collisions would have been greater near the galactic centre, disrupting the proto-galactic fragments first and building the bulge more rapidly than the halo.

The uniformity of globular cluster masses observed today ( $10^5 - 10^6 M_{\odot}$ ) is well explained by this model. Low-mass globular clusters would have had small gravitational binding energies, allowing them to become easily disrupted when the MW was young. High-mass clusters, on the other hand, are more severely affected by dynamical friction and would have rapidly spiralled into the inner regions of the MW while simultaneously dispersing through frequent collisions.

As the gas clouds of disrupted fragments collided, the collapse became largely dissipative, meaning that the gas began to settle slowly toward the central regions of the MW. The presence of some initial angular momentum caused the collapsing gas to

<sup>36</sup> There is still ongoing debate into whether the thin and thick disks are actually distinct. Some models propose that the stars in the thick disk formed outside the MW and only became constituents later, through accretion of satellite galaxies. This model is supported, among other reasons, by the fact that the rotational velocity of the thick disk around the centre of the MW is smaller than that of the thin disk by roughly 50 km/s. Even stronger discrepancies have been observed in other spirals; in one case the thick disk was found to rotate in the opposite direction to the thin disk (Schneider 2002, pg. 50).

become rotationally supported and settle into a disk. Of course, the already-formed halo stars did not participate in this collapse since their collisional cross-sections were too small to allow them to interact significantly, except through gravitational forces.

One model of thick-disk formation suggests that the thick disk may have formed around the galactic midplane with a temperature around  $10^6$  K. In regions where the gas was locally more dense it cooled rapidly, first through thermal bremsstrahlung and Compton scattering, and later at  $10^4$  K through hydrogen recombination. Once H I clouds could form they began collapsing to produce stars. Within a few Myr the most massive stars underwent core-collapse supernova, reheating the gas and increasing its metallicity. About 400 Myr after the first stars formed in this thick disk, the supernovae heating nearly ceased subsequent star formation. Only a few percent of the mass of the gas was converted into stars during this thick-disk-producing period of the MW's evolution<sup>37</sup>.

After the formation of the thick disk, cool molecular gas continued to settle onto the midplane with a scale-height of  $\sim 600$  pc. During the next several Gyr, star formation occurred within the thin disk. The scale-height of the thin disk was maintained by a self-regulating process. If the disk became thinner, its mass density would increase, thereby increasing the SFR, producing more SNe which then reheated and expanded the disk. The ensuing expansion would then decrease the SFR suppressing SNe explosions, so the disk would cool and shrink. Eventually, despite this self-regulating process, as the gas was depleted in the ISM the SFR decreased by an order of magnitude. Because of the decrease in the SFR, the thickness of the disk shrank to 350 pc, its current value. Finally, as the remaining gas continued to cool, it settled into an inner, metal-rich and gas-rich component of the thin disk (young thin disk) in which ongoing star formation is taking place. The Sun is located within the young thin disk.

In this scenario the presence of young stars in the galactic bulge is explained by recent mergers with gas-rich satellite galaxies. When those galaxies were disrupted by tidal interactions with the MW, their gas settled into the disk and bulge, forming new stars. It also appears that the MW's central bar aids the migration of dust and gas into the inner portion of the galaxy by generating dynamical instabilities as it rotates.

<sup>37</sup> A modified version of the thick-disk formation model suggests that the falling gas was initially much cooler allowing it to settle with a much smaller scale-height, similar to today's thin disk. Star formation was then able to proceed due to greater local density of gas and dust. Then, through some strong merger event 10 Gyr ago, the disk was reheated and puffed up to the present thick-disk scale-height of 1 kpc.

**QUESTION 18**

**How can you determine the temperature of H II regions?**

## QUESTION 18

### How can you determine the temperature of H II regions?

The populations of excited states of atoms and ions depend on the local density and temperature. Therefore, if we can determine the level populations from observations, we can use atoms and ions as probes of interstellar space. In particular, there are two general methods that can be used separately to probe the temperature and density of photoionized gas; we discuss these below. To be a useful probe, an atom or ion must be sufficiently abundant to observe, must have energy levels that are at suitable energies, and must have radiative transitions that allow us to probe these levels, either through emission or absorption lines (Draine 2011, pg. 203).

#### TEMPERATURE DIAGNOSTICS

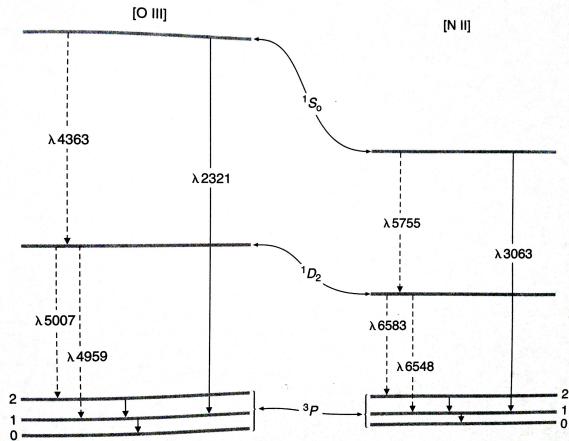


FIG. 71.— Energy-level diagrams for lowest terms of [O III] and [N II], all from ground  $2p^3$  configuration. Splitting of the ground  $^3P$  term has been exaggerated for clarity. Emission lines in the optical regime are indicated by dashed lines while solid lines show IR and UV transitions. Only the strongest transitions are shown. Image taken from Osterbrock & Ferland (2006).

ton either in  $\lambda 5007$  or  $\lambda 4959$ , with relative probabilities given by the ratio of the two transition probabilities, which is very close to 3 to 1. Every excitation of  $^1S$  is followed by emission of a photon in either  $\lambda 4363$  or  $\lambda 2321$ , with the relative probabilities again given by the transition probabilities. Each emission of a  $\lambda 4363$  photon further results in the population of  $^1D$ , and follows the aforementioned radiative cascade; but this contribution is small in comparison with the direct excitation of  $^1D$  and can be neglected (Osterbrock & Ferland 2006, pg. 108). As long as we remain in the low-density limit ( $n_e \ll n_{\text{crit}}$  for the  $^1D$  level) the line ratio  $(j_{\lambda 4959} + j_{\lambda 5007})/j_{\lambda 4363}$  is independent of the density, and only depends on the temperature (the ratio decreases with increasing temperature as  $^1S$  becomes more populated) (Draine 2011, pg. 205). However, at higher densities collisional deexcitation begins to play a role. The lower  $^1D$  term has a considerably longer radiative lifetime than the  $^1S$  term, so it is collisionally deexcited at lower electron densities than  $^1S$ , thus weakening  $\lambda 4959$  and  $\lambda 5007$ . In addition, under these conditions collisional excitation of  $^1S$  from the excited  $^1D$  level begins to strengthen  $\lambda 4363$ . The line ratio therefore decreases as the electron density is increased (Osterbrock & Ferland 2006, pg. 109). Fortunately, the values of  $n_{\text{crit}}$  for [O III] and [N II] ( $\sim 10^5 \text{ cm}^{-3}$ ) are high enough that approximation schemes are justified in many ionized nebulae (e.g., the Orion Nebula, with  $n_e \approx 3000 \text{ cm}^{-3}$ ) (Draine 2011, pg. 205).

Note that this technique requires no information on the distance to the nebula, the amount of  $O^{++}$  present, and so on, as all these factors cancel out (Osterbrock & Ferland 2006, pg. 111). On the other hand, a fundamental assumption is that the levels producing the lines are populated only by collisional excitation. The  $^1D$  level is at sufficiently high energy that as the temperature is lowered below  $\sim 5000\text{K}$ , the rate of collisional excitations becomes very small. This means that the line becomes very weak and difficult to observe; it also means that if the next ionization state (O IV, N III) has an appreciable abundance, radiative recombination with electrons may make a significant contribution to population of the  $^1D$  level. As a result, this method may not be suitable at low temperature (Draine 2011, pg. 207).

Another temperature diagnostic comes from the strengths of the discontinuities in the recombination continuum relative to strengths of recombination lines. The most commonly used discontinuity is the **Balmer jump**<sup>38</sup> at  $\lambda = 364.5 \text{ nm}$ :  $BJ \equiv I_\lambda(\lambda_{\text{BJ,blue}}) -$

<sup>38</sup> The Balmer jump is also used to measure the temperature of stellar atmospheres. In brief, since any photon with  $\lambda < 364.5 \text{ nm}$  is capable of ionizing

The first temperature diagnostic uses ions with two excited levels that are both energetically accessible at the temperatures of interest, but with an energy difference between them that is comparable to  $kT$ , so that the populations of these levels are sensitive to the gas temperature. A few ions, of which [O III] and [N II] are the best examples, have energy-level structures that result in emission lines from two different upper levels with considerably different excitation energies occurring in the observable wavelength region. The energy-level diagrams of these two ions is shown in Figure 71, where it can be seen that, for instance, [O III]  $\lambda 4363$  occurs from the upper  $^1S$  level, while  $\lambda 4959$  and  $\lambda 5007$  occur from the intermediate  $^1D$  level ( $^3P_0 - ^1D_2$   $\lambda 4931$ , which can only occur by an electric-quadrupole transition, has much smaller transition probability and is so weak that it can be ignored). It is clear that the relative rates of excitation of the  $^1S$  and  $^1D$  levels depend very strongly on  $T$ , so the relative strength of the lines emitted by these levels may be used to measure electron temperatures (Osterbrock & Ferland 2006, pg. 108). More specifically, since the  $^1S$  state is at a low enough energy ( $E/k \lesssim 70 \times 10^4$ ), so that the rate for collisional excitation in gas with  $T \sim 10^4$  is not prohibitively slow, then observable line emission will occur from both the  $^1D$  and  $^1S$  levels. Because these levels are at very different energies, the relative strengths of the emission lines will be very sensitive to the temperature of the gas (Draine 2011, pg. 204).

In the low-density limit (collisional deexcitations are negligible), every excitation to the  $^1D$  level results in emission of a pho-

ton either in  $\lambda 5007$  or  $\lambda 4959$ , with relative probabilities given by the ratio of the two transition probabilities, which is very close to 3 to 1. Every excitation of  $^1S$  is followed by emission of a photon in either  $\lambda 4363$  or  $\lambda 2321$ , with the relative probabilities again given by the transition probabilities. Each emission of a  $\lambda 4363$  photon further results in the population of  $^1D$ , and follows the aforementioned radiative cascade; but this contribution is small in comparison with the direct excitation of  $^1D$  and can be neglected (Osterbrock & Ferland 2006, pg. 108). As long as we remain in the low-density limit ( $n_e \ll n_{\text{crit}}$  for the  $^1D$  level) the line ratio  $(j_{\lambda 4959} + j_{\lambda 5007})/j_{\lambda 4363}$  is independent of the density, and only depends on the temperature (the ratio decreases with increasing temperature as  $^1S$  becomes more populated) (Draine 2011, pg. 205). However, at higher densities collisional deexcitation begins to play a role. The lower  $^1D$  term has a considerably longer radiative lifetime than the  $^1S$  term, so it is collisionally deexcited at lower electron densities than  $^1S$ , thus weakening  $\lambda 4959$  and  $\lambda 5007$ . In addition, under these conditions collisional excitation of  $^1S$  from the excited  $^1D$  level begins to strengthen  $\lambda 4363$ . The line ratio therefore decreases as the electron density is increased (Osterbrock & Ferland 2006, pg. 109). Fortunately, the values of  $n_{\text{crit}}$  for [O III] and [N II] ( $\sim 10^5 \text{ cm}^{-3}$ ) are high enough that approximation schemes are justified in many ionized nebulae (e.g., the Orion Nebula, with  $n_e \approx 3000 \text{ cm}^{-3}$ ) (Draine 2011, pg. 205).

Note that this technique requires no information on the distance to the nebula, the amount of  $O^{++}$  present, and so on, as all these factors cancel out (Osterbrock & Ferland 2006, pg. 111). On the other hand, a fundamental assumption is that the levels producing the lines are populated only by collisional excitation. The  $^1D$  level is at sufficiently high energy that as the temperature is lowered below  $\sim 5000\text{K}$ , the rate of collisional excitations becomes very small. This means that the line becomes very weak and difficult to observe; it also means that if the next ionization state (O IV, N III) has an appreciable abundance, radiative recombination with electrons may make a significant contribution to population of the  $^1D$  level. As a result, this method may not be suitable at low temperature (Draine 2011, pg. 207).

Another temperature diagnostic comes from the strengths of the discontinuities in the recombination continuum relative to strengths of recombination lines. The most commonly used discontinuity is the **Balmer jump**<sup>38</sup> at  $\lambda = 364.5 \text{ nm}$ :  $BJ \equiv I_\lambda(\lambda_{\text{BJ,blue}}) -$

<sup>38</sup> The Balmer jump is also used to measure the temperature of stellar atmospheres. In brief, since any photon with  $\lambda < 364.5 \text{ nm}$  is capable of ionizing

$I_\lambda(\lambda_{\text{BJ,red}})$ , where  $\lambda_{\text{BJ,blue}}$  is chosen to be just blueward of the jump, and  $\lambda_{\text{BJ,red}}$  is chosen to be slightly redward of the jump, and to be located between H recombination lines. The jump discontinuity is produced by recombining electrons with zero kinetic energy, and is therefore proportional to the electron energy distribution at  $E = 0$  (Draine 2011, pg. 212). A difficulty in using this measure is that continuous radiation emitted by stars involved in the nebulae and scattered by interstellar dust may have a sizeable Balmer discontinuity, which is difficult to disentangle from the true nebular recombination Balmer discontinuity (Osterbrock & Ferland 2006, pg. 115).

Another completely independent temperature determination can be made from radio-continuum observations. This idea is quite straightforward – namely, that at sufficiently low frequencies any nebula becomes optically thick, and therefore, at these frequencies the emergent intensity is a blackbody distribution, allowing a temperature determination, assuming an isothermal nebula. The difficulty with applying this method is that the wavelengths at which nebula become optically thick ( $\lambda \sim 1 \text{ m}$ ), even the largest radio telescopes have beam sizes that are comparable to or larger than the angular diameters of typical H II regions. Therefore, the nebula does not completely fill the beam, and a correction must be made for the projection of the nebula onto the antenna pattern (Osterbrock & Ferland 2006, pg. 117). Temperature can also be estimated by comparing the ratio of the 21-cm absorption line to the Ly $\alpha$  absorption line. This measures the excitation temperature (often called the spin temperature) of the hyperfine structure levels that produce the 21-cm line and is equal to the gas kinetic temperature if the level populations are determined mainly by collisions. This is usually the case when the electron density is above the critical density ( $\sim 10^{-7} \text{ cm}^{-3}$ ) of the transition which is true in most conditions (Osterbrock & Ferland 2006, pg. 120). Finally, for some ions it is possible to observe both collisionally excited lines and lines emitted following **dielectronic recombination**<sup>39</sup>. For example, electrons colliding with C IV can produce collisionally excited states of C IV, but can also produce excited levels of C III by dielectronic recombination. Because the rate coefficients for collisional excitation and for dielectronic recombination will have different temperature dependencies, the ratio of dielectronic lines to collisionally excited lines will be temperature-sensitive, and therefore useful as a temperature diagnostic (Draine 2011, pg. 213). Moreover, since both depend on the product of densities  $n(\text{C}^{4+})n_e$  this effect cancels out in their ratios (Osterbrock & Ferland 2006, pg. 113).

#### DENSITY DIAGNOSTICS

We can measure the density of a nebula using ions with two or more energetically accessible energy levels that are at nearly the same energy, so that the relative rates for populating these levels by collisions are nearly independent of temperature. The ratio of the level populations will have one value in the low-density limit, where every collisional excitation is followed by spontaneous radiative decay, and another value in the high-density limit, where the levels are populated in proportion to their degeneracies. Generally, the relative level populations in these two limits differ and can therefore be used to determine the density in the emitting region (Draine 2011, pg. 203). The best example of lines that may be used to measure the electron density are [O II]  $\lambda\lambda 3729/\lambda 3726$  and [S II]  $\lambda 6716/\lambda 6731$ , with energy-level diagrams shown in Figure 72.

Consider the example of [O II] in the low-density limit  $n_e \rightarrow 0$ , in which every collisional excitation is followed by emission of a photon. Since the relative excitation rates of the  $^2\text{D}_{5/2}$  and  $^2\text{D}_{3/2}$  levels are proportional to their statistical weights, the ratio of strengths of the two lines is  $j_{\lambda 3729}/j_{\lambda 3726} = 1.5$ . On the other hand, in the high-density limit,  $n_e \rightarrow \infty$ , collisional excitations and deexcitations dominate and set up a Boltzmann populations ratio, yielding  $j_{\lambda 3729}/j_{\lambda 3726} = (g_{\lambda 3729}/g_{\lambda 3726})(A_{\lambda 3729}/A_{\lambda 3726}) \sim 0.34$ , where  $g$  is the degeneracy of the level (3 and 2 respectively) and  $A$  is its transitional probability. The transition between the high- and low-density limits occurs in the neighbourhood of the critical densities, which are  $\sim 10^3 \text{ cm}^{-3}$  and  $\sim 10^4 \text{ cm}^{-3}$  for  $^2\text{D}_{5/2}$  and  $^2\text{D}_{3/2}$ , respectively. In principle, we can obtain a full solution of the equilibrium equations, taking into account all transitions, including excitation to the  $^2\text{P}$  levels with subsequent cascading downward. Note that from collisional transition rates the main dependence of this ratio is on  $n_e/T^{1/2}$  (Osterbrock & Ferland 2006, pg. 122).

Another independent density diagnostic is to measure fine-structure lines such as [O III]  $^3\text{P}_0 - ^3\text{P}_1$  ( $\lambda 88 \mu\text{m}$ ) and  $^3\text{P}_1 - ^3\text{P}_2$  ( $\lambda 52 \mu\text{m}$ ). These FIR lines have much smaller excitation potentials than the optical lines such as  $^3\text{P}_2 - ^1\text{D}_2$   $\lambda 5007$ . Thus, a ratio like  $j_{\lambda 5007}/j_{\lambda 88\mu\text{m}}$  depends strongly on temperature but, since the  $^3\text{P}_2$  level has a much lower critical density than  $^1\text{D}_2$ , the ratio depends on density also. On the other hand, the ratio  $j_{\lambda 54\mu\text{m}}/j_{\lambda 88\mu\text{m}}$  hardly depends on temperature at all (since both excitation

hydrogen from its first excited state ( $n = 2$ ), the opacity of stellar material suddenly increases at  $\lambda$  below this value and the radiative flux measured for a star decreases accordingly. The size of this abrupt drop in the continuous spectrum of a star depends on the fraction of hydrogen atoms in the first excited state, which is controlled by temperature via the Boltzmann equation (Carroll & Ostlie 2007, pg. 247).

<sup>39</sup> This is an alternative to radiative recombination. In this mechanism, an electron is captured by an ion  $X^{i+}$  and the excess energy of the recombination is taken up by a second electron which then also occupies an excited state; thus, the initial recombination is radiationless. The doubly excited ion  $X^{(i+1)+}$  relaxes either by reversing the process or by radiatively cascading down. The dielectronic recombination is a resonant process, because of the discrete energy nature of the bound electron orbits. See <http://www.astro.umd.edu/~chris/publications/thesis/node148.html> and <http://www.nist.gov/pml/div684/grp01/research-324.cfm>.

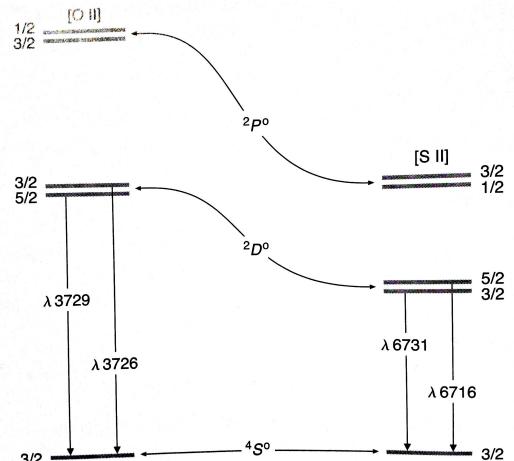


FIG. 72.— Energy-level diagrams of the  $2\text{p}^3$  ground configuration of [O III] and  $3\text{p}^3$  ground configuration of [S II]. Image taken from Osterbrock & Ferland (2006).

potentials are so low in comparison with typical nebular temperatures), but does depend strongly on density (since the two upper levels have different critical densities). Hence, by measuring two [O III] ratios, we can determine the average values of the two parameters,  $T$  and  $n_e$  (Osterbrock & Ferland 2006, pg. 127).

Finally, the **Balmer decrement** can be used to probe  $n_e$  in the high-density limit. This refers to the sequence of lines  $I(\text{H}\alpha)/I(\text{H}\beta)$ , 1,  $I(\text{H}\gamma)/I(\text{H}\beta)$ ,  $I(\text{H}\delta)/I(\text{H}\beta)$ , and so on. These line ratios are relatively insensitive to the electron temperature, and at low density are independent of density. Therefore, comparison of the observed line ratios to theoretical line ratios is usually used to determine the degree of reddening due to dust. However, at high densities the line ratios are affected by collisional effects, with systematic enhancement of the high- $n$  levels relative to H $\alpha$  and H $\beta$ , and these line ratios can therefore be used to constrain the electron density when  $n_e > 10^4 \text{ cm}^{-3}$  (Draine 2011, pg. 213).

#### ABUNDANCE DIAGNOSTICS

The abundance of He relative to H is determined from comparison of the strengths of radiative recombination lines of H and He in regions ionized by stars that are sufficiently hot ( $T_{\text{eff}} \gtrsim 4 \times 10^4 \text{ K}$ ) so that He is ionized throughout the H II zone. The abundances relative to H of elements heavier than He can be inferred by comparing the strengths of emission lines excited by collisions with electrons to emission resulting from recombination of electrons with H $^+$ . For instance, the abundance of O $^{++}$  relative to H can be obtained from the ratio of [O III]  $\lambda 5007$  to H $\beta$ :

$$\frac{I(\text{[O III]})}{I(\text{H}\beta)} \approx \text{const.} \frac{n_e n(\text{O III})}{n_e n(\text{H}^+) \alpha_{\text{H}\beta}}., \quad (253)$$

where the constant arises from the quantum-mechanical transition processes of each atomic species (it is actually temperature-dependent) and the recombination coefficient to H $\beta$  introduces a temperature dependency. Therefore, if  $T$  is known, the abundance ratio of O $^{++}$  to H $^+$  can be obtained from the measured line ratio. To determine the total abundance of oxygen to hydrogen, we must sum over all important ion stages (Draine 2011, pg. 214).

The previous abundance determination from collisionally excited lines requires knowledge of  $T$ , which can be somewhat uncertain. On the other hand, recombination lines of an ion X $^{+r}$  are the result of radiative recombination of X $^{+r+1}$ , which will depend on  $T$  and  $n_e$  in a way very similar to radiative recombination of H $^+$ , thus allowing straightforward comparison of the X $^{+r+1}/\text{H}^+$  ratio. Care must be taken to select lines that will not be excited by optical pumping in the nebula; this can be done, for example, by using lines from levels with different total spin than the ground electronic state of X $^{+r}$ . Another caution is that if the ground term of the recombining species X $^{+r+1}$  has fine structure, the recombination spectrum will depend on the relative populations of the different fine-structure levels, and therefore on  $n_e$  (Draine 2011, pg. 215).

---

STARS AND PLANETS (INCLUDES COMPACT OBJECTS)

---

**QUESTION 1**

**Sketch out an H-R diagram. Indicate where on the main sequence different spectral classes lie. Draw and describe the post main-sequence tracks of both low- and high-mass stars.**

### QUESTION 1

**Sketch out an H-R diagram. Indicate where on the main sequence different spectral classes lie. Draw and describe the post main-sequence tracks of both low- and high-mass stars.**

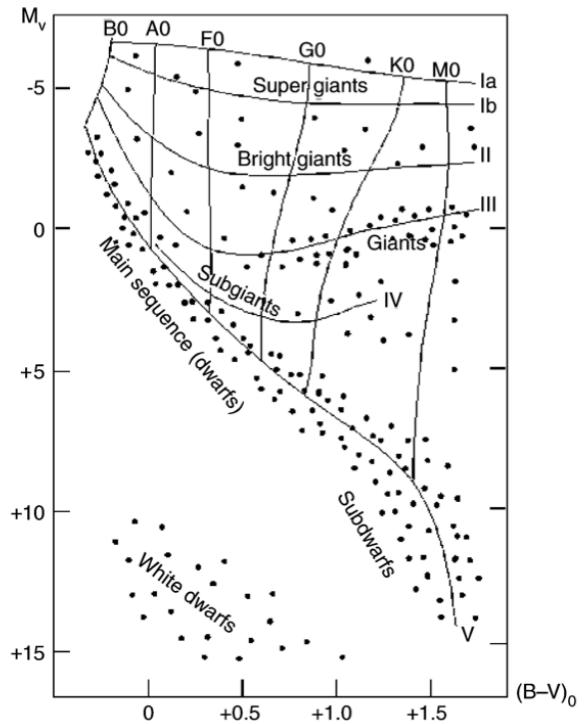
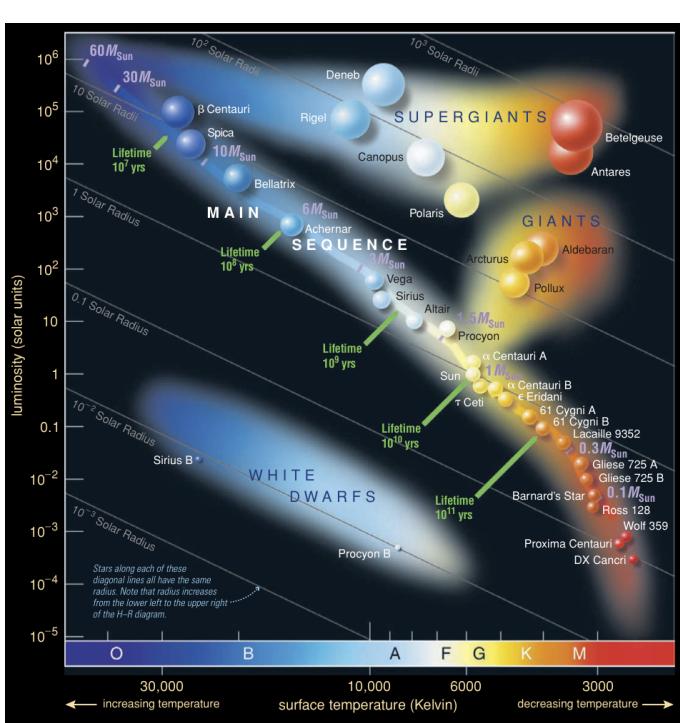


FIG. 73.— (left) An HR diagram with the main regions labelled and example stars identified. Stellar masses and lifetimes are shown on the main sequence and lines of constant radius sweep diagonally from the bottom right to the upper left in the image. Note that spectral classes and associated colours are identified on the temperature axis. Image taken from Bennett et al. (2012). (right) Spectral types and luminosity classes overlaid on a colour-magnitude diagram. Image taken from Schneider (2002).

An HR diagram shows the relationship between stellar luminosity and temperature<sup>40</sup> for stars spanning a wide range in luminosity ( $10^{-4} \lesssim L_{\odot} \lesssim 10^6$ ) and a modest range in temperature ( $3000 \text{ K} \lesssim T \lesssim 30,000 \text{ K}$ ). Alternatively, a colour-magnitude diagram can be plotted with luminosity replaced by absolute magnitude and temperature replaced with colour index, typically  $B-V$  or  $V-I$ <sup>41</sup> (Schneider 2002, pg. 425). As shown in the left panel of Figure 73, there are a variety of regions usually identified on an HR diagram:

- **Main Sequence:** The MS is the most prominent feature on an HR diagram and is traced by stars currently burning hydrogen in their cores. Roughly 85% of stars in the solar neighbourhood are observed to be MS stars since evolution in this stage is governed by the long nuclear timescale ( $\sim 10$  Gyr for the Sun) which is considerably smaller than the free-fall ( $\sim 1$  Myr) and Kelvin-Helmholtz ( $\sim 10$  Myr) timescales. We are thus more likely to observe stars on the MS simply because that stage of evolution requires the most time; earlier and later stages proceed more rapidly (Carroll & Ostlie 2007, pg. 446). The MS is not a line but, rather, has finite width, owing to the changes in a star's temperature and luminosity that occur while it is on the MS and to slight difference in the compositions of stars (Carroll & Ostlie 2007, pg. 221). That fact that most stars are arranged along this one-dimensional sequence tells us that the properties of stars are determined almost solely by their mass. The slope of the MS shows us that the luminosity is a steep function of stellar mass, roughly following the relation  $L \propto M^{3.5}$  (Schneider 2002, pg. 426). Combining the radii and masses known for MS stars, we can calculate their average density, with the results being roughly similar to the density of water. Moreover, moving up the MS, we find that larger, more massive, early-type<sup>42</sup> stars have a lower average density (Carroll & Ostlie 2007, pg. 224).
- **Giants:** Giants represent a portion of the post-MS evolution of low- and intermediate-mass stars that have inflated to relatively large sizes after core hydrogen exhaustion. A variety of important regions lie within this broad portion of the HR diagram: the sub giant branch of hydrogen shell burning; the red giant branch along the Hayashi track, prior to helium core burning; the horizontal branch during helium core burning; the asymptotic giant branch during hydrogen and helium

<sup>40</sup> Here the **effective temperature** is used which is defined to be the temperature a blackbody of the same radius would need to have to emit the same luminosity as the star.

<sup>41</sup> The temperature of a star can be estimated from its colour. From the flux ratio at two different wavelengths or equivalently, from the colour index  $X-Y \equiv m_X - m_Y$  in two filters  $X$  and  $Y$ , the temperature  $T_c$  is determined such that a blackbody at  $T_c$  would have the same colour index. If the spectrum of a star is a Planck spectrum, then the equality  $T_c = T_{\text{eff}}$  would hold, but in general the two differ.

<sup>42</sup> It was once believed that stars begin their lives as young, hot, bright blue O stars, and as they age they become less massive through the exhaustion of nuclear fuel, gradually evolving toward cool, dim, red M stars. Hence the terminology of *early* and *late* spectral types was adopted (Carroll & Ostlie 2007, pg. 220).

TABLE 6  
LUMINOSITY CLASSES

Class	Type of Star
Ia	Bright supergiants
Ib	Normal supergiants
II	Bright giants
III	Normal giants
IV	Subgiants
V	Main-sequence (dwarf) stars
VI	Subdwarfs
D	White dwarfs

shell burning; the Hertzsprung gap marked by the rapid evolution of intermediate-mass stars across the sub giant branch (Carroll & Ostlie 2007, pg. 476).

- **Supergiants:** Supergiants represent a portion of the post-MS evolution of massive stars that have been inflated to extremely large sizes through a succession of core and shell burning. The lack of supergiants observed in the colour-magnitude diagrams of globular clusters is indicative of the rapid MS and post-MS evolution of massive stars.
- **White Dwarfs:** WDs represent the final end product of low- and intermediate-mass stars. The WD mass distribution is fairly sharply peaked (and He WDs cannot be made by single star evolution in a Hubble time). This implies that the vast majority of WDs have the same initial mass and therefore the same initial radius (as from the mass-radius relationship of degenerate matter). The curve traced by cooling WDs is constrained by equation (254) and thus follows a line on the HR diagram (Charles).

Since stars exist which have, for the same spectral type and hence the same temperature, very different luminosities, we can deduce immediately that these stars have very different radii, as can be read from the **Stefan-Boltzmann equation**:

$$L = 4\pi\sigma_{SB}R^2T_{\text{eff}}^4. \quad (254)$$

Therefore, stars on the red giant branch, with their much higher luminosities compared to MS stars of the same spectral class, have a much large radius than the corresponding MS stars. This size effect is also observed spectroscopically: the gravitational acceleration on the surface of a star (i.e. surface gravity) is  $g = GM/R^2$ . We know from models of stellar atmospheres that the width of spectral lines depends on the gravitational acceleration on a star's surface: the lower the surface gravity, the narrower the stellar absorption lines<sup>43</sup>. Hence, a relation exists between the line width and the stellar radius. Since the radius of a star – for a fixed spectral type – specifies the luminosity, this luminosity can be derived from the width of the lines (Schneider 2002, pg. 426). As a result, a star is classified by a **luminosity class**, designated by a Roman numeral appended to its Harvard spectral type. Luminosity classes are overlaid on a colour-magnitude diagram in the right panel of Figure 73 and are summarized in Table 6. Note that subdwarfs reside slightly to the left of the MS because they are deficient in metals (Carroll & Ostlie 2007, pg. 225).

#### EVOLUTION OF LOW- AND INTERMEDIATE-MASS STARS

We will discuss the post-MS evolution of low- and intermediate-mass stars using  $1 M_{\odot}$  stars and  $5 M_{\odot}$  stars as representative case studies. The first difference to note between these two stars is that the former has a radiative core and fuses with the pp chain while the latter has a convective core due to the strong temperature dependence of the CNO cycle.

We will begin by discussing the solar mass star. The end of the MS phase of evolution is defined to occur when hydrogen burning ceases in the core of the star. With the depletion of hydrogen in the core, the generation of energy via the pp chain must stop, with a resultant drop in thermal pressure. For low-mass stars ( $M \lesssim 1.4 M_{\odot}$ ) the contraction of the core after hydrogen burning is stopped by electron degeneracy pressure. However, by now the core temperature has increased to the point that nuclear fusion continues to generate energy in a thick hydrogen-burning shell around a small, predominantly helium core. At the same time, the helium core is nearly isothermal since the luminosity is roughly zero across it. Of course, it is impossible to have a completely isothermal star in hydrostatic equilibrium with a polytrope index of  $\gamma = 1$ . However, an isothermal core is possible with the requirement that, averaged over the entire star  $\bar{\gamma} > 4/3$ , and hence, for a star in hydrostatic equilibrium, only a relatively small fraction of its mass can be in an isothermal core (Marten AST320). For an isothermal core to support the material above it in hydrostatic equilibrium, the required pressure gradient must be the result of a continuous increase in density as the centre of the star is approached. At this point the luminosity being generated in the shell actually exceeds what was produced by the core during hydrogen burning. However, not all of the energy generated reaches the surface; some of it goes into a slow expansion of the envelope. Consequently, the effective temperature begins to decrease slightly and the evolutionary track bends to the right, as shown in Figure 74. As the hydrogen shell-burning continues to consume its nuclear fuel, the resultant ash causes the isothermal helium core to grow in mass while the star moves further red in the HR diagram (Carroll & Ostlie 2007, pg. 451). The star continues a redward trajectory on the HR diagram known as the **subgiant branch (SGB)** on the nuclear timescale of the hydrogen-burning shell (Marten AST320).

<sup>43</sup> The reason for this has to do with **collisional broadening** whereby orbitals of an atom are perturbed through collisions with neutral atoms or close encounters involving the electric field of an ion. The narrower lines observed for the more luminous giant and supergiant stars are due to the lower number densities in their extended atmospheres. Collisional broadening broadens the lines formed in the denser atmospheres of MS stars, where collisions occur more frequently (Carroll & Ostlie 2007, pg. 270).

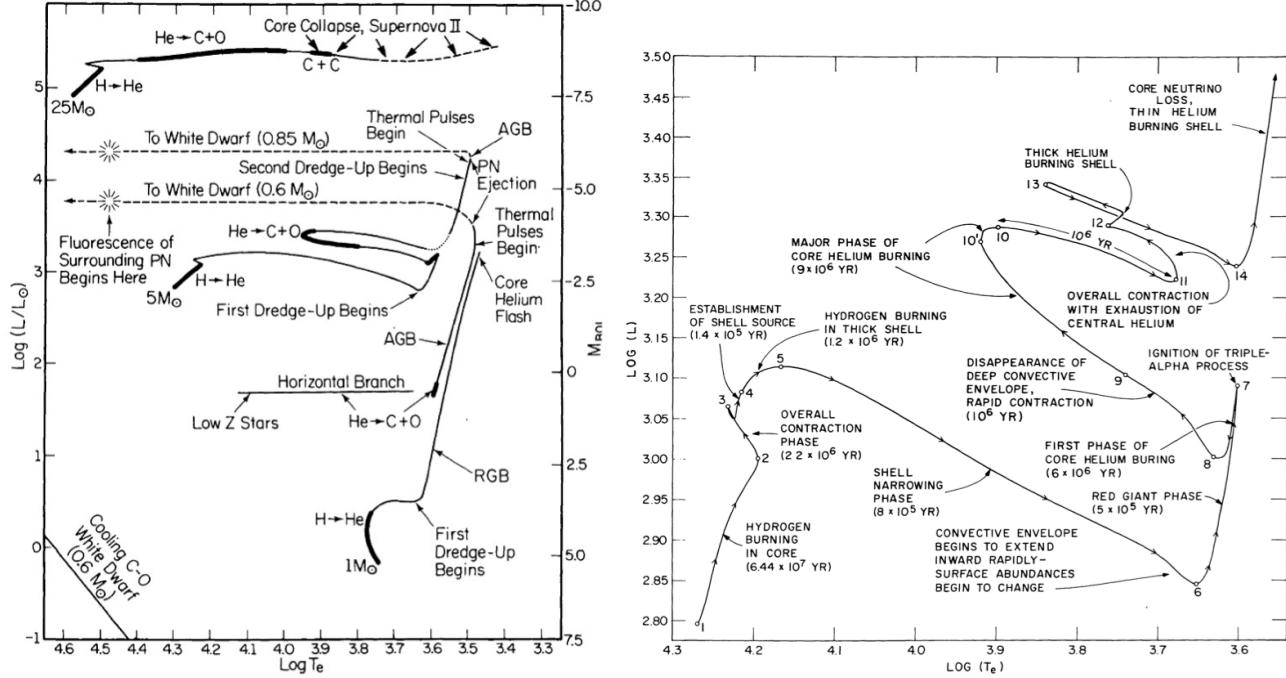


FIG. 74.—(left) Post-MS evolutionary tracks for low-mass ( $1 M_\odot$ ), intermediate-mass ( $5 M_\odot$ ) and high-mass ( $25 M_\odot$ ) stars. Also see Figures 13.4 and 13.5 of Carroll & Ostlie (2007) for the low- and intermediate-mass cases. Image taken from Marten's transient notes. (right) A close-up on the evolutionary track of a  $5 M_\odot$  star with the main stages labelled. Image taken from Marten's AST320 notes.

The core becomes continually more massive as ash is dumped on it from the hydrogen-burning shell. The core becomes more and more dominant as the star evolves since the core grows in mass and shrinks in size, while the envelope becomes more and more tenuous. In the limit that the envelope can be considered weightless, and the shell contains a mass much smaller than that of the core, the properties of the shell depend only on the mass and radius of the core. In particular, one finds that the temperature scales with  $M_{\text{core}}/R_{\text{core}}$ , and thus, for a degenerate core with  $R_{\text{core}} \propto M_{\text{core}}^{-1/3}$ , we have that  $T_{\text{shell}} \propto M_{\text{core}}^{4/3}$ . Hence, the energy production within the shell and its output luminosity increase very steeply with the core mass (Marten AST320). This forces the stellar envelope to keep increasing as it absorbed some of the energy produced by the shell before it reaches the surface. With the expansion of the envelope and decrease in effective temperature, the photospheric opacity increases due to the additional contribution of the  $H^-$  ion. The result is that a convection zone develops near the surface with its base reaching deep into the interior of the star. With the nearly adiabatic temperature gradient associated with convection throughout much of the stellar interior, and the efficiency with which the energy is transported to the surface, the star begins to rise rapidly upward along the **red giant branch (RGB)** of the HR diagram. This path is essentially the same one followed by pre-MS stars descending the Hayashi track prior to the onset of core hydrogen burning (Carroll & Ostlie 2007, pg. 460).

As the star climbs the RGB, its convection zone deepens until the base reaches down into regions where the chemical composition has been modified by nuclear processes. In particular, because of the rather large nuclear reaction cross section, lithium burns via collisions with protons at relatively cool temperatures ( $\sim 10^6$  K). This means that because of the evolution of the star at this point, lithium has become nearly depleted over most of the interior of the star. At the same time, nuclear processing has increased the mass fraction of  $^3\text{He}$  over the middle third of the star as well as altered the abundance ratios of various species in the CNO cycle. When the surface convection zone encounters this chemically modified region, the processed material becomes mixed with the material above it. The effect is observable changes in the composition of the photosphere; the amount of lithium at the surface will decrease and the amount of  $^3\text{He}$  will increase. This transport of material from the deep interior to the surface is referred to as the **first dredge-up** phase (Carroll & Ostlie 2007, pg. 461). Reduced  $^{12}\text{C}/^{13}\text{C}$  and C/N ratios at the stellar surface are also observed and are found to be in fairly good agreement with computational models (Marten AST320).

While the core grows it remains approximately isothermal and at the temperature of the shell surrounding it. In principle, the increase in temperature goes toward lifting the degeneracy, but this is more than compensated for by the increase in core density. Eventually, the density and temperature of the core increase to the point that helium ions (which are not degenerate) start approaching each other close enough that quantum-mechanical tunnelling allows them to overcome their Coulomb barrier, initiating the triple alpha process. This occurs at  $T_{\text{core}} \sim 10^8$  K when the core mass increases to  $M_{\text{core}} = 0.45 M_\odot$ , again independent of the total mass of the star. The fusion will increase the temperature within the core, but will not reduce the density at first, since the pressure exerted by the helium ions is small compared to the electron degeneracy pressure. With increasing temperature and constant density, energy generation increases exponentially, until the thermal pressure becomes high enough to force the core to expand (Marten AST320). By this time the luminosity generated by the helium-burning core reaches  $10^{11} L_\odot$ , comparable to that of an entire galaxy! However, this tremendous energy release lasts for only a few seconds and most of the energy never reaches

the surface. Instead, it is absorbed by the overlying layers of the envelope, possibly causing some mass loss. This short-lived phase of evolution of low-mass stars is referred to as the **helium core flash**. The origin of the explosive energy release is in the very weak temperature dependence of electron degeneracy pressure and the strong temperature dependence of the triple alpha process. The energy generated must first go into “lifting” the degeneracy. Only after this occurs can the energy go into thermal energy required to expand the core, which deceases the density, lowers the temperature, and slows the reaction rate (Carroll & Ostlie 2007, pg. 462).

From detailed models, it turns out that as the degenerate core grows hotter, in its centre the pressure and temperature are sufficiently high that energy is lost in neutrino creation. As a result, the centre will be slightly cooler, and helium core flash ignition will be in a shell around it. Burning will move inwards as the core is heated (possibly in a succession of mini-core-flashes following the main core flash), until degeneracy is lifted throughout the core (Marten AST320).

The evolution during the helium flash is not very well understood, but it appears to be followed by a phase of quiet helium burning in a non-degenerate core. The core will still have  $M_{\text{core}} \sim 0.45 M_{\odot}$ , but its radius will have increased significantly, meaning that the luminosity of the hydrogen-burning shell will have reduced. During this time the position of the star on the HR diagram depends on its metallicity, which determines the opacity in the envelope as well as the efficiency of the CNO cycle. For solar metallicity, stars remain near the Hayashi track, in the so-called **red clump**. For lower metallicities, stars will move to the **horizontal branch** with a generally blueward evolution that is essentially the helium-burning analog of the hydrogen-burning MS, but with a much shorter timescale (Marten AST320).

The situation described above is somewhat different for intermediate-mass ( $1.4 M_{\odot} \lesssim M \lesssim 6 M_{\odot}$ ). Similar to low-mass stars, an isothermal core forms once hydrogen is exhausted in the centre. However, rather than a thick hydrogen-burning shell immediately producing energy with the cessation of hydrogen burning in the core, the entire star participates in an overall contraction on a thermal timescale. This contraction phase releases gravitational potential energy, causing the luminosity to increase slightly, the radius of the star to contract, and the effective temperature to increase (see Figure 74). Eventually, the temperature outside the helium core increases sufficiently to cause a thick shell of hydrogen to burn. Because the ignition of the shell is quite rapid, the overlying envelope is forced to expand slightly, absorbing some of the energy released by the shell. As a result, the luminosity decreases momentarily and the effective temperature drops (Carroll & Ostlie 2007, pg. 457).

For these stars, this phase of evolution ends when the mass of the isothermal core has become too great that the core is no longer capable of supporting the material above it. The maximum fraction of a star’s mass that can exist in an isothermal core and still support the overlying layers is known as the **Schönberg-Chandrasekhar limit** which is derived from the virial theorem and is a function of the mean molecular weights of the core and envelope; it is roughly 0.08. When the mass of the isothermal helium core exceeds this limit, the core collapses on the relatively rapid thermal timescale. The gravitational energy released by the rapidly contracting core again causes the envelope to expand and the effective temperature to drop, resulting in redward evolution on the SGB (Carroll & Ostlie 2007, pg. 458). The rapid progression of intermediate-mass stars across the SGB leads to the existence of the **Hertzsprung gap** in the colour-magnitude diagrams of young star clusters (Carroll & Ostlie 2007, pg. 477). This is in contrast to the much longer nuclear timescale that low-mass stars take across the SGB.

The star then proceeds in a similar manner as discussed before as it moves up the RGB and stabilizes at the tip when helium is ignited in the core. At this phase, the core (which initially has a mass of  $M_{\text{core}} \sim 0.75 M_{\odot}$  for the  $5 M_{\odot}$  star) hardly notices that there is another  $4 M_{\odot}$  of shell and envelope around it, and its structure and luminosity are very similar to what they would have been if the core had been an isolated helium MS star. This reflects the fact that the envelope has become so dilute that it exerts negligible pressure. Like for the low-mass stars, the conditions in the hydrogen-burning shell depend almost completely on the properties of the helium-burning core. When the helium core evolves, its effective temperature will at first, like that of a hydrogen MS star, become slightly lower, and its radius will become slightly bigger. As a result, the hydrogen shell becomes less luminous. Since the shell produces most of the star’s luminosity, the luminosity will drop somewhat. The mass of the helium core, however, will increase, and this causes the core to move upward in mass along the helium MS, towards somewhat larger radius and higher temperature. The higher temperature causes an increase in the energy production in the shell, and therewith a rise in the star’s luminosity (Marten AST320).

When the intermediate-mass star reaches its most blueward point on the HB, the mean molecular weight of the core has increased to the point that the core begins to contract, accompanied by the expansion and cooling of the star’s envelope. Shortly after beginning the redward portion of the HB loop, the core helium is exhausted, having been converted to carbon and oxygen. Again the redward evolution proceeds rapidly as the isothermal C-O core contracts, much like the rapid evolution across the SGB following the extinction of core hydrogen burning. This contraction is again the result of reaching the Schönberg-Chandrasekhar limit where now the mass of the C-O core should be measured relative to the mass of the helium star. With the increase in core temperature associated with its contraction, a thick helium-burning shell develops outside the C-O core. As the core continues to contract, the helium-burning shell narrows and strengthens, forcing the material above the shell to expand and cool. This results in a temporary turn-off of the hydrogen-burning shell. Along with the contraction of the core, neutrino production increases to the point that the core cools a bit. As a consequence of the increasing central density and decreasing temperature, electron degeneracy pressure becomes an important component of the total pressure in the C-O core (Carroll & Ostlie 2007, pg. 463).

When the redward evolution reaches the Hayashi track, the evolutionary track bends upward along a path referred to as the **asymptotic giant branch (AGB)**. Although two shell sources exist at this stage, the helium-burning shell dominates the energy output during the AGB; the hydrogen-burning shell is nearly inactive at this point. The expanding envelope initially absorbs much of the energy produced by the helium-burning shell. As the effective temperature continues to decrease, the convective envelope deepens again, this time extending downward to the chemical discontinuity between the hydrogen-rich outer layer and the helium-rich region above the helium-burning shell. The mixing that results during this **second dredge-up** phase increases the helium and nitrogen content of the envelope; the increase in N is due to the previous conversion of carbon and oxygen

into nitrogen in the intershell region (Carroll & Ostlie 2007, pg. 463). Low-mass stars, where the hydrogen-burning shell is not extinguished, do not experience second dredge-up. Eventually, the core becomes degenerate and as the helium shell eats outward, it comes close to the position where second dredge-up has left hydrogen-rich material, and the hydrogen shell is reignited. From here on, the evolution between low- and intermediate-mass stars become similar (Marten AST320).

The upper portion of the AGB is known as the **thermal-pulse AGB (TP-AGB)**. At this point, the dormant hydrogen-burning shell eventually reignites and again dominates the energy output of the star. However, during this phase of evolution, the narrowing helium-burning shell begins to turn on and off quasi-periodically. These intermittent **helium shell flashes** occur because the hydrogen-burning shell is dumping helium ash onto the helium layer below. As the mass of the helium layer increases, its base becomes slightly degenerate. Then, when the temperature at the base increases sufficiently, a helium shell flash occurs, analogous to the earlier helium core flashes, although much less energetic. This drives the hydrogen-burning shell outward, causing it to cool and turn off for a time. Eventually the burning in the helium shell diminishes, the hydrogen-burning shell recovers, and the process repeats. The period between pulses is a function of the mass of the star, ranging from thousands of years for stars near  $5 M_{\odot}$  to hundred of thousands of years for  $0.5 M_{\odot}$ , with the pulse amplitude of growing with each successive event. This phase of periodic activity in the deep interior is evident in abrupt changes in luminosity at the surface. In particular, the luminosity rapidly decreases when a helium flash occurs since the energy-dominant hydrogen-burning ceases and the star expands. After a period of time, the energy output of the helium shell diminishes when degeneracy is lifted and the hydrogen-burning shell moves deeper into the star, again dominating the luminosity (Carroll & Ostlie 2007, pg. 464).

Because of the sudden increase in energy flux from the helium-burning shell during a flash episode, a convection zone is established between the helium and hydrogen-burning shells. At the same time, the depth of the envelope convection zone increases with the pulse strength of the flashes. For  $M > 2 M_{\odot}$ , the convection zones will merge and eventually extend down into regions where carbon has been synthesized. In the region between the hydrogen- and helium-burning shells, the abundance of carbon exceeds that of oxygen by a factor of five to ten. This is in sharp contrast to the general excess of oxygen over carbon in the atmospheres of most stars. During this **third dredge-up** phase, the carbon-rich material is brought to the surface, decreasing the ratio of oxygen to carbon. This can eventually lead to carbon-rich giants known as **carbon stars**. Since CO is a tightly bound molecule, almost all of the oxygen will be locked up in this phase while the abundant carbon is free to form exotic species like SiC. This is in contrast to M-type stars where all of the carbon is locked up in CO while the abundant O is free to form species like SiO. The interesting spectral features associated with carbon stars have given them the designation of C spectral type (Carroll & Ostlie 2007, pg. 466).

AGB stars are known to lose mass at a rapid rate. The effective temperatures of these stars are also quite cool ( $\sim 3000$  K). As a result, dust grains form in the expelled matter with silicate grains and graphite grains forming in oxygen-rich and carbon-rich material, respectively. The mechanisms that cause this mass loss are poorly known. Some have suggested it is linked to the helium shell flashes while others believe it stems from the high luminosities and low surface gravities coupled with radiation pressure on the dust grains, dragging the gas with them. As one might expect, the rate of mass loss accelerates with time because the luminosity and radius are increasing while the mass is decreasing during continued evolution up the AGB. The decreasing mass and increasing radius of the star imply less surface gravity, so that the surface material is becoming progressively less tightly bound. In the latest stages of evolution on the AGB, a **superwind** develops with  $\dot{M} \sim 10^{-4} M_{\odot}\text{yr}^{-1}$ . These observed high mass loss rates seem to be responsible for the existence of a class of objects known as OH/IR sources. These objects appear to be stars shrouded in optically thick dust clouds that radiate IR light and also show the presence of OH molecules through **maser emission**<sup>44</sup> (Carroll & Ostlie 2007, pg. 467-468).

As stars with initial masses below  $6 M_{\odot}$  continue to evolve up the AGB, the helium-burning shell converts more and more of the helium into carbon and then into oxygen, increasing the mass of the carbon-oxygen core. At the same time, the core continues to contract slowly, causing its central density to increase. Depending on the star's mass, neutrino energy losses may decrease the central temperature somewhat during this phase. In any event, the densities in the core become large enough that electron degeneracy pressure begins to dominate. For  $M < 4 M_{\odot}$  the C-O core will never become large and hot enough to ignite nuclear burning. For higher mass, the stars experience additional nucleosynthesis in their cores, leading to core compositions of oxygen, neon and magnesium (ONeMg cores) with final masses remaining below the Chandrasekhar limit of  $1.4 M_{\odot}$ <sup>45</sup> (Carroll & Ostlie 2007, pg. 467).

During the ensuing final phase of mass loss, the remainder of the star's envelope is expelled thereby extinguishing the hydrogen- and helium-burning shells. As a result, the luminosity of the star rapidly drops. The hot central object, now revealed, will cool to become a **white dwarf**, which is essentially the old red giant's degenerate C-O core (or ONeMg core for more massive stars), surrounded by a thin layer of residual H and He. The expanding shell of gas around a WD progenitor is called a **planetary nebula**. A PN owes its appearance to the UV light emitted by the hot, condensed central star. The UV photons are absorbed by the gas in the nebula, causing the atoms to become excited or ionized, emitting visible light when they relax. The bluish-green colouration of many PN is due to the 5007 Å and 4959 Å forbidden lines of [O III] while the red comes from ionized H and N. PN have characteristic temperatures of  $10^4$  K, lengths of 0.3 pc, and lifetimes of  $10^4$  yr (Carroll & Ostlie 2007, pg. 472).

#### EVOLUTION OF HIGH-MASS STARS

For massive ( $M \gtrsim 6 M_{\odot}$ ) the convective core at hydrogen exhaustion already exceeds the Schönberg-Chandrasekhar limit and therefore an isothermal core cannot form. Instead, the core contracts on a thermal timescale until helium fusion ignites. Of course, there is no helium flash for massive stars since their temperatures are high enough that thermal pressure remains strong,

<sup>44</sup> A maser is the molecular analog of a laser; electrons are pumped up from a lower energy level into a higher, longer-lived metastable energy state. The electron then makes a downward transition back to a lower state when it is stimulated by a photon with an energy equal to the difference in energies between the two states.

<sup>45</sup> If not for the excessive mass loss during the AGB phase, stars with  $M > 4 M_{\odot}$  would have cores exceeding the maximum limit and would suffer catastrophic core collapse.

preventing degeneracy pressure from becoming a factor. Helium core fusion therefore ignites gradually and within only a few hundred thousand years most of the helium has been converted into an inert carbon core (Bennett et al. 2012, pg. 348). For the  $25 M_{\odot}$  star shown in Figure 74, helium is exhausted while the star is only midway over to the RGB. At that point, the core contracts further until carbon is ignited. Prior to this both helium-burning and hydrogen-burning shells had been established above the core. Despite the dramatic events taking place in the interior, the star's outer appearance changes slowly. As each stage of fusion ceases, fusion in the surrounding shells intensifies and further inflates the outer layers. Each time the core flares up, the outer layers contract somewhat, but the overall luminosity remains nearly constant (Bennett et al. 2012, pg. 348). The result is that the evolutionary track zigzags across the top of the HR diagram. Further evolution beyond carbon fusion proceeds very rapidly and the star soon explodes as a SN (Marten AST320).

The evolution of massive stars is complicated greatly by mass loss, even on the MS. Due to mass loss, the whole hydrogen-rich envelope may disappear, in which case the star becomes a helium star, and moves to high temperatures in the HR diagram. Indeed, for very massive stars, this is virtually unavoidable, as their luminosity on the way to the red supergiant branch exceeds the Eddington luminosity limit where the force due to radiation pressure may equal or exceed the force of gravity on surface layers of the star. As a consequence, their envelopes are rapidly blown off in what creates an empty region in the top right of the HR diagram, above and to the right of the **Humphreys-Davidson limit**. Stars close to this limit indeed are observed to have extremely large and variable mass-loss rates; these are the so-called **luminous blue variables (LBVs)** (Marten AST320). Closely related to LBVs are **Wolf-Rayet stars (WRs)** which are characterized by their presence of broad emission lines and extremely large effective temperatures up to  $10^5$  K. Whereas LBVs are all very mass stars with  $M \gtrsim 85 M_{\odot}$ , WRs can have progenitor masses as low as  $20 M_{\odot}$ . WRs also do not demonstrate variability that is characteristic of LBVs (Carroll & Ostlie 2007, pg. 521).

What really sets WRs apart from other stars is their unusual spectra. Not only are the spectra dominated by board emission lines, but they also reveal a composition that is decidedly atypical. WRs are generally separated into three classes: WN, WC, and WO. The spectra of WNs are dominated by emission lines of H and N whereas WC stars exhibit emission lines of He and C, with a distinct absence of H and N lines. On the other hand, the WO stars, which are much rarer than either WNs or WCs, have spectra containing predominantly O lines, with some contribution from highly ionized species. This strange trend in composition has been recognized to be a direct consequence of the mass loss of these stars. WNs have lost virtually all of their hydrogen-dominated envelopes, revealing material synthesized by nuclear reactions in the core. Convection in the core of the star has brought equilibrium CNO cycle-processed material to the surface. Further mass loss results in the ejection of the CNO processed material, exposing helium-burning material generated by the triple alpha process. Then, if the star survives long enough, mass loss will eventually strip away all but the oxygen component of the triple-alpha ash (Carroll & Ostlie 2007, pg. 522).

#### EXPANDING ENVELOPE

In the previous discussions, we have generally ignored the tenuous envelope since it does not play an important role in stellar evolution. Nevertheless, it is worthwhile to investigate what physical mechanisms dominate the expansion of the envelope. Quite generally, as a star becomes more luminous, its radius increases and its effective temperature decreases slightly. This in itself is not enough to bring the star over to the red giant regime. As the temperature in the outer layers decreases, however, the opacity there increases strongly, since it is dominated by bound-free processes (the lower temperature leads to lower ionization states of the metals, which therefore can absorb photons more easily). Therefore, the luminosity cannot easily be transported anymore and part of it is trapped, leading to further expansion. At some point, this apparently can become a runaway process, in which the envelope cools more and more, becomes more and more opaque, traps more and more of the luminosity, and expands to larger and larger radii. It only stops when the star reaches the Hayashi line, where the envelope has become almost completely convective, and energy can be transported more easily. In comparison, when the luminosity decreases, it appears the inverse instability can happen, where the envelope heats a little, becomes less opaque, therefore shrinks slightly, releasing energy which increases the temperature, etc. This deflation instability might be responsible for the blue loops seen in the evolutionary tracks of intermediate-mass stars (Marten AST320).

**QUESTION 2**

**Sketch a plot of radius versus mass for various “cold” objects, including planets, brown dwarfs and white dwarfs.  
Explain the mass-size relationship for rocky and gaseous objects.**

## QUESTION 2

**Sketch a plot of radius versus mass for various “cold” objects, including planets, brown dwarfs and white dwarfs. Explain the mass-size relationship for rocky and gaseous objects.**

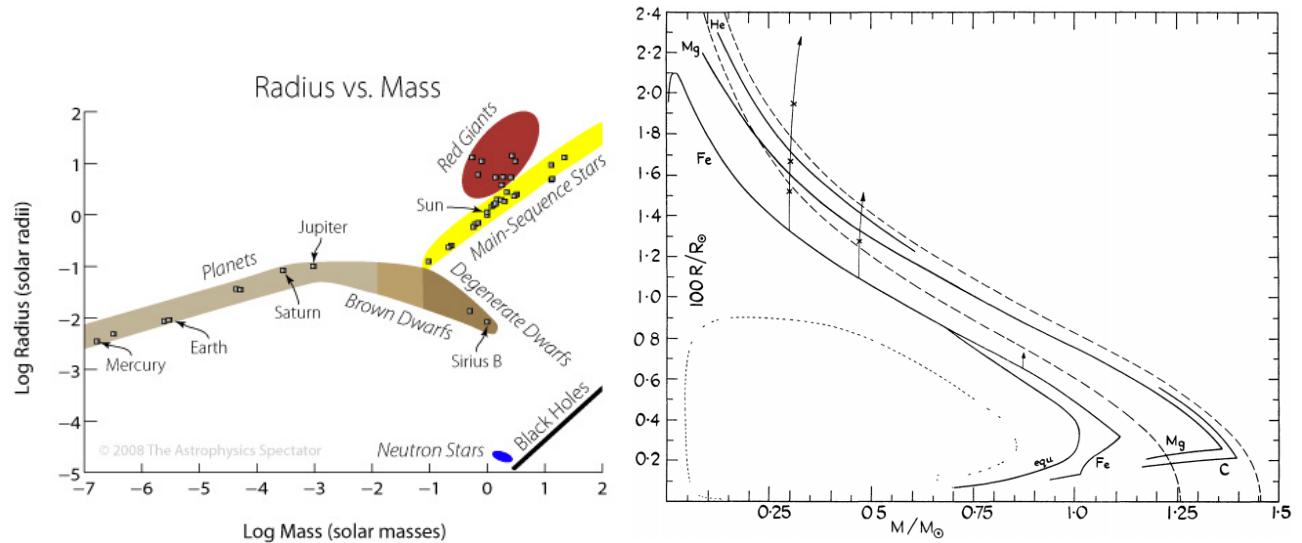


FIG. 75.— (left) Radius versus mass for various astronomical objects. Planets, brown dwarfs, and white dwarfs (labelled as degenerate dwarfs here) lie on a continuous band extending from the lowest masses to  $1.4 M_{\odot}$ . MS stars, red giants, neutron stars, and black holes are also labelled. Image taken from <http://www.astrophysicsspectator.com/topics/overview/SizeStarsPlanets.html>. (right) Mass-radius relation for white dwarfs of various compositions. The dashed curves indicate Chandrasekhar models for  $\mu = A/Z = 2$  (upper) and  $\mu = 2.15$  (lower). These models deviate from the idealized curves (solid lines) because elements are not completely ionized, and at very high densities, inverse beta decay becomes important. The arrows indicate the effects of adding a hydrogen atmosphere. Image taken from Hamada & Salpeter (1961).

The sizes of various astronomical objects is determined by the detailed physical processes affecting their interior structures. Simply stated, for astronomical objects, sizes are dictated by the balance between the opposing forces of gravitational self-attraction and an interior pressure gradient. This is summarized by the condition of **hydrostatic equilibrium**:

$$\frac{dP}{dr} = -G \frac{M(r)\rho}{r^2} = -\rho g, \quad (255)$$

where  $g \equiv GM(r)/r^2$  is the local acceleration of gravity at radius  $r$ . This equation is valid for spherically symmetric objects that are sufficiently static so that accelerations within the interior are negligible. The form of equation (255) indicates that a pressure gradient,  $dP/dr$ , must exist to counteract the force of gravity; it is not the pressure that supports the object, but the change in pressure with radius (Carroll & Ostlie 2007, pg. 287). Various forms of pressure support are available for astronomical objects, though when discussing “cold” objects, we can assume that thermal and radiation pressure are unimportant.

We will begin by discussing the physics of electron degeneracy pressure, which provides support for both brown and white dwarfs. We begin by noting that any system consists of quantum states that are identified by a set of quantum numbers. For example, a box of gas particles is filled with standing de Broglie waves that are described by three quantum numbers, specifying the particle’s component of momentum in each of the three spatial dimensions. If the gas particles are fermions, then the Pauli exclusion principle allows at most one fermion in each quantum state because no two fermions can have the same set of quantum numbers. In an everyday gas at standard temperature and pressure, only one in every  $10^7$  quantum states is occupied by a gas particle, and the limitations imposed by the Pauli exclusion principle are unimportant. However, as energy is removed from the gas its temperature falls, forcing an increasingly larger fraction of the particles into lower energy states. For a fermion gas, as the temperature of the gas is lowered, the fermions will fill up the lowest available unoccupied states, starting with the ground state, and then successively occupy the excited states with the lowest energy. Even in the limit  $T \rightarrow 0$  K, the vigorous motion of all the fermions in excited states produces a pressure in the fermion gas, known as **degeneracy pressure**. The maximum energy of any fermion in a completely degenerate gas at  $T = 0$  K is called the **Fermi energy**. The Fermi energy is  $\epsilon_F \propto n_e^{2/3}$  and the average energy per fermion at absolute zero is  $3/5\epsilon_F$  (Carroll & Ostlie 2007, pg. 563).

At any temperature above absolute zero, some of the states with an energy less than  $\epsilon_F$  will become vacant as fermions use their thermal energy to occupy other, more energetic states. Although the degeneracy will not be precisely complete when  $T > 0$  K, the assumption of complete degeneracy is a good approximation at the densities encountered in the interior of a WD. To understand how the degree of degeneracy depends on both the temperature and density of a WD, we first express the Fermi energy in terms of the density of the electron gas. For full ionization, the number of electrons per unit volume is:

$$n_e = \left( \frac{\# \text{ electrons}}{\text{nucleon}} \right) \left( \frac{\# \text{ nucleons}}{\text{volume}} \right) = \left( \frac{Z}{A} \right) \frac{\rho}{m_H}, \quad (256)$$

where  $Z$  and  $A$  are the number of protons and nucleons, respectively, in the WD's nuclei, and  $m_H$  is chosen to be representative of the mass of the proton and neutron. This relation can be used to compute  $\epsilon_F$  noting that it is proportional to the  $2/3$  power of  $n_e$ . Then the Fermi energy can be compared to the average thermal energy of an electron,  $3/2kT$ . In rough terms, if  $3/2kT < \epsilon_F$ , then an average electron will be unable to make a transition to an unoccupied state, and the electron gas will be degenerate. Hence, the condition for degeneracy may be written as

$$\frac{T}{\rho^{2/3}} < \text{const.}, \quad (257)$$

where the constant can be expressed in terms of fundamental quantities like  $\hbar$ ,  $k$ ,  $m_e$ , and  $m_H$  (see Carroll & Ostlie 2007, pg. 566).

We can determine how electron degeneracy pressure scales with density. Note that **Heisenberg's uncertainty principle**,

$$\Delta x \Delta p_x \approx \hbar, \quad (258)$$

states that an electron confined to a small space must have a correspondingly high uncertainty in its momentum. In a completely degenerate electron gas, the electrons are packed as tightly as possible, and for a uniform number density  $n_e$ , the separation between neighbouring electrons is about  $n_e^{-1/3}$ . However, to satisfy the Pauli exclusion principle, the electrons must maintain their identities as different particles. That is, the uncertainty in their positions cannot be larger than their physical separation. Identifying  $\Delta x \approx \hbar n_e^{-1/3}$  for the limiting case of complete degeneracy, we can use equation (258) to estimate the momentum of an electron as  $p_e \approx \Delta p \approx \hbar n_e^{1/3}$  (Carroll & Ostlie 2007, pg. 568). In addition, we know that the total pressure exerted by the electron gas is equal to

$$P_D = (\text{kinetic energy per unit volume}) = n_e E_e, \quad (259)$$

where  $E_e$  is the average kinetic energy per electron. For nonrelativistic electrons, we know that  $E_e = p_e^2/2m_e$ , while for relativistic electrons we have  $E_e = p_e c$  (Townsend 1997). Using our result that  $p_e$  is proportional to the  $1/3$  power of  $n_e$  and that  $n_e \propto \rho$ , we find that

$$P_D \propto \begin{cases} \rho^{5/3} & \text{nonrelativistic electrons} \\ \rho^{4/3} & \text{relativistic electrons} \end{cases} \quad (260)$$

To determine the **mass-radius relation** for WDs, we first use equation (255) to estimate how the pressure of an object in hydrostatic equilibrium scales with mass  $M$  and radius  $R$ , and then relate this pressure to that found in equation (260). If we use the (unrealistic) assumption of uniform density then from equation (255) we see that

$$\frac{P}{R} \propto \frac{M}{R^2} \frac{M}{R^3} \rightarrow P \propto \frac{M^2}{R^4}. \quad (261)$$

If we use the result obtained in equation (260) we find that

$$R \propto M^{-1/3}, \quad (262)$$

for the nonrelativistic case. For the relativistic case, one finds that  $M$  and  $R$  become a constant, which of course are just the Chandrasekhar mass and radius. Equation (262) shows that the volume of a WD is inversely proportional to its mass, so more massive WDs are actually smaller. This behaviour is a result of the star deriving its support from electron degeneracy pressure. The electrons must be more closely confined to generate the larger degeneracy pressure required to support a more massive star. According to this relation, however, piling more and more mass onto a WD would eventually result in shrinking the star down to zero volume as the mass becomes infinite. Of course, at some point the density of the WD will become large enough that the electrons will become relativistic. When the relativistic regime is entered, the electrons will be moving more slowly than the nonrelativistic approximation would suggest. Hence, less electron pressure will be available to support the star. Consequently, a massive WD is *smaller* than predicted by equation (262). Indeed, zero volume occurs for a finite value of the mass; in other words, there is a limit to the amount of matter that can be supported by electron degeneracy pressure (Carroll & Ostlie 2007, pg. 570). The right panel of Figure 75 shows the mass-radius relation for WDs of various chemical compositions.

Now that we have a good understanding of the physics behind the mass-radius relation for WDs, we will move on to discuss planets and brown dwarfs. First, we discuss the least-massive planets for which pressure support is accomplished by the Coulomb interactions between nuclei within the interior. For this case, Coulomb effects set a fixed density that is nearly independent of mass since the interparticle spacing within the interior ( $\sim \text{\AA}$ ) does not change appreciably as the pressure increases<sup>46</sup>. That is, the rocky interior of small planets resists compression under the application of an external source of pressure, like the mass of overlying layers. Because we can assume these planets have a constant density  $\rho$ , this suggests that the radius is a weakly increasing function of mass, with  $R \propto M^{1/3}$  (Hubbard et al. 2002).

The  $R \propto M^{1/3}$  is generally valid for small terrestrial planets that are made predominantly of rocks and metals. Once we begin to look at larger planets we enter the regime of extrasolar giant planets (EGPs), which have overall compositions much more similar to the composition of stars than to terrestrial planets. In particular, we find that EGPs are made almost entirely of hydrogen and helium, with just a few percent of their masses coming from hydrogen compounds and even smaller amounts of rock and metals

<sup>46</sup> See <http://www.astrophysicsspectator.com/topics/overview/SizeStarsPlanets.html> and the links therein for additional information pertaining to this question.

(Bennett et al. 2012, pg. 236). Within the interiors of these massive planets, the pressure and temperature can reach such extreme values that at some point molecular hydrogen makes a transition from being gaseous to almost liquid, and then at  $T \sim 10^3$  K and  $P \sim 1$  Mbar, has a phase transition to **metallic hydrogen**<sup>47</sup> that is electron-degenerate (Hubbard et al. 1997). As an example, Figure 76 shows the theoretically expected interior structure of Jupiter.

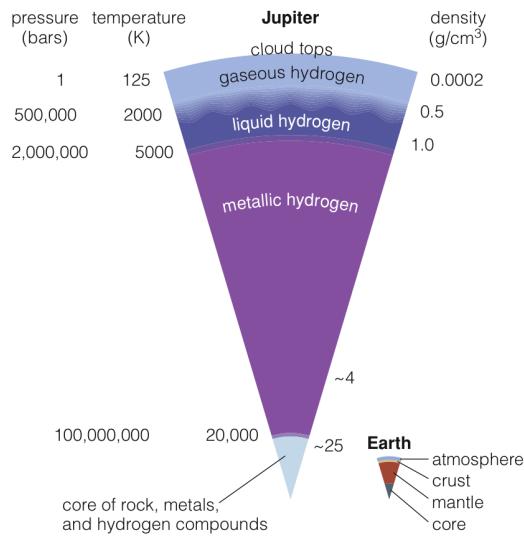


FIG. 76.— Jupiter’s interior structure, labeled with the pressure, temperature, and density at various depths. Earth’s interior structure is shown to scale for comparison. The gaseous and liquid hydrogen layers are primarily molecular with hydrogen in the form of  $\text{H}_2$  whereas the metallic hydrogen may consist of both liquid and solid metallic states. Note that 1 bar is roughly the atmospheric pressure at sea level on Earth and that the density of liquid water is 1  $\text{g}/\text{cc}$ . Image taken from Bennett et al. (2012).

Since EGPs are predominantly metallic hydrogen and helium mixtures, and have fully convective cores, they are the low-mass end of a continuum of objects, known as substellar-mass objects (SMOs), that includes brown dwarfs (BDs) with masses up to  $\sim 0.08 M_\odot \approx 80 M_J$ . Ideally, EGPs and BDs would be distinguished on the basis of mode of formation, though in practice this is not easily determined. At the high-mass end of SMOs, the electron degeneracy established by the metallic interiors produces the same  $R \propto M^{-1/3}$  relation that we saw for WDs. However, over a broad range of intermediate masses spanning  $0.3 M_J$  to  $70 M_J$  we observe a competing effect between the Coulomb and electron-degeneracy effects. The resulting competition between effects that produce opposite results renders the radius nearly constant over roughly two orders of magnitude in mass, near the radius of Jupiter ( $\sim 0.1 R_\odot$ ). However, there can be some scatter around this depending on the age of the planet. In particular, if the planet is younger than its cooling timescale, then thermal pressure will contribute to the overall pressure support, and a larger radius will be observed than would be expected in the absence of thermal effects (Hubbard et al. 1997).

<sup>47</sup> Metallic hydrogen is a state of hydrogen which results when it is sufficiently compressed and undergoes a phase transition; it is an example of degenerate matter. Solid metallic hydrogen is predicted to consist of a crystal lattice of hydrogen nuclei (namely, protons), with a spacing which is significantly smaller than the Bohr radius. Indeed, the spacing is more comparable with the de Broglie wavelength of the electron. The electrons are unbound and behave like the conduction electrons in a metal. In liquid metallic hydrogen, protons do not have lattice ordering; rather, it is a liquid system of protons and electrons. See [http://en.wikipedia.org/wiki/Metallic\\_hydrogen](http://en.wikipedia.org/wiki/Metallic_hydrogen) for more information.

**QUESTION 3**

**Describe the physical conditions which lead to the formation of absorption lines in stars' spectra. What leads to emission lines?**

### QUESTION 3

**Describe the physical conditions which lead to the formation of absorption lines in stars' spectra. What leads to emission lines?**

Generally speaking, the formation of spectral lines is typically expressed in terms of what are called Kirchhoff's Laws: (1) A hot, dense gas or hot solid object produces a continuous spectrum with no dark spectral lines; (2) A hot, diffuse gas produces bright spectral lines; (3) A cool, diffuse gas in front of a source of a continuous spectrum produces dark spectral lines in the continuous spectrum. The first law refers to the phenomena of blackbody radiation emitted by any body above absolute zero. The second law refers to the production of emission lines

which occur when an electron makes a downward transition from a higher to a lower energy orbit. Finally, the third law refers to the production of absorption lines when an electron makes a transition from a lower to a higher energy orbit. In this case only the photons from the background continuum with exactly the right amount of energy are consumed (Carroll & Ostlie 2007, pg. 126).

#### ABSORPTION LINES

In order to understand stellar absorption lines we must consider the treatment of the propagation of light through a gas. For absorption of a beam of wavelength  $\lambda$  the resultant change in its intensity,  $dI_\lambda$ , is proportional to its intensity,  $I_\lambda$ , the distance traveled through the gas,  $ds$ , and the density of the gas,  $\rho$ . That is,

$$dI_\lambda = -\kappa_\lambda \rho I_\lambda ds, \quad (263)$$

where the proportionality constant,  $\kappa_\lambda$  [ $\text{cm}^2 \text{g}^{-1}$ ], is called the absorption coefficient, or opacity (Carroll & Ostlie 2007, pg. 241). The opacity is the cross-section for absorbing photons of wavelength  $\lambda$  per unit mass of stellar material. The total opacity involves a sum over the contributions from all processes responsible for absorbing photons. This involves both true absorption through transitions of atomic electrons (**bound-bound** and **bound-free absorption**) and scattering processes (free-free absorption<sup>48</sup>, **Thomson scattering**<sup>49</sup>, **Compton scattering**<sup>50</sup>, and **Rayleigh scattering**<sup>51</sup>) (see Carroll & Ostlie 2007, pg. 245, for details). It is the bound-bound opacity that is responsible for forming absorption lines in stellar spectra. For these transitions the opacity is small except at the discrete wavelengths capable of producing an upward transition.

Let's now consider the light we observe with a telescope after being emitted from a stellar atmosphere. For this purpose it is useful to define a quantity called the optical depth that is related to the opacity via

$$d\tau_\lambda \equiv -\kappa_\lambda \rho ds, \quad (264)$$

where  $s$  is the distance measured along the photon's path in its direction of motion. We will take  $\tau_\lambda = 0$  at the outermost layer of the stellar atmosphere in the assumption that the light travels unimpeded through empty space on its way to our telescope. Then, by solving the differential equation in equation (263) and making use of equation (264) we see that for pure absorption the decrease in intensity of a ray of light as it travels through the stellar atmosphere from a depth  $\tau_\lambda$  to our telescope is

$$I_\lambda = I_{\lambda,0} e^{-\tau_\lambda}, \quad (265)$$

where  $I_{\lambda,0}$  is its intensity at a depth of  $\tau_\lambda$ . Obviously, the intensity of the ray decreases with increasing optical depth. Indeed, the optical depth may be thought of as the number of mean free paths from the original position to the surface, as measured along the ray's path. As a result, we typically see no deeper into an atmosphere at a given wavelength than  $\tau_\lambda \approx 1$ . With a more precise treatment in Carroll & Ostlie (2007, pg. 263) it can be shown that the average depth we can peer into a star is at  $\tau_\lambda = 2/3$ . Hence, when referring to the effective temperature of a star, we are referring to the temperature at a depth of  $\tau_\lambda = 2/3$ , not at  $\tau_\lambda = 0$ . Thus, the "surface" of a star, which by definition has temperature  $T_{\text{eff}}$ , is *not* at the top of the atmosphere, but deeper down.

Therefore, if the opacity increases at some wavelength, then the actual distance back along the ray to the level where  $\tau_\lambda = 2/3$  decreases for that wavelength. One cannot see as far into murky material, so an observer will not see as deeply into a stellar atmosphere at wavelengths where the opacity is greater than average (i.e., greater than the continuum opacity). This implies that if the temperature of the stellar atmosphere decreases outward, then these higher regions of the atmosphere will be cooler and correspondingly dimmer. As a result, the intensity of the radiation at  $\tau_\lambda = 2/3$  will decline the most for those wavelengths at which the opacity is greatest. It is for this reason that dark absorption lines arise in the continuous spectrum of a star. Hence, the temperature *must* decrease outward for the formation of absorption lines (Carroll & Ostlie 2007, pg. 254). This is the analog for stellar atmospheres of Kirchhoff's third law.

<sup>48</sup> This is a scattering process that takes place when a free electron in the vicinity of an ion absorbs a photon, causing the speed of the electron to increase. In this process the nearby ion is necessary in order to conserve both energy and momentum (an isolated free electron cannot absorb a photon). It may also happen that as it passes near an ion, the electron loses energy by emitting a photon, which causes the electron to slow down. This reverse process is known as free-free emission, or bremsstrahlung.

<sup>49</sup> Here a photon is scattered by a free electron through the process of Thomson scattering. In this process, the electron can be thought of as being made to oscillate in the electromagnetic field of the photon, causing it, in turn, to emit radiation at the same frequency as the incident wave, and thus the wave is scattered. However, because the electron is tiny, it makes a poor target for an incident photon, producing a cross-section roughly a billion times smaller than the hydrogen cross-section for photoionization. Hence, this process is only effective as a source of opacity when the electron density is very large, which requires high temperatures (i.e. interior of stars and atmospheres of the hottest stars).

<sup>50</sup> This is roughly the high-energy analog of Thomson scattering that occurs when a photon is scattered by an electron that is loosely bound to an atomic nucleus. Compton scattering occurs when the wavelength of the photon is much smaller than the atom. If the electron is ejected from the atom, then the energy of the scattered wave is reduced, to compensate for overcoming the work function.

<sup>51</sup> Similar to Compton scattering, but occurs when the photon's wavelength is much larger than the size of the atom. Rayleigh scattering can be neglected in most stellar atmospheres, but it is important in the UV for the extended envelopes of supergiants, and in cool MS stars. It is also responsible for the interstellar reddening.

TABLE 7  
HARVARD SPECTRAL CLASSIFICATION OF MS STARS

Type	$T_{\text{eff}}$ [K]	$L$ [ $L_{\odot}$ ]	Mass [ $M_{\odot}$ ]	$R$ [ $R_{\odot}$ ]	Apparent Colour	Features
O	$\gtrsim 33,000$	$\gtrsim 10^5$	$\gtrsim 20$	$\gtrsim 10$	Blue	Strong He II absorption lines
B	10,000–33,000	50– $10^5$	3–20	2–10	Blue white	Strong He I absorption lines
A	7,500–10,000	5–50	2–3	1.5–2	White	Strong Balmer absorption lines
F	6,000–7,500	1.5–5	1.05–2	1.1–1.5	Yellow-white	Neutral metal absorption lines (Fe I and Cr I)
G	5,000–6,000	0.6–1.5	0.8–1.05	0.9–1.1	Yellow	Ca II and other neutral metal lines becoming stronger
K	4,000–5,000	0.1–0.6	0.5–0.8	0.7–0.9	Yellow-orange	Ca II H and K lines strongest
M	$\lesssim 4,000$	$\lesssim 0.1$	$\lesssim 0.5$	$\lesssim 0.7$	Orange-red	Dominated by molecular absorption bands, especially TiO and VO

#### EMISSION LINES

Since the stellar atmosphere is optically thick below the base of the photosphere we know that emission lines must be originating elsewhere. In particular, they must originate at higher altitudes and, from Kirchhoff's second law, a hot, low-density gas must be present (Carroll & Ostlie 2007, pg. 365). This occurs, for example, in the chromosphere and corona of the Sun. The chromosphere is particularly strong in  $H\alpha$  emission while the high temperature and low density of the corona allows for the production of strong X-ray emission lines as well as a variety of forbidden emission lines (Carroll & Ostlie 2007, pg. 368).

#### SPECTRAL CLASSIFICATION

The study of stellar spectra allows for the classification of stellar spectral types in the familiar Harvard classification scheme (OBAFGKM). In this scheme, stars are ranked in terms of their effective temperatures where O stars are the hottest and M stars are the coolest. In addition, each class is further subdivided into ten parts (e.g., A0-A9) where a larger number denotes a smaller temperature (Carroll & Ostlie 2007, pg. 204). The general characteristics of each spectral type are outlined in Table 7 using data from Carroll & Ostlie (2007).

The distinctions between the spectra of stars with different temperatures are due to electrons occupying different atomic orbitals in the atmospheres of these stars. In order to understand which orbitals will be occupied at a given temperature we must learn some statistical mechanics. In particular, the Boltzmann equation, states that at temperature  $T$  we have

$$\frac{N_b}{N_a} = \frac{g_b}{g_a} e^{-(E_b - E_a)/kT}, \quad (266)$$

where  $N_a$  and  $N_b$  are the number of atoms with electrons in states with energies  $E_a$  and  $E_b$ , respectively, with corresponding degeneracies  $g_a$  and  $g_b$  (Carroll & Ostlie 2007, pg. 212). The form of the Boltzmann equation is easy to understand: for a thermal distribution of particles with energies on the order of  $kT$ , higher-energy states are increasingly less energetically accessible, with a probability of occupying them dropping off exponentially in energy; the degeneracies enter simply as statistical weights. If equation (266) is used to determine the temperature at which a gas of neutral hydrogen atoms has equal occupancy in the ground ( $n = 1$ ) and first excited ( $n = 2$ ) state we arrive at  $T = 8.5 \times 10^4$  K. However, we find that **Balmer absorption lines** (transition upward from  $n = 2$  orbital) are strongest in spectral type A0 with a temperate of 9520 K and diminish in strength at higher and lower temperatures (Carroll & Ostlie 2007, pg. 213). Clearly, this is inconsistent with the Boltzmann equation, which seems to argue that Balmer absorption lines should become increasingly more prominent as the temperature is raised.

The solution to this problem is to consider the relative number of atoms in different stages of ionizations. In particular, we must invoke the **Saha equation**, which states that

$$\frac{N_{i+1}}{N_i} = \frac{2Z_{i+1}}{n_e Z_i} \left( \frac{2\pi m_e kT}{h^2} \right)^{3/2} e^{-\chi_i/kT}, \quad (267)$$

where  $N_i$  and  $N_{i+1}$  are the number of atoms in ionization state  $i$  and  $(i+1)$  (for example, H I and H II),  $n_e$  is the number density of free electrons with mass  $m_e$ , and  $\chi_i$  is the energy required to lift the atom form ionization state  $i$  to  $(i+1)$ . The quantities  $Z_i$  and  $Z_{i+1}$  are called partition functions, which are simply the degeneracy-weighted sums of the number of ways in which an atom can arrange its electrons with the same energy. Partition functions are necessary because the initial and final ions will not necessarily be in their respective ground states. The partition functions are simply

$$Z = \sum_{j=1}^{\infty} g_j e^{-(E_j - E_1)/kT}, \quad (268)$$

where  $g_j$  is the degeneracy of the state with energy  $E_j$  and  $E_1$  is the energy of the ground state. Note the presence of  $n_e$  in the equation (267). The reason for this is that as  $n_e$  increases, the number of atoms in the higher ionization state  $N_{i+1}$  decreases since there are more electrons available for recombination to take place. The factor of 2 in front of  $Z_{i+1}$  reflects the two possible spins of the free electron while the term in parenthesis is roughly equal to the number density of electrons for which the quantum energy is roughly equal to the characteristic thermal energy  $kT$  (Carroll & Ostlie 2007, pg. 214).

We can now combine the Boltzmann and Saha equations to determine the temperature dependence of absorption line strength. Returning to our example of Balmer absorption lines of hydrogen, equation (267) allows us to calculate the ionized proportion,  $N_{\text{II}}/N_{\text{total}}$ , of hydrogen as a function of  $T$ . In working out the details of the partition functions ( $Z_{\text{II}} = 1$  because an ionized

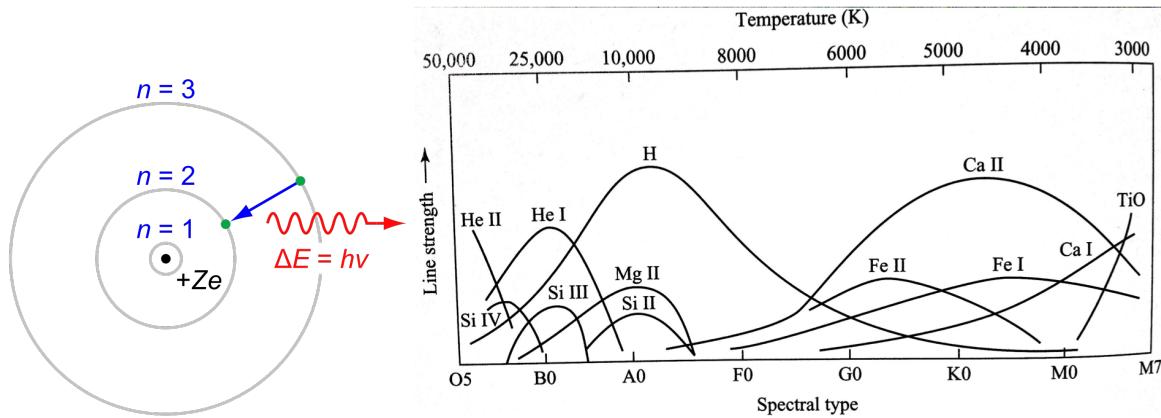


FIG. 77.— (left) A reminder of the Bohr model of the hydrogen atom with the ground state ( $n = 1$ ) and first two excited states ( $n = 2$  and  $3$ ) drawn. Balmer transitions occur when an electron jumps between the first excited state ( $n = 2$ ) and higher-energy levels. For example, the transition from  $n = 3$  to  $n = 2$  produces an  $H\alpha$  photon, as shown in the image. In contrast, Ly $\alpha$  emission arises when the electron makes the transition from  $n = 2$  downwards to  $n = 1$ . Image taken from <http://en.wikipedia.org/wiki/H-alpha> (right) The dependence of various absorption line strengths on temperature and spectral type. Image taken from Carroll & Ostlie (2007).

hydrogen atom is simply a proton, and has no degeneracy;  $Z_1 \approx g_1 = 2$  because the energy of the first excited state is 10.2 eV  $\gg kT$  for the temperature regime under consideration, so that the Boltzmann factor  $e^{-(E_2 - E_1)/kT} \ll 1$ , implying that nearly all the H I atoms are in the ground state.) we find that at 8300 K about 5% of hydrogen is ionized, about half is ionized at 9600 K, and all but 5% is ionized at 11300 K. This explains why Balmer lines are most prominent around 9520 K and diminish at higher and lower temperatures. At higher temperatures the rapid ionization of hydrogen prevents the electron absorption process and at lower temperatures there is insufficient thermal energy to raise many hydrogen atoms to the first excited state.

Of course, stellar atmospheres are not composed of pure hydrogen, and our discussion above depends on the appropriate value for the electron number density. In real stars, the presence of ionized helium provides more electrons with which the hydrogen ions can recombine. Thus, when helium is added, it takes a *higher* temperature to achieve the same degree of hydrogen ionization. Furthermore, it should be emphasized that the Saha equation can only be applied to gas in thermodynamic equilibrium so that the Maxwell-Boltzmann velocity distribution is obeyed. In addition, the density of gas must not be too great, or the presence of neighbouring ions will distort an atom's orbitals and lower its ionization energy (Carroll & Ostlie 2007, pg. 217). Nevertheless, in Figure 77 we show how the strength of various spectral lines varies with spectral type and effective temperature. As the temperature changes, a smooth variation from one spectral type to the next occurs, indicating that there are only minor differences in the composition of stars, as inferred from their spectra.

#### ADDITIONAL INFORMATION

When interpreting stellar spectra it is necessary to understand that the spectrum is formed only in the surface layers of a star, called the *stellar photosphere*, although in cool stars the chromosphere and corona also contribute to the emergent spectrum, especially in the FUV and EUV. The radiation emerging from the photosphere represent a collection of photons that have been reprocessed on their outward journey through the star. The interior plasma of a star interacts with the radiation field through the physical processes of electron scattering and free-free and bound-free (photoionization) absorption and emission by ions. Free-free absorption occurs when a free electron becomes able to absorb a photon in the vicinity of an ion. Thus a  $\gamma$ -ray photon produced in the core of the star random walks its way to the surface and, in the process, is degraded into hundreds of lower energy photons. The high densities within the interior mean that the mean free path of photons is small and that collisions are extremely effective in coupling the radiation field with the thermal state of the gas. In this condition, the interior is very nearly in a perfect state of thermodynamic equilibrium and, as a consequence, the material in the interior of the star radiates at the local temperature of a blackbody radiator. When we approach the surface the mean free path of photons increases drastically to the point that only the final few interactions with stellar material are important in the formation of the emerged spectrum (Gray & Corbally 2009, pg. 50).

The physical processes of electron scattering and free-free and bound-free absorption by ions are referred collectively as sources of *continuous opacity*, as they can scatter or absorb and emit photons over a wide range of wavelengths. In the stellar atmosphere other forms of opacity come into play. In particular, bound-bound absorption, or *line opacity*, is important. As a consequence, opacity in the stellar photosphere is a strong function of wavelength; in the core of a spectral line the opacity is considerably higher than in the surrounding continuum, where the opacity is due only to the continuum processes. As we already know, the temperature gradient in the stellar photosphere is responsible for the emergence of absorption lines. In the continuum region around an absorption line the total opacity is relatively low, as only continuum processes contribute. This means that we can see at those wavelengths deeply into the photosphere to relatively hot layers and thus the emerging radiation flux is high. In the core of a spectral line the opacity is high, and thus the majority of the photons emerge from higher, cooler layers, leading to a lower radiation flux (Gray & Corbally 2009, pg. 51).

**QUESTION 4**

**Why do some stars pulsate while some others do not? Consider Cepheids as an example.**

#### QUESTION 4

Why do some stars pulsate while some others do not? Consider Cepheids as an example.

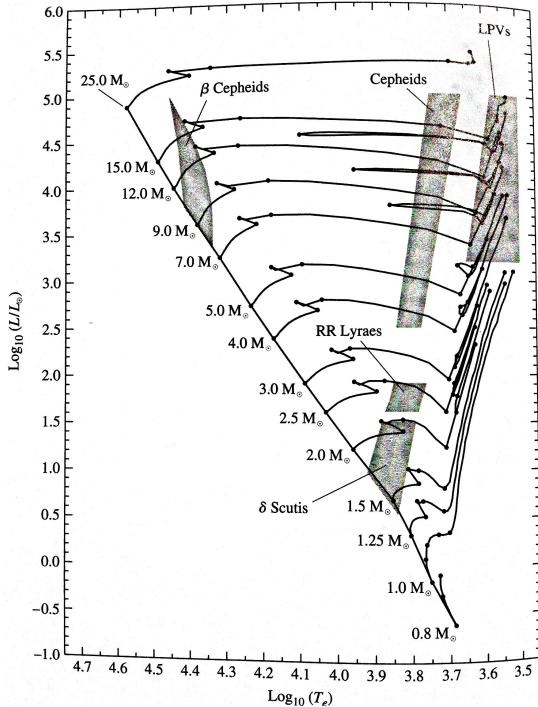


FIG. 78.— Image taken from Carroll & Ostlie (2007).

**density relation** explains why the pulsation period decreases as we move down the instability strip from the very tenuous supergiants to the very dense WDs<sup>52</sup>. It also explains the tight **period-luminosity relation** (the longer the period, the higher the luminosity) observed for the different types of pulsating stars. In particular, the period-luminosity relations exist because the instability strip is roughly parallel to the luminosity axis of the HR diagram, with small deviations in the relation related to the finite width of the instability strip (Carroll & Ostlie 2007, pg. 492).

Note that the timescale of pulsations is often referred to as the **dynamical timescale**,

$$\tau_{\text{dyn}} \approx \frac{1}{\sqrt{G\rho}}, \quad (271)$$

which is a measure of the e-folding time for changes in radius as the star makes dynamic adjustments in structure; for the Sun  $\tau_{\text{dyn}}$  is about an hour (Hansen et al. 2004, pg. 13).

#### DYNAMICAL INSTABILITY

We begin our investigation into how transient stellar pulsation arises by discussing **dynamical instability**. This concept refers to whether or not a star can remain stable to a density perturbation resulting from something like compression. We know from equation (261) that in hydrostatic equilibrium  $P \propto R^{-4}$ . Moreover, for a polytrope equation of state we have that  $P \propto \rho^\gamma \propto R^{-3\gamma}$ . Hence, in order for a star to be stable against contraction, we must have that  $\gamma \geq 4/3$ , since otherwise thermal pressure cannot rise quickly enough to maintain hydrostatic equilibrium, and the star will collapse. For  $\gamma > 4/3$ , the pressure increases enough that it exerts a net force outwards that expands the overlaying mass until hydrodynamic equilibrium is restored. Thus, we have the requirement that  $\gamma \geq 4/3$  in order to satisfy dynamic stability. If  $\gamma$  is not constant within a star, for instance because of ionization which drives  $\gamma \rightarrow 1$ , then marginal stability occurs if a suitably averaged  $\gamma_{\text{mean}}$  over the star reaches the critical value of  $4/3$  (Kippenhahn & Weigert 1990, pg. 241).

Note that the critical value of  $4/3$  depends strongly on spherical symmetry and Newtonian gravitation. The 4 in the numerator comes from the fact that the weight of the envelope in Newtonian mechanics varies as  $r^{-2}$  and has to be distributed over the surface of the sphere, introducing another  $r^{-2}$ . The 3 in the denominator comes from the  $r^3$  in the formula for the volume of a sphere. Therefore, effects of general relativity change the critical value of  $\gamma$  and make models less stable. Moreover, rotation causes the

<sup>52</sup> This relation actually breaks down at WDs since they exhibit non-radial oscillations with longer periods than those predicted by equation (270).

Cepheid variables are an example of a general class of stellar objects, known as **pulsating stars**, that periodically dim and brighten as their surfaces rhythmically expand and contract. Stellar pulsation is actually a transient phenomenon, with the positions of pulsating variables on the HR diagram confirming this. Rather than being located on the MS, where stars spend most of their lives, the majority of pulsating stars occupy a narrow ( $\sim 1000$  K wide), nearly vertical **instability strip** on the right-hand side of the HR diagram. Figure 78 shows the location of pulsating stars on the HR diagram with the instability strip marked by the location of Cepheids, RR Lyraes, and  $\delta$  Scutis. Theoretical evolutionary tracks for stars of various masses are also shown in this diagram. As stars evolve along these tracks, they begin to pulsate as they enter the instability strip and cease their oscillations upon leaving (Carroll & Ostlie 2007, pg. 489). Table 8 summarizes the properties of each of the classes of pulsating stars.

The radial oscillations of a pulsating star are the result of sound waves resonating in the star's interior. A rough estimate of the pulsation period,  $\Pi$ , may be obtained by considering how long it would take a sound wave to cross the diameter of a model star of radius  $R$  and constant density  $\rho$ . The adiabatic sound speed is

$$v_s = \sqrt{\frac{\gamma P}{\rho}}, \quad (269)$$

where  $P$  can be approximated by integrating equation (255) for hydrostatic equilibrium assuming constant density. From this it is easy to see that

$$\Pi \approx \sqrt{\frac{3\pi}{2\gamma G\rho}}. \quad (270)$$

Qualitatively, this shows us that the pulsation period of a star is inversely proportional to the square root of its mean density. This **period-mean-density relation** explains why the pulsation period decreases as we move down the instability strip from the very tenuous supergiants to the very dense WDs<sup>52</sup>.

It also explains the tight **period-luminosity relation** (the longer the period, the higher the luminosity) observed for the different types of pulsating stars. In particular, the period-luminosity relations exist because the instability strip is roughly parallel to the luminosity axis of the HR diagram, with small deviations in the relation related to the finite width of the instability strip (Carroll & Ostlie 2007, pg. 492).

TABLE 8  
PULSATING STARS

Type	Range of Periods	Pop.	(Non-)Radial	Notes
Long-Period Variables	100-700 days	I, II	R	RG or AGB stars
Cepheids	1-50 days	I	R	Produced by instability in partial ionization zones
W Virginis	2-45 days	II	R	Metal-deficient Cepheids with smaller luminosities
RR Lyrae	1.5-24 hrs	II	R	HB stars found in globular clusters. Used as standard candles
$\delta$ Scuti	1-3 hrs	I	R, NR	Evolved F stars found near the MS
$\beta$ Cephei	3-7 hrs	I	R, NR	Similar to Cepheids, but due to iron partial ionization zone
ZZ Ceti	100-1000 s	I	NR	Pulsating WDs on the instability strip

star to deviate from spherical symmetry, causing the critical value of  $\gamma$  to increase, making models more stable (Kippenhahn & Weigert 1990, pg. 241).

#### CEPHEID VARIABLES

We now consider Cepheid variables as a case study for stellar pulsation. Simply stated, pulsations in Cepheid variables result when the partial ionization zones of hydrogen and helium drive  $\gamma_{\text{mean}}$  below the critical value of  $4/3$  in the star. We can obtain a better understanding of this process by discussing a generalized model of how this works. In particular, if a layer of a star becomes more opaque upon compression, it could “dam up” the energy flowing toward the surface and push the surface layers upward. Then, as this expanding layer becomes more transparent, the trapped heat can escape and the layer will fall back down to repeat the cycle. In most regions of a star, however, the opacity actually decrease with compression. Recall that Kramer’s law states that the opacity  $\kappa$  depends on the density and temperature of the stellar material as  $\kappa \propto \rho/T^{7/2}$ . As the layers of a star are compressed, their density and temperature both increase. But because the opacity is more sensitive to the temperature, the opacity usually decreases upon compression (Carroll & Ostlie 2007, pg. 496).

We therefore require special circumstances to overcome the damping effect of most stellar layers, which explains why stellar pulsation is observed for only one of every  $10^5$  stars. One of these special circumstances deals with the **partial ionization zones** within a stellar interior. In these layers of the star where the gases are partially ionized, part of the work done on the gases as they are compressed produces further ionization rather than raising the temperature of the gases; in this case,  $\gamma \rightarrow 1$ . With a smaller temperature rise, the incase in density with compression produces a corresponding increase in opacity within this region. Similarly, during expansion, the temperature does not decease as much as expected since the ions now recombine with electrons and release energy. Again, the density term in Kramer’s law dominates, and the opacity decreases with decreasing density during expansion. This layer of the star can thus absorb heat during compression, be pushed outward to release the heat during expansion, and fall back down again to begin another cycle. In this way, partial ionization zones are the pistons that drive the oscillations of stars; they modulate the flow of energy through the layers of the star and are the direct cause of stellar pulsation (Carroll & Ostlie 2007, pg. 497). This process is known as the  **$\kappa$ -mechanism** since it is the modulated absorption of radiation that yields the **vibrational instability** of the star. For this case, the instability occurs if during adiabatic compression the absorption coefficient increases. Then in the compressed state more energy is absorbed than in equilibrium and the ensuing expansion is slightly enhanced. For analogous reasons the state of maximum expansion is followed by an enhanced compression. For stars, the pulsation takes place on the hydrostatic timescale, which is shorted compared to the Kelvin-Helmholtz timescale, so deviations from adiabaticity should be small in most parts of the stellar interior (Kippenhahn & Weigert 1990, pg. 408).

In most stars there are two main ionization zones. The first is a broad zone where the ionization of neutral hydrogen ( $H I \rightarrow H II$ ) occurs first and first ionization of helium ( $He I \rightarrow He II$ ) follows immediately in layers with a characteristic temperature of  $\sim 10^4$  K. Those layers are collectively referred to as the **hydrogen partial ionization zone**. The second, deeper zone involves the second ionization of helium ( $He II \rightarrow He III$ ), which occurs at  $T \sim 4 \times 10^4$  K, and is called the **He II partial ionization zone** (Carroll & Ostlie 2007, pg. 497). The exact location of these zones can be determined by considering the Saha equation, given by equation (267), which is basically the Boltzmann equation extended to ionization. Of course, there are a couple of fundamental differences between excitation and ionization. In particular, excitation concerns ions and bound electrons distributed over discrete states only. In the case of ionization, the upper state consists of two separate particles, the ion and the free electron, where the latter possesses a continuous distribution in kinetic energy; this is the primary reason behind using the Saha equation. Note that the degree of ionization increases with temperature, but decreases with pressure. This is easily understood: with increasing temperature the collisions become more violent, the photons more energetic, and successful ionization more frequent. On the other hand, at constant temperature, an increase in pressure means an enhanced probability of ions and electrons recombining (Kippenhahn & Weigert 1990, pg. 111). Obviously, the determination of the ionization zones is very sensitive to the properties of the stellar interior, and must be determined numerically.

In order for excitation of the  $\kappa$ -mechanism to win over damping from Kramer’s law, it is necessary that the ionization zones contain a sufficient part of the mass of the star. This means that these zones must be situated at suitable depths, and since ionization is mainly a function of temperature, we conclude that it is essentially a question of the surface temperature that decides whether a star is vibrationally stable or unstable via the  $\kappa$ -mechanism (Kippenhahn & Weigert 1990, pg. 413). For example, if the star is too hot ( $\sim 7500$  K), the ionization zones will be located very near the surface. At this position, the density is quite low, and there is not enough mass available to drive the oscillations effectively. This accounts for the hot blue edge of the instability strip on the HR diagram. In a cooler star ( $\sim 6500$  K), the characteristic temperatures of the ionization zones are found deeper in the star’s interior where there is enough mass for the ionization zone “piston” to push around, and oscillations can be excited.

However, if the surface temperature is too low ( $\sim 5500$  K), the onset of efficient convection in its outer layers may dampen the oscillations. Because the transport of energy by convection is more effective when the star is compressed, the convecting stellar material may lose heat at minimum radius. This could overcome the damming up of heat by the ionization zones – and so quench the pulsation of the star. Hence, the cool red edge of the instability strip is the result of the damping effects of convection (Carroll & Ostlie 2007, pg. 498).

Detailed numerical calculations show that it is the He II partial ionization zone that is primarily responsible for driving the oscillations of stars within the instability strip. The hydrogen ionization zone plays a more subtle role. As a star pulsates, the hydrogen ionization zone moves toward or away from the surface as the zone expands and contracts in response to the chaining temperature of the stellar gases. It happens that the star is brightest when the *least mass* lies between the hydrogen ionization zone and the surface. The luminosity incident on the bottom of the hydrogen ionization zone is indeed a maximum at minimum radius, but this merely propels the zone outward (through mass) most rapidly at that instant. The emergent luminosity is thus greatest after minimum radius, when the zone is nearest the surface. This delaying action of the hydrogen partial ionization zone produces the phase lag observed for classical Cepheids and RR Lyrae stars. Thus, the star is brightest when its surface is expanding outward most rapidly, after it has passed through its minimum radius (Carroll & Ostlie 2007, pg. 498).

#### OTHER PULSATING STARS

All stellar models evolving into the instability strip will become vibrationally unstable via the  $\kappa$ -mechanism and start to pulsate. For  $M \sim 5 - 20 M_{\odot}$ , stars undergoing the phase of helium burning loop away from and back to the Hayashi line, thereby passing through the instability strip at least twice; these correspond to the Cepheid variables we have just discussed. Of much smaller masses are the helium-burning stars located on the HBs of the HR diagrams of globular clusters. Where these branches intersect the downward continuation of the instability strip, we find the RR Lyrae variables. Like the classical Cepheids these are pulsating stars driven by the  $\kappa$ -mechanism. It seems, however, that some of them oscillate in the first overtone, as opposed to Cepheids which pulsate in the fundamental mode (see Figure 14.9 of Carroll & Ostlie (2007)). Even further down in the HR diagram, in the region of the MS, the instability strip is marked by another group of pulsating stars known as  $\delta$  Scuti stars or dwarf Cepheids. Above the location of the RR Lyrae stars in the HR diagram of globular clusters we sometimes find W Virginis stars. In contrast to classical Cepheids, which belong to Pop I, these stars are of Pop II. It is not surprising that they do not obey the same period-luminosity relation as Cepheids. These stars are apparently low-mass stars in an evolutionary stage later than the HB (Kippenhahn & Weigert 1990, pg. 415). The  $\beta$  Cepheis stars are B stars with effective temperatures around 20,000 K and where hydrogen is completely ionized, and the helium ionization zone is too near the surface to effectively drive pulsations in these stars. However, these stars are still being driven by the  $\kappa$ -mechanism, except that the element responsible is iron. Although the abundance of iron is low in all stars, it still contributes significantly to stellar opacities at temperatures around  $10^5$  K. The depth of this iron ionization zone is sufficient to produce net positive pulsational driving in these stars (Carroll & Ostlie 2007, pg. 499). Pulsating WDs with surfaces temperatures  $\sim 10^4$  K lie within the instability strip and are known as ZZ Ceti variables. The pulsations correspond to non-radial g-modes that resonate within the WD's surface layers of hydrogen and helium. Because these g-modes involve almost perfectly horizontal displacements, the radii of those compact stars hardly change. Their brightness variations (typically a few tenths of a magnitude) are due to temperature variations on their surfaces. For ZZ Ceti stars, it is actually the hydrogen partial ionization zone that is responsible for driving these oscillations (Carroll & Ostlie 2007, pg. 562).

The mechanisms responsible for the pulsation of stars outside the instability strip are not always as well understood. The long-period variables are red supergiants (AGB stars) with huge, diffuse convective envelopes surrounding a compact core. Their spectra are dominated by molecular absorption lines and emission lines that reveal the existence of atmospheric shock waves and significant mass loss. While we understand that the *hydrogen* ionization zone drives the pulsation of a long-period variable star, many details remain to be explained, such as how its oscillations interact with its outer atmosphere (Carroll & Ostlie 2007, pg. 498).

#### NONRADIAL STELLAR PULSATION

As some types of stars pulsate, their surfaces do no move uniformly in and out in a simple “breathing” motion. Instead, such a star executes a more complicated type of non-radial motion in which some regions of its surface expand while other areas contract. As we have just learned, the radial pulsation of stars is attributed to standing sound waves in the stellar interior. For the case of non-radial oscillations, the sound waves can propagate horizontally as well as radially to produce waves that travel around the star. Because pressure provides the restoring force for sounds waves, these non-radial oscillations are called **p-modes**. Another class of non-radial oscillations are known as **g-modes** since here *gravity* provides the restoring force. These waves involve a “sloshing” back and forth of stellar gases, which is ultimately connected to the buoyancy of stellar material. Because “sloshing” cannot occur for purely radial motion, there are no radial analogs for the g-modes (Carroll & Ostlie 2007, pg. 506).

All of the ideas of non-radial pulsation come into play in the science of helioseismology, the study of the oscillations of the Sun. A typical solar oscillation mode has very low amplitude and results in a luminosity variation of only one part in  $10^6$ . With an incoherent superposition of roughly  $10^7$  modes rippling through the surface and interior, the Sun is “ringing” like a bell. The oscillations observed on the Sun have modes with periods between 3 and 8 minutes and very short horizontal wavelengths. These so-called **five-minute oscillations** have been identified as p-modes that are concentrated below the Sun’s photosphere within the convection zone. Of course, the Sun is not a normal pulsating star. It lies far beyond the red edge of the instability strip on the HR diagram where turbulent convection overcomes the tendency of the ionization zones to absorb heat at maximum compression. Thus, the  $\kappa$ -mechanism cannot be responsible for the solar oscillations. However, the timescale for convection near the top of the convection zone is a few minutes, and it is strongly suspected that the p-modes are driven by tapping into the turbulent energy of the convection zone itself, where the p-modes are confined. It is also suspected that g-modes are present within the Sun, but

these only exist within the deep radiation zone. Since they dwell below the convection zone, their amplitudes are significantly diminished at the Sun's surface, explaining why none have been identified to date (Carroll & Ostlie 2007, pg. 512).

**QUESTION 5**

**Define the terms “thermal equilibrium” and “hydrostatic equilibrium”. How do they apply to stars? If a solar-type star is out of dynamical equilibrium, how long does it take to restore it? What is the time scale to restore thermal equilibrium?**

### QUESTION 5

**Define the terms “thermal equilibrium” and “hydrostatic equilibrium”. How do they apply to stars? If a solar-type star is out of dynamical equilibrium, how long does it take to restore it? What is the time scale to restore thermal equilibrium?**

#### *HYDROSTATIC EQUILIBRIUM*

Most stars are in long-lasting phases of their evolution that no changes can be observed at all. Then the stellar matter cannot be accelerated noticeably, which means that all forces acting on a given mass element of the star compensate each other. This mechanical equilibrium in a star is called **hydrostatic equilibrium**, since the same condition also governs the pressure stratification, say, in a basin of water. Assuming that we are dealing with gaseous stars in the absence of strong rotation, magnetic fields, or close companions, the only forces within the star are due to gravity and to the pressure gradient (Kippenhahn & Weigert 1990, pg. 6).

For a given moment of time, we consider a thin spherical mass shell with thickness  $dr$  at a radius  $r$  within the star. Per unit area of the shell, the mass is  $\rho dr$  and the weight of the shell is  $-g\rho dr$ . This weight is the gravitational force acting towards the centre (as indicated by the minus sign). In order to prevent the mass elements of the shell from being accelerated in this direction, they must experience a net force due to pressure of the same absolute value, but acting outwards. This means that the shell must feel a larger pressure at its interior (lower) boundary than the pressure at its outer (upper) boundary. The sum of the forces arising from pressure and gravity has to be zero, implying that

$$\frac{dP}{dr} = -\rho g \rightarrow \frac{dP}{dr} = -\frac{Gm}{r^2} \rho. \quad (272)$$

We can rewrite equation (272) by taking  $m$  to be the independent variable (convenient for spherical symmetry), for which we derive

$$\frac{dP}{dm} = -\frac{Gm}{4\pi r^4}, \quad (273)$$

using the identity that  $dm = 4\pi r^2 \rho dr$  (Kippenhahn & Weigert 1990, pg. 7).

The equation for hydrostatic equilibrium is a special case of conservation of momentum. If the (spherical) star undergoes accelerated radial motions, we have to consider the inertia of the mass elements, which introduces an additional term to equation (273). In particular, we know that the force per unit area acting on an infinitesimally thin shell due to the pressure gradient and gravity are

$$f_P = -\frac{dP}{dm} dm \quad \text{and} \quad f_G = -\frac{gdm}{4\pi r^2} = -\frac{Gm}{r^2} \frac{dm}{4\pi r^2}, \quad (274)$$

respectively. If the sum of the two forces is not equal to zero, the mass shell will be accelerated according to

$$\frac{dm}{4\pi r^2} \frac{d^2r}{dt^2} = f_P + f_G \rightarrow \frac{1}{4\pi r^2} \frac{d^2r}{dt^2} = -\frac{dP}{dm} - \frac{Gm}{4\pi r^4}. \quad (275)$$

The signs in equation (275) are such that the pressure gradient alone would produce an outward acceleration, while the gravity alone would produce an inward acceleration (Kippenhahn & Weigert 1990, pg. 10).

We can use this result to model the derivation of a star from hydrostatic equilibrium. For example, we can consider what would happen if the pressure term in equation (275) were to suddenly disappear. In this case we define a characteristic timescale  $\tau_{ff}$  for the ensuing collapse of the star by setting

$$\left| \frac{d^2r}{dt^2} \right| = \frac{R}{\tau_{ff}^2} \quad \text{which from equation (275) leads to} \quad \tau_{ff} \approx \left( \frac{R}{g} \right)^{1/2}. \quad (276)$$

This is some kind of a mean value for the timescale of the free fall time over a distance of order  $R$  following the sudden disappearance of pressure. We can correspondingly determine a timescale  $\tau_{expl}$  for the explosion of our star for the case that gravity were to suddenly disappear:  $R/\tau_{expl}^2$ , where we have replaced  $dP/dr$  with  $P/R$  after writing  $4\pi r^2(dP/dm)$  as  $(dP/dr)/\rho$ . We then find that

$$\tau_{expl} \approx R \left( \frac{\rho}{P} \right)^{1/2}. \quad (277)$$

Since the term in brackets in equation (277) is of order the mean velocity of sound in the stellar interior, one can see that  $\tau_{expl}$  is of the order of the time a sound wave needs to travel from the centre to the surface (Kippenhahn & Weigert 1990, pg. 10).

If the star is near hydrostatic equilibrium then the two terms on the right side of equation (275) have about equal absolute value and  $\tau_{ff} \approx \tau_{expl}$ . We then call this the **hydrodynamical timescale**,  $\tau_{hydro}$ , since it gives the typical time after which a slightly perturbed star can again reach hydrostatic equilibrium. From equation (276) we find that

$$\tau_{hydro} \approx \left( \frac{R^3}{GM} \right)^{1/2} \approx \frac{1}{\sqrt{4G\rho}}. \quad (278)$$

In the case of the Sun we find the surprisingly small value  $\tau_{\text{hydro}} \approx 27$  minutes. Even in the case of a red giant,  $\tau_{\text{hydro}} \approx 18$  days, while for a WD  $\tau_{\text{hydro}} \approx 5$  seconds. In most phases of their life the stars change slowly on a timescale that is very long compared to  $\tau_{\text{hydro}}$ , supporting the idea that they are very close to hydrostatic equilibrium (Kippenhahn & Weigert 1990, pg. 11).

#### THERMAL EQUILIBRIUM

**Thermal equilibrium** is a condition in thermodynamics where a system has a uniform temperature throughout that does not change in time. In a system where different system elements can transfer thermal energy (e.g., by radiation), thermal equilibrium requires that the net transfer of thermal energy to any system element is zero (the energy may eventually be dumped into reservoirs external to the system). More loosely, thermal equilibrium may also refer to a time-invariant thermal distribution that is non-uniform in space (Charles). If we represent a star as an “ideal box” then the confined gas particles and blackbody radiation will come into equilibrium, individually and with each other, and can be described by a single well-defined temperature. However, this cannot be the case for a real star since a net outward flow of energy occurs through the star, and the temperature varies with location. Gas particles and photons at one position in the star may have arrived there from other regions, either hotter or cooler. The distribution in particle speeds and photon energies thus reflects a range of temperatures. Despite this, the idealized case of a single temperature can still be employed if the distance over which the temperature changes significantly is large compared with the mean free paths of interacting particles and photons. For this reason, stars are often modelled as being in **local thermodynamic equilibrium**<sup>53</sup> (LTE) whereby the particles and photons cannot escape the local environment and so are effectively confined to a limited volume of nearly constant temperature (Carroll & Ostlie 2007, pg. 239).

The energy a star radiates away so profusely from its surface is generally replenished from reservoirs situated in the very hot central region. This requires an effective transfer of energy through the stellar material, which is possible owing to the existence of a non-vanishing temperature gradient in the star. Depending on the local physical situation, the transfer can occur mainly via radiation, conduction, and convection. In any case, certain particles (photons, atoms, electrons) are exchanged between hotter and cooler parts, and their mean free path together with the temperature gradient will play a decisive role (Kippenhahn & Weigert 1990, pg. 27).

Rough estimates show important features of the radiative transfer in stellar interiors and justify an enormous simplification of the formalism. Let us first estimate the mean free path  $\lambda_{\text{ph}}$  of a photon at an average point inside a star like the Sun:  $\lambda_{\text{ph}} = 1/\kappa\rho$ , where  $\kappa \sim 1 \text{ cm}^2 \text{ g}^{-1}$  and  $\rho \sim 1.4 \text{ g cm}^{-3}$  are typical stellar values; then  $\lambda_{\text{ph}} \sim 2 \text{ cm}$  (i.e. stellar material is very opaque). The typical temperature gradient in the star can be roughly estimated by averaging between centre ( $T_c \approx 10^7 \text{ K}$ ) and surface ( $T_s \approx 10^4 \text{ K}$ ) values:  $dT/dr \approx (T_c - T_s)/R_\odot \approx 1 \text{ K cm}^{-1}$ . The radiation field at a given point is emitted from a small, nearly isothermal surrounding, with temperature differences being on the order of  $\Delta T = \lambda_{\text{ph}}dT/dr \approx 3 \times 10^{-4} \text{ K}$ . Since the energy density of radiation is  $U \sim T^4$ , the relative anisotropy of the radiation at a point with  $T = 10^7 \text{ K}$  is  $4\Delta T/T \sim 10^{-10}$ . The situation in stellar interiors must obviously be very close to thermal equilibrium, and the radiation very close to that of a blackbody. Nevertheless, the small remaining anisotropy can easily be a carrier of a star’s huge luminosity: this fraction of  $10^{-10}$  near the centre is still roughly  $10^3$  times larger than the flux at the stellar surface. Radiative transfer of energy occurs via the the non-vanishing net flux; that is, via the surplus of outwards-going radiation (emitted from somewhat hotter material below) over the inwards-going radiation (emitted from cooler material above) (Kippenhahn & Weigert 1990, pg. 28).

The above estimates have shown that for radiative transport in stars, the mean free path of the transporting photons is very small compared to the stellar radius over which the transport extends. For this reason, the transport can be treated as a diffusion process, for which the diffusive flux  $F$  of radiative energy is given by

$$F = -\frac{1}{3}c\lambda_{\text{ph}}\nabla U \rightarrow F = -\frac{4ac}{3}\frac{T^3}{\kappa\rho}\frac{dT}{dr} \quad (279)$$

where  $U$  is the energy density of radiation,  $U = aT^4$ , and its gradient simplifies to a radial derivative in the spherically symmetric case. Note that this can be interpreted formally as an equation for heat conduction:  $F = -k_{\text{rad}}\nabla T$  (Kippenhahn & Weigert 1990, pg. 29).

In heat conduction, energy transfer occurs via collisions during the random thermal motion of particles. A basic estimate similar to those done above show that in ordinary stellar material (i.e. non-degenerate gas) conduction has no chance of taking over an appreciable part of the total energy transport. Although the collisional cross-sections of charged particles is small at high temperatures ( $\approx 10^{-19} \text{ cm}^2$  per particle), the large density results in a mean free path several orders of magnitude less than those for photons; and their velocities are only a few percent of  $c$ . Therefore, the diffusion of particles is much less efficient than for photons<sup>54</sup> (Kippenhahn & Weigert 1990, pg. 31).

We now have enough information to determine the timescale for thermal adjustment within a star. It turns out that equation (279), which is valid for both radiative and conductive energy transport, can be written in the form of

$$\frac{d}{dx} \left( \sigma \frac{dT}{dx} \right) = c \frac{dT}{dt}. \quad (280)$$

This equation governs the variation in temperature along a rod with variable conductivity  $\sigma$  and specific heat  $c$ , where  $x$  is the spatial coordinate along the rod. One can easily estimate the timescale over which equation (280) demands considerable changes

<sup>53</sup> Note that thermodynamic (i.e. complete) equilibrium is a more stringent condition in that it requires both mechanical and thermal equilibrium (Kippenhahn & Weigert 1990, pg. 25).

<sup>54</sup> This situation can be inverted in the cores of evolved stars where the electron gas is highly degenerate. The enhanced densities and thermal speeds raise the mean free path considerably so that heat conduction can become so efficient that it short-circuits radiative transfer

of an initially given temperature profile, the **thermal adjustment timescale**,

$$\tau_{\text{adj}} = \frac{c}{\sigma} d^2, \quad (281)$$

where  $d$  is a characteristic length over which the temperature variation changes. For a star, a rough estimate of this timescale yields  $\tau_{\text{adj}} \approx \tau_{\text{KH}}$ . This means that the Kelvin-Helmholtz timescale can be considered a characteristic time of thermal adjustment of a star (i.e. the time it takes a thermal fluctuation to travel from the centre to the surface of the star). For the Sun,  $\tau_{\text{KH}} \approx 10^7$  years. In spite of this indicated equivalence, it is advisable to consider  $\tau_{\text{adj}}$  separately, in particular if it is applied to parts of a star only. For example, in evolved stars with isothermal cores of high conductivity, the luminosity there is zero so that formally  $\tau_{\text{KH}}$  tends to infinity, though the decisive timescale that enforces the isothermal situation is indeed the very small  $\tau_{\text{adj}}$  (Kippenhahn & Weigert 1990, pg. 34). The actual value of  $\tau_{\text{adj}}$ , of course, depends on the detailed physics of the thermal energy transport involved in the specific scenario.

#### VIRIAL THEOREM

Stars in hydrostatic and thermal equilibrium obey the virial theorem, which can be derived in the following manner (see Kippenhahn & Weigert 1990, pg. 15-17). If we multiply equation (273) by  $4\pi r^3$  and integrate over  $dm$  in the interval  $[0, M]$ , that is from centre to surface, we obtain on the left-hand side an integral which can be simplified by partial integration:

$$\int_0^M 4\pi r^3 \frac{dP}{dm} dm = [4\pi r^3 P]_0^M - \int_0^M 12\pi r^2 \frac{dr}{dm} P dm, \quad (282)$$

where the bracket term vanishes, since  $r = 0$  at the centre and  $P = 0$  at the surface, and the second integral is reduced to  $3P/\rho$ . Therefore, after multiplication of  $4\pi r^3$  and integration of equation (273) we arrive at

$$\int_0^M \frac{Gm}{r} dm = 3 \int_0^M \frac{P}{\rho} dm. \quad (283)$$

Both sides of equation (283) have dimensions of energy and are easily interpreted. The left side is the negative of the **gravitational energy**  $E_{\text{grav}}$ , integrated over all mass elements  $dm$  of the star. We see that  $E_{\text{grav}}$  varies if the configuration undergoes expansion or contraction: if all mass shells expand or contract simultaneously, then  $E_g$  increases or decreases, respectively. And the same must be true for the integral on the right-hand side of equation (283). Note that these radial motions must be slow compared to  $\tau_{\text{hydro}}$  in order for hydrostatic equilibrium to be maintained and equation (283) valid.

In order to understand the meaning of the right-hand side of equation (283) we assume an ideal gas for which  $P/\rho = (\gamma - 1)c_V T$ . If we define the internal energy,  $E_{\text{int}} = c_V T$ , then we arrive at the statement

$$E_{\text{grav}} = -\zeta E_{\text{int}}, \quad (284)$$

where  $\zeta \equiv 3(\gamma - 1)$ . Substituting  $\gamma = 5/3$  for a non-relativistic monoatomic gas yields the familiar  $E_{\text{grav}} = -2E_{\text{int}}$ .

In general, if the configuration slowly expands or contracts,  $E_{\text{grav}}$  and  $E_{\text{int}}$  will vary, the total energy will not remain constant, and the gas, which has a finite temperature, must radiate. If  $L$  is the luminosity of the star, then conservation of energy demands that

$$L = (\zeta - 1) \frac{dE_{\text{int}}}{dt} = -\frac{\zeta - 1}{\zeta} \frac{dE_{\text{grav}}}{dt}. \quad (285)$$

For our usual case this reduces to  $L = -\dot{E}_{\text{grav}}/2 = \dot{E}_{\text{int}}$ , which means that half the energy liberated by contraction (for example) is radiated away while the other half is used to heat the star ( $L > 0, \dot{E}_{\text{int}} > 0$ ). We have to keep in mind that it is the luminosity that causes the shrinking: a configuration in hydrostatic equilibrium has a finite temperature and therefore radiates into the (cold) universe. If we could prevent the star from radiating by illuminating it from all sides so that it absorbed as much energy as it lost, then it would not shrink.

**QUESTION 6**

**Define and describe Type Ia, Type Ib, Type Ic, and Type II supernovae.**

### QUESTION 6

**Define and describe Type Ia, Type Ib, Type Ic, and Type II supernovae.**

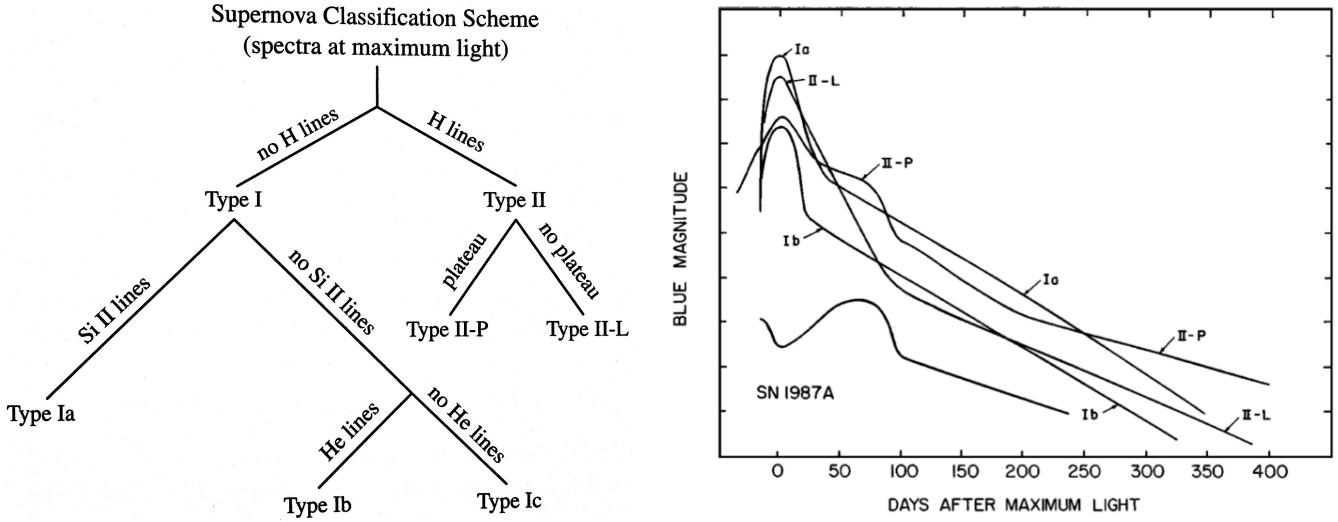


FIG. 79.— (left) The classification of supernovae based on their spectra at maximum light and the existence or absence of a plateau in the Type II light curve. Image taken from Carroll & Ostlie (2007). (right) Typical light curves for different SNe classifications along with that of SN 1987A. The curve for Ib includes Ic as well and represents an average. For Type II-L, SN 1979C was used, though this may be unusually luminous. The light curve of SN 1987A, although unusual, was generically related to those of Type II-P; the initial peak was low because the progenitor was a blue supergiant, much smaller than a red supergiant. Image taken from Filippenko (1997).

Historically supernovae have been classified based on their spectra with two main groups: **Type I** SNe which display an apparent lack of hydrogen in their spectra and **Type II** SNe that contain strong hydrogen lines. Type I SNe are further subdivided into three groups with those whose early-time spectra show strong Si II absorption at 6150Å (**Type Ia**), prominent He I (**Type Ib**), or neither Si II nor He I (**Type Ic**). Type II SNe are also further subdivided into two groups based on the appearance of their light curves. In particular, light curves showing a distinctive plateau in intensity for an extended period of time shortly after maximum brightness are called **Type II-P**, whereas those more closely resembling Type I light curves showing a linear dropoff in intensity after maximum brightness are called **Type II-L** (Filippenko 1997).

A schematic representation of SNe classification is shown in Figure 79 along with typical light curves of the various types. To first order the light curves of Type I SNe are all broadly similar, while Type II SNe exhibit strong dispersions in their properties. The typical peak brightness of a Type Ia is  $M_B = -18.4$ , while the light curves of Types Ib and Ic are fainter by 1.5 to 2 magnitudes in blue light, but are otherwise similar. All Type I SNe show similar rates of decline of their brightness after maximum, about 0.07 magnitude per day at 20 days. After 50 days, the rate of dimming slows and becomes constant, with Type Ia's declining 1.5 times faster than the others (Carroll & Ostlie 2007, pg. 527). The peak absolute magnitudes of Type II-P show a wide dispersion, almost certainly due to differences in the radii of the progenitor stars. Most Type II-L SNe, on the other hand, have a nearly uniform peak absolute magnitude, roughly 2.5 magnitude fainter than Type Ia, although a few exceptionally luminous ones have been observed. At late times ( $t \gtrsim 150$  days) the light curves of most Type II resemble each other, both in shape and absolute flux, with a decline rate close to that expected from the decay of  $\text{Co}^{56}$  to  $\text{Fe}^{56}$  (Filippenko 1997).

The lack of hydrogen lines in Type I SNe indicates that the stars involved have been stripped of their hydrogen envelopes. The differences in the spectral signatures between Type Ia and Types Ib and Ic indicate that different physical mechanisms are at work. This is reflected in the different environments observed for these outbursts, with these locations providing important clues to the nature of the different SNe types and to the masses of their progenitor stars. Type II, Ib, and Ic have never been observed in elliptical galaxies and only rarely in S0 galaxies. They are generally found near spiral arms and H II regions, implying that their progenitors must have started their lives as massive stars ( $M \gtrsim 8 M_\odot$ ). In contrast, Type Ia SNe have been found to occur in all types of galaxies, including ellipticals, and in spirals there is no strong preference for spiral arms. Based on this evidence, it is believed that the progenitors of Type Ia SNe are carbon-oxygen WDs that accrete matter from a companion star and undergo thermonuclear runaway. Although the WDs probably reach the Chandrasekhar limit prior to exploding, this is not yet certain. Type II SNe are thought to arise from evolved, massive progenitors that suffer core collapse (generally iron) and subsequently rebound, leaving behind a NS or BH. It is also believed that Type Ib and Ic operate in the same manner as Type II, except that their progenitors were stripped of their hydrogen (Type Ib and Ic) and possibly helium (Type Ic) envelopes prior to exploding, either via mass transfer to companion stars or through winds. Indeed, the progenitors of some Type II SNe seem to have only a low-mass skin of hydrogen and their spectra gradually evolve to resemble those of Type Ib (Filippenko 1997).

Figure 80 show the early-time and late-time spectra of the different SN types. At early times, the lines are broad owing to the high velocities of the ejecta, and most of them have P Cygni profiles formed by resonant scattering above the photosphere. As we have mentioned, Type Ia SNe are characterized by a deep absorption trough around 6150 Å produced by blueshifted Si II,

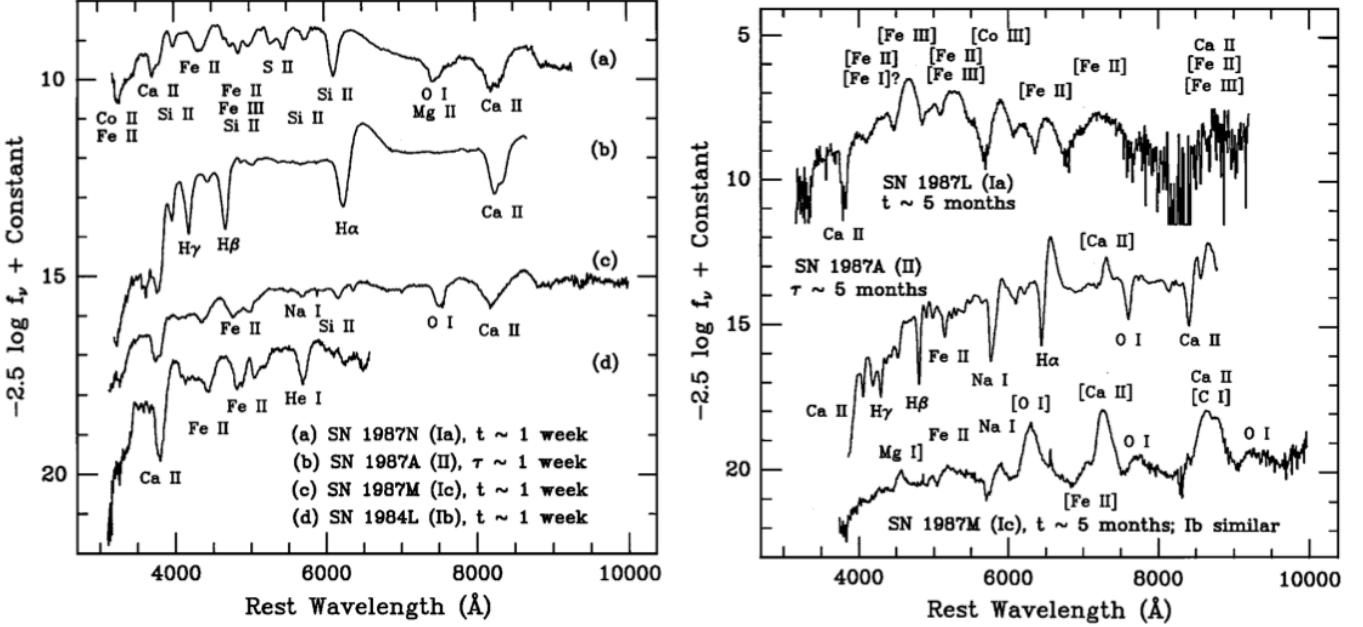


FIG. 80.— (left) Spectra of SNe, showing early-time distinctions between the four major types and subtypes. (right) Spectra of SNe, showing late-time distinctions between the four major types and subtypes. Images taken from Filippenko (1997).

while members of the Ib and Ic subclasses do not show this line. The presence of moderately strong He I lines distinguish Ib from Ic. At late times, Type Ia show blends of dozens of Fe emission lines, mixed with some Co lines. Type Ib and Ic, on the other hand, have relatively unblended emission lines of intermediate-mass elements such as O and Ca. At early times, the spectrum of a Type II SNe is nearly featureless and quite blue, indicating a high colour temperature ( $\gtrsim 10^4$  K). Initially, the widths of the Balmer lines and blueshifts of their P Cygni absorption minima decrease noticeably in some objects, as the photosphere quickly recedes to the inner, more slowly moving layers of the homologously expanding ejecta. Eventually, as the light curve drops to the late-time tail, the spectrum gradually takes on a nebular appearance; the continuum fades, but H $\alpha$  remains very strong and prominent emission lines of [O I] and [Ca II] appear (Filippenko 1997).

#### CORE-COLLAPSE SUPERNOVA

We begin by describing the physical processes involved in generating the magnificent **core-collapse supernovae** associated with Types II, Ib, and Ic. Note that a typical core-collapse SNe releases on the order of  $10^{46}$  J ( $10^{53}$  erg) of energy, with about 1% of that appearing as kinetic energy of the ejected material, less than 0.01% being released as photons, and the remainder carried away by neutrinos (Carroll & Ostlie 2007, pg. 529). The source of energy is provided by the release of gravitational potential energy during the collapse of the core. As we will describe below, the collapse of an iron core, followed by the generation of a shock wave and the ensuing ejection of the star's envelope, are believed to be the general mechanism that creates a core-collapse SNe. The details that result in a Type II rather than a Type Ib or Type Ic have to do with the composition and mass of the envelope at the time of core collapse and the amount of radioactive material synthesized in the ejecta. Type II SNe, which are more common than either Type Ib or Type Ic, are usually red supergiant stars in the extreme upper-right-hand corner of the HR diagram at the time they undergo core collapse. Type Ib's and Ic's have lost various amounts of their envelopes prior to detonation and are believed to be the products of exploded WR stars; probably WN and WC WR stars for Type 1b and 1c, respectively (Carroll & Ostlie 2007, pg. 534). Evidence for the general core-collapse scheme come from observations of SNR, such as the Crab Nebula, which is still expanding at almost  $1500 \text{ km s}^{-1}$  with a luminosity of  $10^5 L_\odot$ . Much of the radiation being emitted is in the form of highly polarized synchrotron radiation, indicating the presence of relativistic electrons that are spiralling around the magnetic field lines of a pulsar. Another source of evidence comes from neutrino observations of SN 1987A, which were detected some three hours before the photon signal. Finally, the chemical composition of the Sun's photosphere show enrichment in  $\alpha$ -elements (i.e. C, O, Ne, Mg, Si, Ar, Ca, Fe, Ni) that are produced by the successive addition of  $\alpha$  particles during the nuclear evolution of massive stars (Carroll & Ostlie 2007, pg. 537-542).

Massive stars evolve up to silicon burning within their cores, which, through a series of reactions, produces different isotopes of iron and nickel. Any further reactions that produce nuclei more massive than  $^{56}\text{Fe}$  are endothermic and cannot contribute the the luminosity of the star. Grouping all of the products together, silicon burning is said to produce an iron core, surrounded by onion-like layers from previous generations of nuclear burning. At the very high temperatures present within the iron core, the photons possess enough energy to destroy heavy nuclei (i.e. the reverse reactions of the silicon-burning sequence), a process known as **photodisintegration**. Particularly important are the photodisintegration of iron and helium:



Of course, this process of stripping iron down to individual protons and neutrons is highly endothermic; thermal energy is removed from the gas that would otherwise have resulted in the pressure necessary to support the core of the star. The core masses for which this process occurs vary from  $1.3 M_{\odot}$  for a  $10 M_{\odot}$  ZAMS star to  $2.5 M_{\odot}$  for a  $50 M_{\odot}$  star (Carroll & Ostlie 2007, pg. 531).

Under the extreme conditions that now exist ( $T \sim 10^{10}$  K and  $\rho \sim 10^{13}$  kg m $^{-3}$ ), the free electrons that had assisted in supporting the star through degeneracy pressure are captured by heavy nuclei and through inverse  $\beta$ -decay by the protons that were produced through photodisintegration; the latter produces a large flux of neutrinos that escape the star. With the reduction in electron degeneracy pressure the core suddenly begins to collapse. In the inner portion of the core, the collapse is homologous, and the velocity of the collapse is proportional to the distance away from the centre of the star. At the radius where the velocity exceeds the local sound speed, the collapse can no longer be homologous and the inner core decouples from the now supersonic outer core, which is left behind and nearly in free-fall. During the collapse, speeds can reach  $\sim 70,000$  km s $^{-1}$  in the outer core, and within about one second a volume the size of the Earth has been compressed to a radius of about 50 km (Carroll & Ostlie 2007, pg. 532).

Since mechanical information will propagate through the star only at the speed of sound and because the core collapses so quickly, there is not enough time for the outer layers to learn about what has happened inside; the outer layers are left in the precarious position of being almost suspended above the catastrophically collapsing core. The homologous collapse of the inner core continues until the density is roughly three times the density of an atomic nucleus. At this point, the inner core now stiffens due to neutron degeneracy pressure. The result is that the inner core rebounds somewhat, sending pressure waves outward into the infalling material from the outer core. When the velocity of the pressure waves reach the sound speed, they build into a shock wave that begins to move outward. The subsequent details are somewhat complicated as the shock wave can become damped due to photodisintegration when it encounters the infalling outer core. However, below the shock a **neutrinosphere** developed from the process of photodisintegration, and since the overlying material is so dense that it is optically thick to neutrinos, this can drive the shock upwards. This may continue until the shock reaches the surface of the star, resulting in the ejection of the stellar envelope, and producing a SNe with a peak photon luminosity of  $\sim 10^9 L_{\odot}$  (Carroll & Ostlie 2007, pg. 533).

### WHITE DWARF SUPERNOVA

Type Ia SNe are classified based on the absence of hydrogen lines in their spectra. Instead, the early-time spectra of Type Ia's are dominated by the strong presence of Si II lines, along with neutral and singly ionized lines of intermediate-mass elements like O, Mg, S, and Ca, with some contribution from iron-peak elements (Fe, Co). The spectral lines also show P-Cygni profiles, representative of mass loss. In addition, the blueshifted absorption features indicate expansion velocities of the ejecta of  $\gtrsim 10^4$  km s $^{-1}$  ( $\sim 0.1c$ ). At later times, the relative contribution of iron-group elements increases as the photosphere quickly recedes into the ejecta. After about 2 weeks the spectrum is dominated by lines of Fe II which is consistent with an iron-rich core, though some lines of intermediate-mass elements like Si II and Ca II are still present (Filippenko 1997; Carroll & Ostlie 2007, pg. 687)

Type Ia SNe are remarkable consistent in their energy output; at maximum light most reach an average maximum in the blue and visual wavelengths of  $\langle M_B \rangle \simeq \langle M_V \rangle \simeq -19.3$ , with a typical spread of less than 0.3 magnitudes. Also, a clear relationship exists between the peak brightness and the rate of decline in the light curve (the brightest Type Ia's decline the slowest), making it possible to accurately determine the maximum luminosity of an individual Type Ia by measuring the rate of decline. This feature, along with their large intrinsic luminosities, make Type Ia SNe useful standard candles in astronomy (Carroll & Ostlie 2007, pg. 686).

The absence of hydrogen lines in Type Ia SNe suggest that they represent a class of evolved objects that have either lost their hydrogen or had it converted to heavier elements. In addition, the remarkable consistency in their spectra indicate that a fairly uniform mechanism must be responsible for these extremely energetic events. The standard model is that these events safe due to the destruction of a WD in a binary system. If sufficient mass falls onto the WD, its mass can be driven near the Chandrasekhar limit, producing a catastrophic explosion. At present, however, it is still unclear of the exact mechanism that triggers the explosion. One proposed scenario is known as the **double-degenerate** model where two WDs in a binary orbit spiral into each other through the emission of GWs. At some point, the less massive WD (which has the larger radius) will eventually spill over its Roche lobe and be completely torn apart in just a few orbits. The resulting thick disk dumps its C-O rich material onto the more massive primary. As the mass of the primary grows and nears the Chandrasekhar limit, nuclear reactions begin in the deep interior, eventually destroying the primary WD. This model appears to predict the right number of mergers, consistent with the observed Type Ia SN rate in galaxies, and it naturally accounts for the lack of hydrogen in their spectra. However, computer simulations show that nuclear ignition is probably off-centre, resulting in ultimate collapse to a NS, rather than a SN. In addition, it appears that the production of heavy elements may be inconsistent with the relative abundances observed in SN spectra. The other scenario, known as the **single-degenerate** model, involves an evolving star orbiting a WD. In this case, the mass falling onto the WD from the companion star produces a SN, though the exact details of this are unclear. One proposed method is that helium from the secondary falls onto the C-O WD, becoming degenerate. When enough helium has accumulated, a helium flash occurs, that will not only burn helium to carbon and oxygen, but will also send a shock wave downward that causes ignition of the degenerate C-O core. Another version of this scenario does not invoke degenerate helium, but simply has carbon and oxygen igniting in the interior of the WD as it nears the Chandrasekhar limit, at which point the degenerate gas is no longer able to support the star. What happens next is also unclear. It is not known if the burning front of carbon and oxygen occurs at subsonic speeds (known as **deflagration**) or if it accelerates and steepens to become a supersonic burning front (known as a **detonation**, or true explosion). The details of this affect the resulting light curve as well as the relative abundances of elements observed in the spectrum (Carroll & Ostlie 2007, pg. 688).

### RADIOACTIVE DECAY IN LIGHTCURVES

A Type II-P SN is the most common type of core-collapse SN. The source of the plateau in the light curve is due largely to the energy deposited by the shock into the hydrogen-rich envelope. As the ejecta cools through adiabatic expansion, it enters a stage of prolonged hydrogen recombination, releasing energy and photons at a nearly constant temperature of about 5000 K; this is the origin of the plateau. The plateau may also be further supported by the energy deposited in the envelope by the radioactive decay of  $^{56}_{28}\text{Ni}$  (half-life of 6 days) into  $^{56}_{27}\text{Co}$ . The energy released by the decay is deposited into the optically thick expanding shell, which is then radiated away from the SNR's photosphere. This holds up the light curve for a time, extending the observed plateau. Eventually, the expanding gas cloud will become optically thin, exposing the central product of the explosion, the NS or BH. Furthermore, the decay product  $^{56}_{27}\text{Co}$  can also decay into  $^{56}_{26}\text{Fe}$  with a half-life of 78 days. This implies that as the luminosity of the SN diminishes in time, it should be possible to detect the contribution to the light being made by this decay process. Type II-L SNe appear to have had progenitor stars with significantly reduced hydrogen envelopes, implying that the signature of radioactive decay becomes evident almost immediately after the event (Carroll & Ostlie 2007, pg. 534).

### NUCLEOSYNTHESIS DURING SUPERNOVA

When nuclei having progressively larger values of  $Z$  (the number of protons) form via stellar nucleosynthesis, it becomes increasingly difficult for other charged particles, such as protons,  $\alpha$ -particles, and so on, to react with them, due to an enhanced Coulomb barrier. However, the same limitation does not exist when neutrons collide with these nuclei. Consequently, nuclear reactions involving neutrons can occur even at relatively low temperatures, assuming, of course, that free neutrons are present in the gas. The reactions with neutrons



result in more massive nuclei that are either stable or unstable against the  $\beta$ -decay reaction,



If the  $\beta$ -decay half-life is short compared to the timescale for neutron capture, the neutron-capture reaction is said to be slow or an **s-process**. Such reactions tend to yield stable nuclei, either directly or secondarily via  $\beta$ -decay. On the other hand, if the half-life for the  $\beta$ -decay reaction is long compared with the timescale for neutron capture, the neutron-capture reaction is termed rapid or an **r-process**, and results in neutron-rich nuclei. s-Process reactions tend to occur in normal phases of stellar evolution, whereas r-processes can occur during a SN when a large flux of neutrinos exists. Although neither process plays significant roles in energy production, they do account for the abundance ratios of nuclei with  $A > 60$  (Carroll & Ostlie 2007, pg. 543).

### GAMMA-RAY BURSTS

About once per day, at some random location in the sky, a shower of gamma-ray photons with energies ranging from about 1 keV to many GeV appears; these events are known as **gamma-ray bursts (GRBs)**, and although the lower end of the energy scale includes X-rays, the majority of the energy is contained in the gamma-ray regime. There are two distinct classes of GRBs: those events that last longer than 2 seconds are referred to as **long-soft** GRBs, while those that are shorter than 2 seconds are **short-hard** events. “Soft” and “hard” refer to having more of the event energy at lower energies and higher energies, respectively. Subsequent observations of GRBs in X-ray, optical, and radio wavelengths have confirmed that these objects are extragalactic in origin (Carroll & Ostlie 2007, pg. 547).

Long-soft GRBs are believed to be produced during the core-collapse SN of very massive (possibly WR) stars. For such massive stars, a black hole with a surrounding debris disk will form during a SN. The collimating effect of the debris disk and associated magnetic fields would lead to a jet emanating from the centre of the SN. Since the jet material will be highly relativistic, it will appear to be further collimated. The jet will plow its way through the overlying material of the infalling stellar envelope producing bursts of gamma-rays. This is known as the **collapsar model** of long-soft GRBs, while a competing theory called the **supernova model** is similar, but assumes that this mechanism is delayed until after the overlying material has cleared (Carroll & Ostlie 2007, pg. 548). Short-hard GRBs, on the other hand, are thought to be the result of mergers of compact objects, either two NSs or a NS and a BH. These types of GRBs emit roughly 1000 times less energy than their long-soft counterparts (Carroll & Ostlie 2007, pg. 705). These ephemeral GRBs are more difficult to study due to their short durations and lack of afterglow in other wavelength bands. Nevertheless, their short durations are indicative of compact objects with small physical dimensions (Wikipedia).

**QUESTION 7**

**Small asteroids are usually odd shaped, while larger celestial bodies are round. The dividing line occurs at around 200 km. Explain what this is determined by. Using the same logic, can you estimate the tallest mountain that can stand on Earth and Mars, respectively?**

## QUESTION 7

**Small asteroids are usually odd shaped, while larger celestial bodies are round. The dividing line occurs at around 200 km. Explain what this is determined by. Using the same logic, can you estimate the tallest mountain that can stand on Earth and Mars, respectively?**

We begin by considering a spherical asteroid of radius  $R = 100$  km and uniform density  $\rho$ . We know that this is the dividing line for which smaller asteroids will have insufficient self-gravity to pull themselves into spheres. For larger asteroids the pressure within its interior is strong enough to deform the rock, allowing it to flow and form a spherical body. To determine the critical pressure at which the rock will deform involves calculating the central pressure of our candidate asteroid.

The pressure profile within the asteroid can be modelled using the formula for hydrostatic equilibrium,

$$\frac{dP}{dr} = -G \frac{M(r)\rho}{r^2}, \quad (289)$$

If we assume that the asteroid has uniform density  $\rho$ , then the mass contained within radius  $r$  is simply  $M(r) = 4\pi\rho r^3/3$ . In this case we can integrate Eq. (289) from the surface to the interior, yielding

$$P_{\max} = -\frac{4\pi G\rho^2}{3} \int_R^0 r dr \rightarrow P_{\max} = \frac{2\pi G\rho^2}{3} R^2. \quad (290)$$

Pressures above  $P_{\max}$  will deform the rock and allow the asteroid to flow into a spherical body.

Based on this information we can now estimate the tallest mountains that can stand on Earth and Mars. To investigate this issue we will crudely model mountains as rectangular prisms with base area  $A$ , height  $h$ , and uniform density  $\rho$ . The pressure at the base of the mountain is simply  $P_{\text{base}} = \rho gh$ , where  $g$  is the gravitational acceleration at the base of the mountain (we assume that  $g$  does not vary much over the total height). The maximum height achievable is that for which the pressure at the base is less than the maximum pressure for which the rock can withstand without deforming. Therefore, we have the condition that

$$P_{\text{base}} < P_{\max} \rightarrow \rho gh < \frac{2\pi G\rho^2}{3} R^2 \rightarrow h < \frac{2\pi G\rho}{3g} R^2. \quad (291)$$

Standard rock densities are  $\rho \sim 2.65$  g cm<sup>-3</sup><sup>55</sup> and the gravitational acceleration on the surface of Earth and Mars are 9.81 and 3.71 m/s<sup>2</sup>, respectively (WolframAlpha). Equation (291) then implies that the maximum height of terrestrial and martian mountains are 377 and 998 m, respectively. The actual heights of Mount Everest and Olympus Mons are 8850 and 21170 m, larger by a factor of roughly 20 from our crude estimates (WolframAlpha).

### ASTEROID SHAPES

The known asteroid population spans some five orders of magnitude in size, up to a diameter of about 1000 km. As a consequence, it happens to include the transition from the shape regime dominated by material strength, to that controlled by self-gravitational forces. Most small asteroids are irregular rocky boulders, generated by catastrophic fragmentation events, while the largest ones instead look like regularly-shaped small planets. However, there are exceptions to this rule, due to variations in self-gravity, material properties, and collisional history. In the intermediate transition range, in particular, a variety of shapes and surface morphologies are present (Farinella & Zappalà 1997).

At sizes of a few hundred km, the transition occurs between celestial bodies whose global shape is *irregular*, namely can keep for long times a memory of their origin and history, and bodies for which self-gravity is intense enough to overcome the solid-state material rigidity, so that – if we neglect small-scale surface topography – their shapes are relaxed to equilibrium figures, moulded exclusively by gravitational, rotational, and (possibly) tidal forces. Lightcurve amplitudes, defined as the peak-to-peak variations of the asteroid brightness over a rotational cycle, provide a crude indicator for asteroid shapes. This is true in particular when lightcurves taken at several different viewing orientations are available: the amplitude then varies from zero, when the body is viewed pole-on and no change in the projected surface area is seen, to a maximum value, which corresponds to the equatorial view and gives an estimate of the global deviation of the shape from an axisymmetric one (Farinella & Zappalà 1997).

Asteroids and comets are basically large chunks of rock, metal, and ice; obviously, rocks and icebergs can take a variety of shapes. Perhaps then, we should ask the question as to why the largest bodies are round. As we now know, the answer is a question of material strength. The larger the asteroid or planet, the greater is the pressure at the centre. If the central pressure exceeds the strength of the rocky material, the material will deform, either by plastic flow<sup>56</sup> or by fracture, as a result of failure of the normal elastic properties of the solid rock. Higher internal temperatures also favour deformation toward an equipotential shape. In the case of an ideal non-rotating body, this shape will be perfectly spherical (Hartmann 2005, pg. 176). As shown in Figure 81 there is a strikingly sharp cutoff in irregular shapes of asteroids at diameters around 400 to 600 km. Those asteroids with diameters above this threshold become round as they begin to obey hydrostatic equilibrium, with internal pressures and temperatures becoming great enough to cause failure of the normal elastic properties of the rock (Hartmann 2005, pg. 192).

<sup>55</sup> This is actually close to the values found for many large asteroids where  $\bar{\rho} \sim 3$  g cm<sup>-3</sup> based on measurements of their sizes and masses. Note that mass is a hard quantity to measure for asteroids with only a few of the largest asteroids allowing this measurement based on their small gravitational perturbations to other asteroids (Hartmann 2005).

<sup>56</sup> Think of asphalt as an example; if you strike it with a hammer it chips, but if you place a cannonball on its surface, it will eventually sink in. These properties are believed to arise at least in part from small cracks and imperfections in the crystal lattices of the material. A sudden force may fracture the lattice entirely, but a prolonged gentle stress may cause imperfections like holes in the lattice to propagate through the material, so that on a macroscopic scale the material appears to flow (Hartmann 2005, pg. 192).

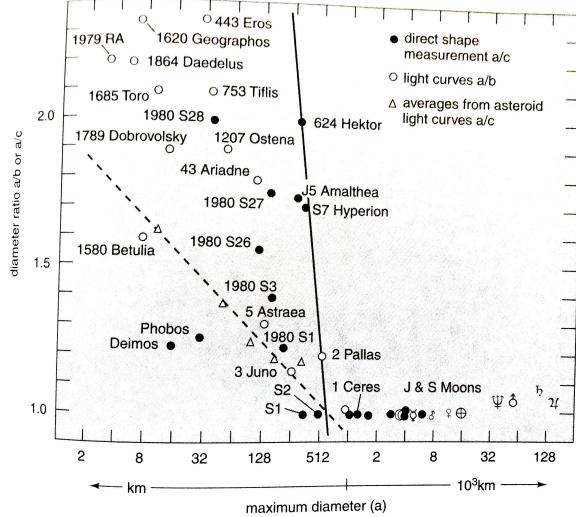


FIG. 81.— Irregular shapes in planetary bodies, as a function of size. Closed circles and planetary symbols show direct measure of  $a/c$ , where  $a$  and  $c$  are maximum and minimum diameters. Open symbols are based on asteroid light curves, which reveal the lower ratio of  $a/b$  where  $b$  is the intermediate diameter. Both the trend of  $a/c$  and  $a/b$  indicate the maximum diameter for irregular bodies is about 600 km. Larger planetary bodies have spheroidal shapes because planetary interiors are not strong enough to support irregularities of such large scale. Image taken from Hartmann (2005).

#### ASTEROIDS AS RUBBLE PILES

The accepted picture of asteroids is that they represent “rubble piles” or gravitational aggregates that lack cohesion while remaining non-fluid. Evidence supporting this claim come from observations of the spin rates of small ( $R \lesssim 20$  km) asteroids. In particular, a class of “fast-rotating asteroids” (FRAs) show interesting characteristics which give constraints to theories about their structure and physical processes. Firstly, roughly 25% of small asteroids are FRAs, while essentially no large asteroids appear to have large spin rates. Secondly, mean lightcurve amplitudes of FRAs are smaller than those of slower rotators. These observations indicate that FRAs are less elongated and more spheroidal than slow rotators, and the tendency to spheroidal shapes is higher for faster spin rates. Finally, there is an apparent cutoff of the spin rate at a period of roughly 2.2 hrs (Pravec & Harris 2000).

The cutoff, as well as the tendency to spheroidal shapes, are evidence that FRAs larger than  $\sim 100$  m are rubble piles. A rotating, homogeneous sphere will be everywhere in a state of compression as long as the rotation frequency does not exceed the surface orbit frequency about the sphere. This is simply the equivalent of saying that the centrifugal acceleration at the equator is less than the acceleration of gravity. By equating the acceleration of gravity at the surface with the centrifugal acceleration at the equator, we can derive a criterion for the critical limit of rotation period depending only on the density of the sphere:

$$\frac{Gm}{r^2} = \omega^2 r \rightarrow P_{\text{crit}} = \frac{3.3 \text{ hr}}{\sqrt{\rho}}, \quad (292)$$

which gives good agreement with the observed  $P_{\text{crit}} = 2.2$  hr for  $\rho \sim 2.5 \text{ g/cm}^3$ . Hence, the sharp truncation in spin rates appears to be strong evidence that most asteroids with  $R \gtrsim 100$  m are “strengthless” rubble piles rather than monolithic bodies. Gravity can be the only force needed to hold an asteroid body together and there is not any indication of a presence of tensile strength in asteroids greater than a few hundred meters. If there are asteroids with significant tensile strength, they must not be common, since it would be unlikely to observe the concentration of spin rates right up to the 2.2 hr cutoff but none beyond, if a large fraction of FRAs were monoliths. Note that the lack of tensile strength does not mean that rubble piles are in hydrostatic equilibrium. In fact, friction and finite particle size effects prevent the bodies from forming equilibrium shapes. If rubble pile bodies formed equilibrium shapes, they would not be FRAs, since rotation periods for such objects with bulk densities plausible for asteroids have rotation periods greater than 4 hrs (Pravec & Harris 2000).

Interestingly, roughly 15% of near-Earth asteroids (NEAs) and main-belt asteroids (MBAs) with  $R < 5$  km have satellites. The angular momentum content from the primary’s rotation and the secondary’s orbit among small binaries suggests that the satellites were formed by rotational disruption after the body was pushed beyond its critical spin limit. Tidal encounters can account for near-critical spin rates and are efficient at forming binaries from rubble piles; however, they are even more efficient at subsequently dissociating those binaries as a result of repeated planetary encounters. One mechanism that operates on both NEAs and MBAs that may lead to the observed binaries is YORP-induced spinup (named after its discoverers), which arises from reflection and/or absorption and re-radiation of sunlight by the surface of an irregularly shaped asteroid. The timescale for YORP spin alteration depends on the size of the body (increasing with  $R^2$ ), the distance from the Sun (increasing with  $a^2$ ), the body’s thermal properties, and the body’s shape and obliquity. The YORP spinup/spindown timescale for km-sized NEAs and MBAs is estimated to be between  $\sim 0.1 - 1$  Myr, depending on the shape and makeup of the asteroid. For this process, mass shed from the equator of a critically spinning body accretes into a satellite and the primary maintains a low equatorial elongation. The satellite forms mostly from material originating near the primary’s surface and enters into a close, low-eccentricity orbit. The observed NEA and MBA binary fraction of  $\sim 15\%$  is probably a balance between YORP spinup and planetary tidal forces that suppress

this affect (Walsh et al. 2008).

**QUESTION 8**

**Why are low mass stars convective in their outer envelopes while high mass stars are convective in their inner cores?**

## QUESTION 8

**Why are low mass stars convective in their outer envelopes while high mass stars are convective in their inner cores?**

Three different energy transport mechanisms operate in stellar interiors: **radiation**, **convection**, and **conduction**. In the latter, heat conduction, energy transfer occurs via collisions during the random thermal motion of particles (electrons and nuclei in completely ionized matter, otherwise atoms or molecules). For ordinary stellar matter (i.e. a non-degenerate gas) conduction has no chance of taking over an appreciable part of the total energy transport. Heat conduction is basically a diffusion process, as in equation (279), and although large densities are present within stars ( $\rho \sim 1.4 \text{ g cm}^{-3}$ ), the collisional cross-sections of colliding particles is rather small ( $\sim 10^{-19} \text{ cm}^{-2}$ ), so that typical mean free paths are several orders of magnitude smaller than those of photons; and the velocity of these particles is only a few percent of  $c$ . Therefore, the conductive diffusion is much smaller than that for photons (Kippenhahn & Weigert 1990, pg. 31). For this reason, we will omit a discussion on conduction, and focus only on radiative and convective heat transport.

### RADIATIVE TRANSPORT

Radiation allows the energy produced by nuclear reactions and gravitation to be carried to the surface via photons, the photons being absorbed and reemitted in nearly random directions as they perform a random walk throughout the interior. This suggests that the opacity of the material must play an important role, as one would expect. Considering this meandering journey through the star, it may seem surprising that the energy ever manages to escape to the surface. This situation is analogous to the motions of air molecules in a closed room. An individual molecule moves with a speed of nearly  $500 \text{ m s}^{-1}$ , and it collides with other air molecules several billion times per second. However, because there is no overall migration of the molecules in a closed room, a person standing in the room feels no wind. On the other hand, opening a window generates a breeze as a pressure gradient is established between one side of the room and the other. The air in the room responds to this pressure gradient, producing a net flux of molecules toward the area of lower pressure. In a star the same mechanism that causes a *breeze* of photons to move toward the surface of the star. Because the temperature of a star decreases outward, the radiation pressure is smaller at greater distances from the centre ( $P_{\text{rad}} \propto T^4$ ). This gradient in the radiation pressure produces the slight net movement of photons toward the suffices that carries the radiative flux (Carroll & Ostlie 2007, pg. 255).

Through a detailed account of integrating the radiative transfer equation, this process can be summarized as

$$\frac{dP_{\text{rad}}}{dr} = -\frac{\kappa\rho}{c} F_{\text{rad}}, \quad (293)$$

which physically states that a net radiative flux is driven by differences in the radiation pressure, with a photon wind blowing from high to low  $P_{\text{rad}}$  (Carroll & Ostlie 2007, pg. 261). By noting that  $P_{\text{rad}} = aT^4/3$ , this can then be transformed into

$$\frac{dT}{dr} = -\frac{3}{4ac} \frac{\kappa\rho}{T^3} \frac{L_r}{4\pi r^2}. \quad (294)$$

From this expression, we see that if either the flux of opacity increases, the temperature gradient must become steeper (more negative) if radiation is to transport all of the required luminosity outward. The same situation holds as the density increases or the temperature decreases (Carroll & Ostlie 2007, pg. 316). Assuming hydrostatic equilibrium, it is common to rewrite equation (294) in terms of  $dT/dP$ , that is the gradient describing the temperature variation with depth, where the depth is expressed in terms of pressure, which is monotonically increasing function with depth (Kippenhahn & Weigert 1990, pg. 32). In this case, we have:

$$\nabla_{\text{rad}} \equiv \left( \frac{d \ln T}{d \ln P} \right)_{\text{rad}} \equiv \frac{P}{T} \frac{dT}{dP} \rightarrow \nabla_{\text{rad}} = \frac{3}{4ac} \frac{\kappa P}{M_r T^4} \frac{L_r}{4\pi G}. \quad (295)$$

### CONVECTIVE TRANSPORT

Convective transport of energy means an exchange of energy between hotter and cooler layers in a dynamically unstable region through the exchange of macroscopic mass elements, or bubbles, the hotter of which move upwards while the cooler ones descend. The moving mass elements will finally dissolve in their new surroundings and thereby deliver their excess (or deficiency) of heat. Owing to the high density in stellar interiors, convective transport can be very efficient. However, this energy transport can operate only if it finds a sufficient driving mechanism in the form of buoyancy forces. Unfortunately, a theoretical treatment of convective motions within stars is extremely difficult, so that most descriptions resort to the so-called **mixing length theory**, which was developed in complete analogy to molecular heat transfer. Equation (295) shows the temperature gradient that would be maintained in a star if the whole luminosity had to be transported outwards by radiation only. If convection contributes to the energy transport, the actual gradient  $\nabla_{\text{act}}$  will be different; namely, it will be suppressed (Kippenhahn & Weigert 1990, pg. 48).

Investigating the question of whether or not convection will occur is equivalent to describing **dynamical instability** within a star. For this kind of instability, we assume that mass elements move fast enough that no appreciable amounts of heat are exchanged with their surroundings, and they therefore move adiabatically. Whether convection occurs within a given region of a star depends on the question of whether small perturbations present will grow or stay small. For this purpose, we will describe a fluctuation as some bubble (subscript  $b$ ) with physical quantities ( $T, P, \rho$ ) that are constant, but somewhat different than the average surroundings (subscript  $s$ ). For any quantity  $A$  we define the difference  $\delta A$  between the bubble and its surroundings as  $\delta A \equiv A_b - A_s$  (Kippenhahn & Weigert 1990, pg. 36).

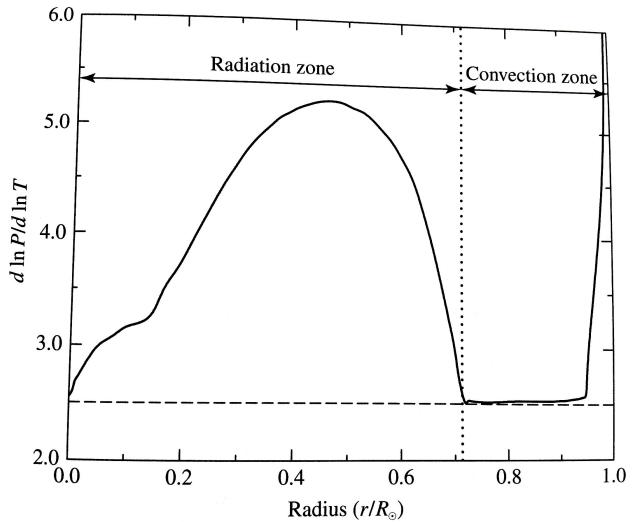
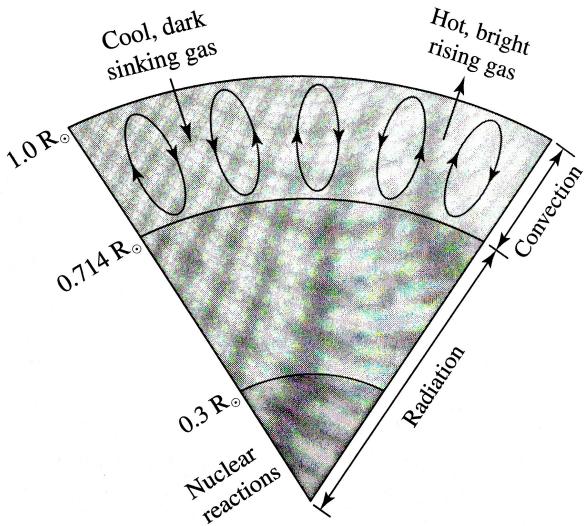


FIG. 82.— (left) A schematic diagram of the Sun's interior showing the radiation and convection zones, along with the main area of nuclear reactions. (right) The convective condition  $d \ln P / d \ln T$  plotted versus  $r / R_{\odot}$  for a numerical simulation of the Sun. The dashed horizontal curve represents the boundary between adiabatic convection and radiation for an ideal monatomic gas. The onset of convection does not exactly agree with this case because of the incorporation of a more sophisticated equation of state and a more detailed treatment of convection physics. Images taken from Carroll & Ostlie (2007).

Now, suppose that we have an initial fluctuation in temperature, for example a slightly hotter bubble with  $\delta T > 0$ . Normally we would also expect an excess  $\delta P > 0$ . However, this bubble will expand immediately until pressure balance with its surroundings is restored, and since this expansion occurs at the sound speed, it is usually more rapid than any other motion of the bubble; thus we will always assume that  $\delta P = 0$ . For an ideal gas at constant pressure, the assumed  $\delta T > 0$  requires that  $\delta \rho < 0$ . Hence, the bubble is lighter than the surrounding material, and the buoyancy forces will lift it upwards. After the bubble moves up some distance  $\Delta r$ , its density will differ from the surroundings by

$$\delta \rho = \left[ \left( \frac{d\rho}{dr} \right)_b - \left( \frac{d\rho}{dr} \right)_s \right] \Delta r, \quad (296)$$

where the first derivative represents the change in the bubble's density after adiabatically rising a distance  $dr$  and the second derivative is the spatial gradient of the surroundings. If  $\delta \rho < 0$  then a buoyancy force will lift the bubble upwards, and we can say that the situation is *unstable* since the bubble has been lifted further. Physically, the condition for convection thus requires that  $\delta \rho < 0$  (i.e. the bubble's density decreases faster than for the surroundings). However, this criterion is not very useful, since it requires knowledge of density gradients. It is therefore preferable to turn them into temperature gradients. Assuming ideal gases and hydrostatic equilibrium it is possible to compute the condition for convection as

$$\nabla_{\text{act}} > \nabla_{\text{ad}} \Leftrightarrow \left| \frac{dT}{dr} \right|_{\text{act}} > \left| \frac{dT}{dr} \right|_{\text{ad}}, \quad (297)$$

where  $\nabla_{\text{act}}$  and  $\nabla_{\text{ad}}$  represent the temperature gradient of the surrounding material and the variation in temperature as the bubble rises adiabatically, respectively. The negative of this relation is known as the **Schwarzschild criterion** for dynamical stability. Another variation of equation (297) involves the addition of an  $\nabla_{\mu}$  term on the right-hand side which describes the spatial variation in the chemical composition of the surroundings. If heavier elements exist below lighter ones then  $\nabla_{\mu} > 0$  and this term acts to stabilize the effect. Physically, this occurs since the bubble will carry its heavier material upwards into lighter surroundings, and gravity will tend to draw it back to its original place (Kippenhahn & Weigert 1990, pg. 39).

If these criteria favour stability, then no convective motions will occur, and the whole flux will be carried by radiation (i.e.  $\nabla_{\text{act}} = \nabla_{\text{rad}}$ ). If they favour instability, then small perturbations will increase to finite amplitude until the whole region boils with convective motions that carry part of the flux – and  $\nabla_{\text{act}}$  must be determined through something like mixing-length theory. This instability can be caused either by the fact that  $\nabla_{\text{rad}}$  has become too high (large flux, or very opaque matter), or else by a depression of  $\nabla_{\text{ad}}$ ; both cases occur in stars (Kippenhahn & Weigert 1990, pg. 40).

#### STELLAR CONVECTIVE ZONES

Analysis of equation (297) allows us to develop some understanding of which conditions are likely to lead to convection over radiation. In general, convection will occur when (1) the stellar opacity is large, implying that an unachievable steep temperature  $\nabla_{\text{act}}$  would be necessary for radiative transport, (2) a region exists where ionization is occurring, causing a large specific heat and a low adiabatic temperature gradient  $\nabla_{\text{ad}}$ , and (3) the temperature dependence of the nuclear energy generation rate is large, causing a steep radiative flux gradient and a large temperature gradient. In the atmosphere of many stars, the first two conditions can occur simultaneously, whereas the third condition would occur only deep in stellar interiors. In particular, the third condition can occur when the highly temperature-dependent CNO cycle or triple alpha processes are occurring (Carroll & Ostlie 2007, pg. 325).

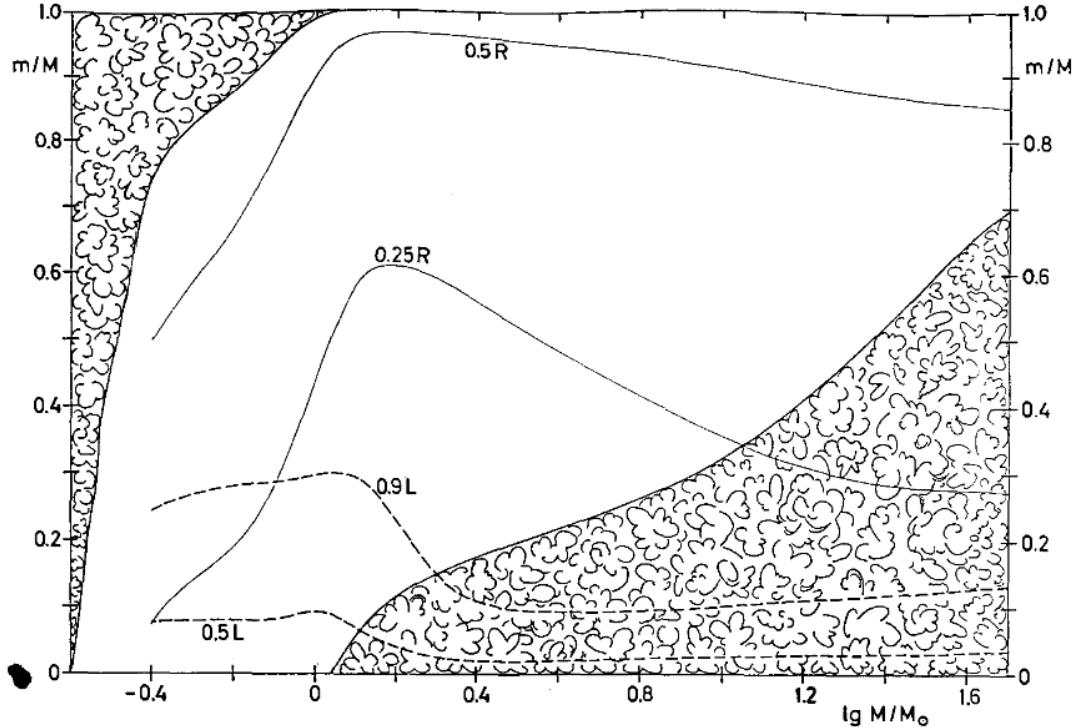


FIG. 83.— The mass values  $m$  from centre to surface are plotted against stellar mass  $M$ . Cloudy regions indicate the extension of convective zones within these ZAMS models. Two solid lines give the  $m$  values at which  $r$  is  $1/4$  and  $1/2$  of the total radius  $R$ . The dashed lines show the mass elements inside which  $50\%$  and  $90\%$  of the total luminosity  $L$  are produced. Image from Kippenhahn & Weigert (1990).

For an ideal gas equation 297 can be rewritten in the equivalent way:

$$\frac{d\ln P}{d\ln T} < \frac{\gamma}{\gamma-1} \quad (= 2.5 \text{ for } \gamma = 5/3). \quad (298)$$

Figure 82 plots  $d\ln P/d\ln T$  as a function of depth for an interior model of the Sun. As can be seen, the Sun is purely radiative below below  $0.7 R_\odot$  and becomes convective above that point. As we know, this occurs because the opacity in the outer proton of the Sun becomes large enough to inhibit the transport of energy by radiation. Throughout most of the region of convective energy transport,  $d\ln P/d\ln T \approx 2.5$ , which is characteristic of the nearly adiabatic temperature gradient of most convection zones. The rapid rise in  $d\ln P/d\ln T$  above  $0.95 R_\odot$  is due to the significant departure of the actual temperature gradient from an adiabatic one (i.e. the adiabatic approximation is invalid near the surface of the Sun). Notice also that  $d\ln P/d\ln T$  decreases to almost 2.5 at the centre of the Sun. Although the Sun remains purely radiative at the centre, the large amount of energy that must be transported outward pushes the temperature gradient in the direction of becoming dynamically unstable. For stars slightly more massive than the Sun, their cores are convective because of the stronger temperature dependence of the CNO cycle as compared to the pp chain (Carroll & Ostlie 2007, pg. 355).

The interior structure of stars along the MS varies with mass, primarily in the location of convection zones. In the upper portion of the MS, where energy generation is due to the strong temperature-dependent CNO cycle, convection is dominant in the core. This occurs because the rate of energy generation changes quickly with radius, and radiation is not efficient to transport all of the energy being released in nuclear reactions. Outside of the hydrogen-burning core, radiation is again capable of handling the flux, and convection ceases. As the stellar mass decreases, so does the central temperature and the energy output of the CNO cycle until, near  $1.2 M_\odot$ , the pp chain begins to dominate and the core becomes radiative. Meanwhile, near the surface of the star, as the effective temperature decreases with decreasing mass, the opacity increases, in part because of the location of the zone of hydrogen ionization. The increase in opacity makes convection more efficient than radiation near the surfaces of stars with masses less than roughly  $1.3 M_\odot$ . This has the effect of creating convection zones near the surfaces of these stars. As we continue to move down the MS, the bottom of the surface convection zone lowers until the entire star becomes convective near  $0.3 M_\odot$  (Carroll & Ostlie 2007, pg. 343). The dependence of the convection zone region on stellar mass is shown schematically in Figure 83.

#### ATMOSPHERE OPACITY

The primary source of the continuum opacity in the atmospheres of low-mass stars (stars later than F0) is the photoionization of  $H^-$  ions. An  $H^-$  ion is a hydrogen atom that possesses an extra electron. Because of the partial shielding that the nucleus provides, a second electron can be loosely bound to the atom on the side of the ion opposite that of the first electron. The binding energy of  $H^-$  is only 0.75 eV, so that any photon with energy in excess of this amount can be absorbed by the ion and liberate

the extra electron. At longer wavelengths than this threshold,  $H^-$  can also contribute to the opacity through free-free absorption (i.e. opposite of bremsstrahlung). Consequently,  $H^-$  ions are an import source of the continuum opacity for stars cooler than F0. Obviously,  $H^-$  ions becomes increasingly ionized at higher temperatures and therefore make less of a contribution to the continuum opacity. For stars of spectral types B and A, the photoionization of hydrogen atoms and its free-free absorption are the main sources of the continuum opacity. At the even higher temperatures encountered in O stars, the ionization of atomic hydrogen means that electron scattering becomes increasingly important, with the photoionization of helium also contributing to the opacity. In the coolest stellar atmospheres, molecules can exist and contribute to the bound-bound and bound-free opacities; the large number of discrete molecular absorption lines is an efficient impediment to the flow of photons (Carroll & Ostlie 2007, pg. 248). Based on this information we may suppose that stellar envelopes are more opaque for low-mass stars. This conclusion is further corroborated from Kramer's opacity law, which states that  $\kappa \propto \rho/T^{3.5}$ . Not only are low-mass stars cooler, but their densities are also larger ( $R \propto M$  on the MS so  $\rho \propto M^{-2}$ ), implying a larger opacity. From condition (1) above, these large opacities lead to convection zones in their outer envelopes, explaining why we observe low-mass stars to be convective there.

#### STELLAR PULSATIONS

The mechanisms responsible for stars outside the instability strip are not well understood. The long-period variables are red supergiants (AGB stars) with huge, diffuse convective envelopes surrounding a compact core. Their spectra are dominated by molecular absorption lines and emission lines that reveal the existence of atmospheric shock waves and significant mass loss. While we understand that the partial hydrogen ionization zone drives the pulsation of these stars, many details need to be explained, such as how its oscillations interact with its outer atmosphere (Carroll & Ostlie 2007, pg. 498).

**QUESTION 9**

**Describe and compare the ages of the surfaces of Mercury, Venus, Earth and Mars.**

### QUESTION 9

**Describe and compare the ages of the surfaces of Mercury, Venus, Earth and Mars.**

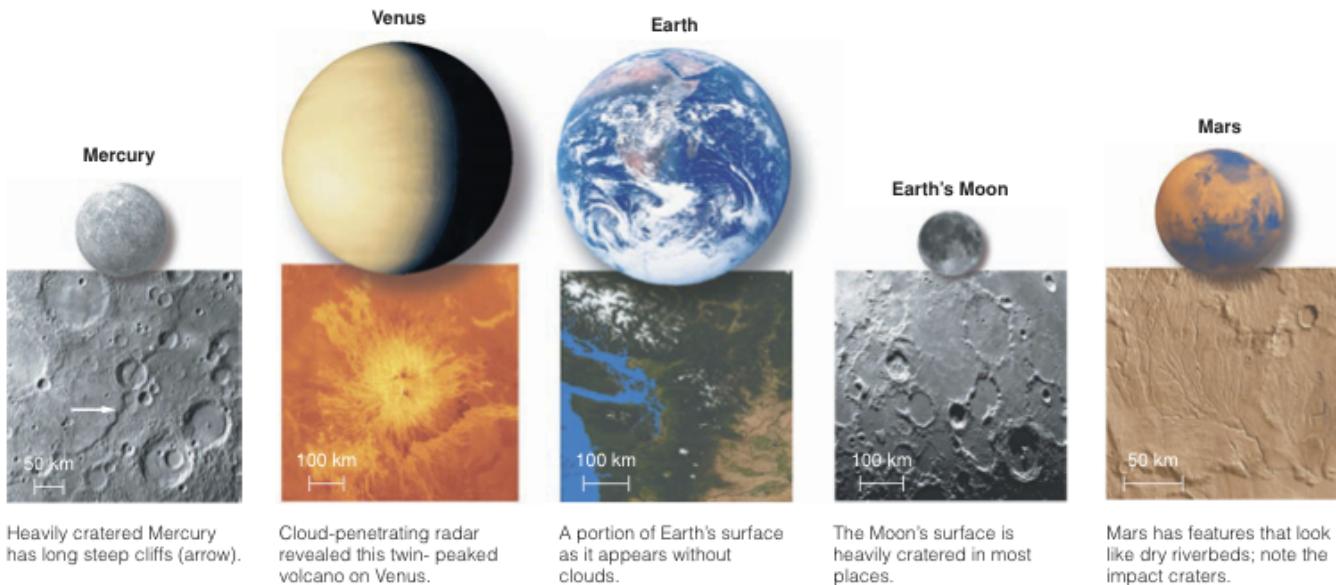


FIG. 84.— The terrestrial planets (and the moon), shown to scale, along with sample surface close-ups from orbiting spacecraft. Image taken from Bennett et al. (2012).

The terrestrial planets are believed to have been built up in the same way and correspondingly display similarities, such as being small, rocky, and slowly rotating. However, upon closer examination there are many stark differences between their appearances and structures, both internally and externally; see, for example, Figure 84 for a comparison between surface features. In short, the different surface features arise from the varying heating and cooling sources available for each planet, as determined by both their location in the solar system and their size. In particular, a planet is termed **geologically active** if ongoing internal processes – such as volcanism, plate tectonics, and weather erosion – are actively reshaping the surface. Most of this geological activity is the result of physics occurring within the interior of the planet (Bennett et al. 2012, pg. 192).

#### THE MOON

Though the Moon is obviously not a planet, it is often grouped amongst the terrestrial worlds due to common similarities. Moreover, its closeness to Earth has made it useful in calibrating the ages of the other terrestrial planets based on cratering. Despite the proximity of the Moon to Earth, the two worlds are very different. Because of its low surface gravity, the Moon has been unable to retain a significant atmosphere. Without a protective atmosphere, the Moon has suffered impacts by meteorites throughout its history. Along with a large number of smaller impacts, a significant number of very large collisions occurred roughly 700 Myr after its formation. These impacts were powerful enough to penetrate its thin crust, allowing molten rock in the interior to flow across the surface. The result was the formation of the many smooth, roughly circular **maria** that can be seen on its Earth-facing surface<sup>57</sup>. Rocks have been collected from both the maria and the **highland** (or mountainous) regions between the maria. Composition analysis of these samples confirm that the maria rocks are basalts, similar to the kind of volcanic rock found on Earth. The lunar basalts are rich in Fe and Mg, and contain glassy structures that are characteristic of rapid cooling. However, unlike Earth basalts, the lunar samples contain no water and a lower percentage of volatile compounds. Radioactive dating of these rocks (e.g., rubidium-87 with a half-life of 48 Gyr) confirm that the maria are relatively young ( $\sim 3.5$  Gyr) compared to the older ( $\sim 4.4$  Gyr) highlands. Furthermore, this dating of lunar samples implies that a spike of **late heavy bombardment (LHB)** occurred roughly 700 Myr after the Moon formed. It was during that time that the majority of the cratering occurred in the highlands, with some of the largest collisions producing the maria. Over the last 3.8 Gyr meteorite impacts have continued, but at a significantly reduced rate. In this way the fairly smooth, relatively uncratered surfaces of the maria have been maintained (Carroll & Ostlie 2007, pgs. 754-760). The only ongoing change to the Moon's surface is a very slow *sand-blasting* of the surface by micrometeorites, sand-sized meteorites that perpetually rain down on the Moon. This process gradually pulverizes the surface rock, producing a lunar surface covered by a thin layer of powdery regolith (Bennett et al. 2012, pg. 203).

The counting of impact craters offers a simple method to estimate the ages of geologic features on planetary surfaces when *in situ* rock samples are lacking. The crater chronology method is based on the simple idea that old surfaces have accumulated more impact craters than more recent ones. The relationship between geologic age and number of lunar craters, based on the radiometric dating of existing lunar rock samples, is found to be approximately linear from the present to  $\sim 3$  Gyr ago, and

<sup>57</sup> At the original time of impact, the Moon's interior had cooled enough that it had solidified. Not for another  $\sim 100$  Myr did the core become molten again through energy released by radioactive decay. There was therefore an intervening period between the original fracture on the Moon's surface and the filling up with lava, explaining why the maria appear smooth today. Since this time the Moon's interior has once again solidified (Bennett et al. 2012, pg. 202).

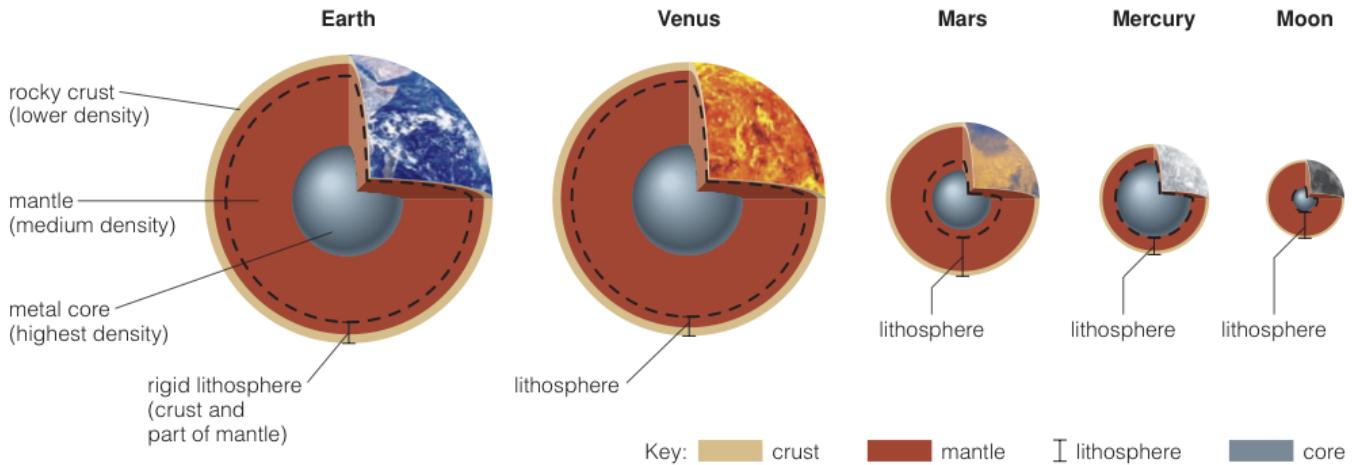


FIG. 85.— Interior structures of the terrestrial worlds, shown to scale and in order of decreasing size. Colour coding shows the core-mantle-crust layering by density; a dashed circle represents the inner boundary of the lithosphere, defined by strength of rock rather than by density. The thicknesses of the crust and the lithosphere of Venus and Earth are exaggerated to make them visible in this figure. Image taken from Bennett et al. (2012).

approximately exponential beyond that time; consistent with the LHB period. Since large numbers of Venusian, Martian or Mercurian rock are not available, the surface ages of these planets are estimated by comparing their cratering to that of the moon. Of course, uncertainties arise based on the differences between the planets cratering histories (i.e. things like gravitational focusing and location within the solar system matter) and geological activities can bias the erasing of craters (e.g., preferentially erase small craters) (Le Feuvre & Wieczorek 2011).

#### MERCURY

The surface of Mercury is reminiscent of the Moon. In particular, Mercury is heavily cratered, indicating that it underwent extensive bombardment during its nearly 4.6 Gyr history. One impact, known as Caloris Basin, was so large that it created ripples that traveled across the planet and converged on the opposite side to produce a jumbled collection of hills. A careful comparison of images of the Moon and Mercury show that Mercury's craters are often separated by regions that are largely devoid of significant cratering. Assuming that the rate of impact was roughly the same on both worlds throughout their histories, and that they formed at approximately the same time, Mercury's surface must have been refreshed more recently, meaning that it is somewhat younger. This is consistent with the conclusion that since Mercury is larger and closer to the Sun, it would have cooled off more slowly after formation, and hot, molten material would have been more likely to reach the surface to cover older impact sites (Carroll & Ostlie 2007, pg. 738). Cratering estimates give Mercury's surface an age of  $\lesssim 3.8$  Gyr (Le Feuvre & Wieczorek 2011).

Given the planet's size and its proximity to the Sun, it is not surprising that Mercury has only a very tenuous atmosphere. Because of the high temperature on its subsolar side (800 K) and its relatively low escape velocity, atmospheric gases quickly evaporate into space. What atmosphere it does possess is due to charged nuclei of hydrogen and helium from the strong solar wind that become trapped in its weak magnetic field, together with atoms of oxygen, sodium, potassium, and calcium that have escaped the surface regolith through blasting by solar wind and micrometeorites. Ironically, radar data suggest that this closest planet to the Sun possesses highly reflective volatile material, probably water ice, in permanently shadowed craters near the polar caps. Because tidal interactions have forced the planet's rotation axis to be almost exactly perpendicular to its orbital plane, the polar regions never get more than a very small amount of sunlight. Moreover, with virtually no atmosphere, Mercury cannot efficiently transport heat away from the equatorial regions. As a result, temperature near the pole probably never exceed 170 K, and in shadowed regions this may be as low as 60 K (Carroll & Ostlie 2007, pg. 739).

One surprising feature of Mercury are the existence of many extended cliffs, which have vertical faces up to 3 km high and typically run for hundreds of km across the surface. The interpretation of these features is that they arose when its large core swelled from heat released during its differentiation, and later contracted as it cooled. The thin mantle around the core followed in contraction, generating internal stresses that formed cliffs as the surface crumpled (Bennett et al. 2012, pg. 205).

Mercury's relatively high average density indicates that it must have lost most of its lighter elements and undergone enough gravitational separation to create a fairly dense core. Based on computer simulations, it appears that Mercury may have experienced a major collision with a large planetesimal early in its history. The collision was sufficiently energetic that much of the outer, lighter silicate material was removed, leaving behind the iron and nickel that had previously settled to the centre of the planet (Carroll & Ostlie 2007, pg. 740). Figure 85 compare the interior structure of the different terrestrial planets.

#### VENUS

Venus is often referred to as Earth's sister planet because its mass ( $0.8 M_E$ ) and radius ( $0.95 R_E$ ) are comparable to our own. Despite these basic similarities, the two planets are markedly different in many of their fundamental features. The most obvious difference is that Venus spins in a retrograde sense, rotating almost upside-down compared to the other planets, with a sluggish rotation period of 243 days (actually longer than its orbital period of 225 days). One consequence of this slow rotation is the

lack of any measurable magnetic field. Detailed analytical and numerical studies of the interactions among Venus attribute its retrograde motion to gravitational perturbations from nearby solar system objects. This is coupled to its thick atmosphere that can be significantly affected by tidal forces, and through friction with the surface can produce a damping effect on the planet's rotation. However, the path to its final configuration is still unclear: this can either occur by flipping the rotation axis to near 180° or the slowing of the spin rate to zero at an axis tilt of 0°, and then tides producing a slow retrograde rotation (Carroll & Ostlie 2007, pg. 741).

Venus is well-known for its hot, thick atmosphere composed primarily of CO<sub>2</sub> (97%) that contains thick clouds of concentrated sulphuric acid. At the base of the atmosphere the temperature is 750 K, sufficient to melt lead, and the pressure is 100 atm, equal to the pressure at a depth of about 1 km below the surface of Earth's oceans. The high surface temperature of Venus far exceeds what is expected from simple blackbody arguments, with the difference attributed to the large abundance of carbon dioxide (a greenhouse gas); the atmosphere is so thick that the optical depth at IR wavelengths is  $\tau = 70$ . The origin of the thick atmosphere is not well understood, though it probably arose from a combination of volcanic outgassing and material deposited by meteorites and comets. In contrast to Venus, which has abundant carbon dioxide and a dearth of water in its atmosphere, the Earth has abundant water and very little atmospheric CO<sub>2</sub>. Since these planets formed near each other in the solar nebula and have common attributes, some mechanism must have been responsible for this discrepancy. The idea is that the surface temperature of Venus was large enough (due to its proximity to the Sun) to sustain hot water on its surface. As the planet was bombarded by meteorites, the surface temperature began to rise and the oceans started to evaporate. The addition of more IR-absorbing water in the atmosphere triggered a *runaway greenhouse effect*, causing the surface temperature to climb to nearly 2000 K, hot enough to vaporize the remaining water and even melt rock. Since H<sub>2</sub>O is lighter than CO<sub>2</sub>, the water migrated to the top of the atmosphere where it was dissociated by solar UV radiation. This process liberated the lighter hydrogen atoms, allowing most of them to escape from the planet. Since the carbon dioxide remained, it became the dominant species in the atmosphere of Venus. This reasoning is corroborated by enhanced deuterium-to-hydrogen ratios in the atmosphere of Venus, consistent with the slower escape rate of deuterium (Carroll & Ostlie 2007, pg. 743). The startling difference between Venus and Earth, therefore, can be attributed to Venus' closer proximity to the Sun, which coupled to a runaway greenhouse effect, removed its water content. Of course, both the large abundances of H<sub>2</sub>O and CO<sub>2</sub> still exist on Earth; the former primarily in liquid form within oceans and the latter primarily in solid form, dissolved within the water to produce carbonate rocks (rocks rich in carbon and oxygen) such as limestone (Bennett et al. 2012, pg. 216).

By far the greatest amount of information about the surface of Venus has come from radar imaging, because radio signals can easily penetrate the atmosphere even though visible and UV light cannot. Venus' surface features are composed primarily of rock (and not soil) and its overall landscape resembles that of the Earth with lowland plains and upland hills that have been contorted by tectonic forces. In addition, nearly one thousand volcanic features with extended lava flows have been identified on its surface, with craters dotted in between in a mostly uniform fashion. This suggests that the entire surface is roughly the same age, with cratering estimates putting this at 250 Myr (Le Feuvre & Wieczorek 2011). Based on this information it has been concluded that large-scale lava flows must have recently repaved the surface. Sulphuric acid concentration within the atmosphere suggest that volcanism is still occurring on Venus (Carroll & Ostlie 2007, pg. 744). In contrast to the Earth, Venus does not show strong signs of weather erosion. We can trace the lack of erosion on Venus to two facts. First, Venus is far too hot for any type of rain or snow on its surface. Second, Venus has virtually no wind or weather because of its slow rotation. Without any glaciers, rivers, rain, or strong winds, there is very little erosion on Venus. Furthermore, it appears that plate tectonic activity on Venus has recently ceased. On Earth, plate tectonics resculpts the surface gradually, so that different regions have different age; Venus, on the other hand, has a uniform age across its surface. This indicates the absence of plate tectonics, which suggests that either Venus has suppressed mantle convection, or that its lithosphere somehow resists fracturing. The former is unlikely since Venus should have similar convection as the Earth, while the latter is the more plausible option since the reduced water content in Venus would have acted to stiffen its rock (Bennett et al. 2012, pg. 215).

#### EARTH

The Earth shows a large abundance of liquid water oceans and the presence of extensive life. Most of the carbon dioxide that would otherwise exist within its atmosphere is bound up in carbonate rocks; if all the CO<sub>2</sub> trapped within rock today were released into Earth's atmosphere, the amount would be comparable to that currently contained in the atmosphere of Venus. The present-day atmosphere of Earth is made up of 78% N<sub>2</sub>, 21% O<sub>2</sub>, 1% H<sub>2</sub>O, and traces of Ar, CO<sub>2</sub>, and other constituents. The atmosphere owes its current composition in part to the development of life on the planet (e.g., plants process carbon dioxide into oxygen as a by-product of photosynthesis). It is important to note that the Earth's surface would have been cooler in the past and its water would have been in the form of ice, even as recently as 2 Gyr ago. However, geologic evidence, including fossil records, suggests that Earth's oceans were liquid water as early as 3.8 Gyr ago. The resolution of this paradox probably lies in the details of the greenhouse effect and a different atmospheric composition that exists today (Carroll & Ostlie 2007, pg. 747).

Although the presence of volcanos is a feature that Earth shares with Venus and Mars, Earth's present-day tectonic activity is unique among the terrestrial planets. This activity has its origin in the dynamic interior of Earth. Earth's surface layer, known as the **lithosphere**, encompasses both the oceanic and continental crust as well as the outer portion of the **mantle**. The lithosphere is fractured into crustal plates and rides on the convective, somewhat plastic, underlying upper mantle. As the plates move across the surface of the planet, they crash into or grind against one another, carrying the continents with them. Earth's plate boundaries are generally the sites of active volcanism, mountain building, and frequent earthquakes. For example, when two plates collide, the lighter continental crust overrides the heavier oceanic crust and a subduction zone develops. This creates ocean trenches that can generate volcanic islands as a result of the heat generated by friction as the oceanic crust descends into the mantle. If two plates collide that contain continental crust, neither plate will overrun the other; instead, buckling occurs and a mountain range such as the Himalayas is generated (Carroll & Ostlie 2007, pg. 751). As a result of this ongoing activity 90% of the Earth's crust

is younger than 600 Myr, with the oldest rocks found on Earth dating to 3.8 Gyr. Plate tectonic activity is constantly reclining the surface, carrying old crust down into the mantle and forming new crust to replace it (Carroll & Ostlie 2007, pg. 759).

#### MARS

At first inspection, images taken on Mars give the impression of a dry, dusty world. However, closer inspection has revealed a fascinating world that, although dry today, once clearly had water flowing across its surface. This idea is supported by channels that are characteristic of water erosion found on Earth. There is also evidence that huge flash floods may have occurred on the surface of the planet. It appears that lakes of water may have been present on Mars in the distant past as well, as seen by apparent sediment deposits from ancient Martian lakes within impact craters. With present-day surface temperatures varying between -140° and 20°, combined with the very low atmospheric pressure, it appears that liquid water that was present on Mars is now either trapped in a layer of permafrost or frozen in its polar ice caps (Carroll & Ostlie 2007, pg. 764).

Although water ice is certainly present today in the polar caps, the caps are composed primarily of dry ice (frozen carbon dioxide). Mars' axis tilt of 25° and its orbital period of 1.9 yr means that the planets' seasonal variations are similar to Earth's but are roughly twice as long. Consequently, Mars experiences winter and summer seasons corresponding to observed variations in the sizes of the ice caps. It is the dry ice that sublimates during the Martian summer and freezes back out again during the winter. The small residual cap that remains during the summer is composed of water ice. Furthermore, it appears that the spin axis of Mars fluctuates wildly between about 0° to 60° on timescales as short as a few Myr; the variations are due to gravitational interactions with the Sun and other planets. This implies that at various times in the past, the polar ice caps could completely melt (high tilt angle), while at other times the planet's atmosphere might actually freeze out (low axis tilt). Evidently, the Earth's axis has been stabilized by strong tidal interactions with the Moon (Carroll & Ostlie 2007, pg. 767).

Mars' thin atmosphere is composed of 95% carbon dioxide with the rest mainly molecular nitrogen. Unlike the case of Venus, however, the greenhouse effect has very little influence on the current equilibrium temperature of Mars; the atmosphere is simply too thin. In the past, the atmosphere of Mars may have been much more dense, causing the greenhouse effect to be more efficient than it is today. The water that is currently trapped in the ice caps and permafrost would then have been flowing freely, maybe even resulting in rainfall (Carroll & Ostlie 2007, pg. 767). The presence of volcanos on Mars seem to suggest that outgassing would have produced a thicker atmosphere in the past. The mystery then lies in considering what happened to its atmosphere. Although this is still unclear, it is believed to be related to the cooling of Mars and the corresponding shutdown of its magnetic field. Early in its history, Mars probably had molten convecting metals in its core, much like Earth today. The combination of this core convection and the planet's rotation should have produced a magnetic field and a protective magnetosphere around Mars. The magnetic field would have weakened as Mars cooled and the core ceased to convect, leaving the atmosphere vulnerable to solar wind particles. These solar wind particles could have stripped gases out of the Martian atmosphere and into space (Bennett et al. 2012, pg. 212). Even though the present-day atmospheric density is quite low, it is sufficient to produce huge dust storms that sometimes cover the entire surface of Mars. The seasonal storms are driven by high winds and are responsible for the varieties in surface hues that can be seen from Earth (Carroll & Ostlie 2007, pg. 768).

Mars' main global feature is the difference between the northern and southern hemispheres of the planet, known as a crustal dichotomy. The southern hemisphere is heavily cratered and elevated by several kilometres, while the northern hemisphere is more smooth and lies several kilometres below. Two theories for the crustal dichotomy invoke large-scale mantle convection and a giant impact. The differences in cratering indicate that the southern hemisphere is older, with an age of about 4.4 Gyr, while the northern hemisphere is younger with an age of 3.5 Gyr (Charles).

The dust on the surface appears reddish in colour and contains a relatively high abundance of iron, which oxidizes (rusts) when exposed to the atmosphere. Apparently Mars did not undergo the same degree of gravitational separation than Earth did, possibly because the smaller, more distant planet cooled more rapidly following its formation. However, averaging the density over its volume reveals that the planet is actually underabundant in iron compared to the other terrestrial planets; the reason for this is not understood. The lack of significant gravitational separation is also consistent with the absence of an appreciable magnetic field. If an iron core is present, presumably it is quite small and probably not molten (Carroll & Ostlie 2007, pg. 768).

Even if Mars may not be geologically active today, it certainly has been in the past. A 3000-km-long network of canyons near the planet's equator have apparently been formed by **faulting** (or fracturing of the crust) in order to relieve stresses that built up in the interior. In addition, numerous inactive volcanos have been spotted on its surface. The most famous is Olympus Mons, a shield volcano that covers an area roughly the size of Utah that rises 24 km above the surrounding surface, and has a huge volcanic crater. Its enormous size is attributed to a process known as **hot-spot volcanism**, where a weak spot in the crust has allowed molten material to rise to the surface; this process produced the Hawaiian islands on Earth (Carroll & Ostlie 2007, pg. 768).

#### SOLAR SYSTEM FORMATION

Here we outline a proposed model for the formation of the solar system, following Carroll & Ostlie (2007, pg. 863-866). Within an interstellar gas and dust cloud (perhaps a GMC), the Jeans condition was satisfied locally, and a portion of the cloud began to collapse and fragment. The most massive segments evolved rapidly into massive stars that, within a few Myr, lived their lives and died in SN explosions. These events would have enriched surrounding areas that had not yet collapsed, and may have even triggered their collapse. In any case, the solar nebula formed from the collapse of some small region. Assuming that the solar nebula possessed some initial angular momentum, it would have spun up as it collapsed, producing a protosun surrounded by a disk of gas and dust. In fact, the disk itself probably formed more rapidly than the star did, causing much of the mass of the growing protosun to be funnelled through the disk first.

Within the nebular disk, small grains with icy mantles were able to collide and stick together randomly. When objects of appreciable size were able to develop in the disk, they began to gradationally influence other material in their areas. As the

low-energy collisions continued, progressively larger planetesimals were able to form. In the innermost regions of the disk the accreting particles were composed of **calcium-aluminium-rich inclusions (CAIs)**, silicates, iron, and nickel; relatively volatile materials were unable to condense out of the nebula because of the high temperatures in that region. At distances beyond the snow line, at about 5 AU, the nebula became sufficiently cool that water-ice could form as well. The result was that water-ice could also be included in the growing planetesimals beyond that distance. Even farther out ( $\sim 30$  AU) methane-ice also participated in the development of planetesimals.

The object that grew fastest was Jupiter. Thanks to the presence of water-ice along with rocky materials, and with a nebula that was sufficiently dense in that region, Jupiter's core reached a mass of  $\sim 10 M_E$ . At that point the planet's gravitational influence became great enough that it started to collect the gases in its vicinity (principally hydrogen and helium). The outcome was the formation of the massive planet we see today, together with the Galilean satellites. The entire process of forming Jupiter probably took about 1 Myr, halting once the gas was depleted. The other Jovian planets also likely developed large cores, but could not capture as much gas as Jupiter, simply because their were further out in the solar nebula where the density was lower.

At the inner regions of the solar nebula cooled, the most refractory elements were able to condense out to form the CAIs. Next to condense were the silicates and the other equally refractory materials. The slow relative velocities of silicate grains in nearly identical orbits resulted in low-energy collisions that promoted grain growth. Eventually, a hierarchy of planetesimal sizes developed; there were probably as many as 100 the size of the Moon, 10 the size of Mercury, and several as large as Mars. However, during the accretion process, most of the large planetesimals became incorporated into Venus and Earth. When the forming planets became massive enough, internal heat that was generated by radioactive decay, together with energy released during collisions, started the process of gravitational separation. With the formation of the massive Jupiter just beyond 5 AU from the Sun, gravitational perturbations began to influence the orbits of planetesimals in the region. In particular, most of the objects in the present-day asteroid belt had their orbits "pumped up" into progressively more and more eccentric orbits until some of them were absorbed by Jupiter or the other developing planets or were sent crashing into the Sun, while most were jettisoned from the solar system entirely; this process stole material from the feeding zone of Mars.

Long before the terrestrial planets finished "feeding" on planetesimals in their regions of the disk, however, the evolving Sun reached the stage of thermonuclear ignition in its core, initiating the T-Tauri phase. At this point, the infall of material from the disk was reversed by the strong stellar wind that ensued, and any gases and dust that had not yet collected into planetesimals were driven out of the inner solar system.

**QUESTION 10**

**What is the Eddington Luminosity? Give examples where it is important.**

### QUESTION 10

**What is the Eddington Luminosity? Give examples where it is important.**

Carroll & Ostlie (2007, pgs. 289-290) consider a cylinder that is filled with point particles that interact through perfectly elastic collisions (i.e. as an ideal gas). For such a scenario it can be shown using simple physics that the pressure exerted on the walls of the container is

$$P = \frac{1}{3} \int_0^\infty n_p p v d p, \quad (299)$$

where  $n_p dp$  is the number density of particles having momenta between  $p$  and  $p+dp$ . Physically, the  $1/3$  term arises since the total momentum of the collection of particles will be shared over three spatial dimensions. For the special case of massive, nonrelativistic particles, equation (299) can be integrated using  $p = mv$ , and taking  $n_\nu d\nu = n_p dp$  to be the Maxwell-Boltzmann velocity distribution for an ideal gas. In this case, we derive the familiar ideal gas law:

$$P_{\text{gas}} = nkT = \frac{\rho kT}{\mu m_H} \left( \text{where } \mu \equiv \frac{\bar{m}}{m_H} \right). \quad (300)$$

Because photons possess momentum  $p_\gamma = h\nu/c$ , they are capable of delivering an impulse to other particles during absorption or reflection. Hence, electromagnetic radiation results in another form of pressure, that is separate from the gas pressure derived above. Substituting the speed of light for the velocity in equation (299), using the expression for photon momentum, and using an identity for the distribution function,  $n_p dp = n_\nu d\nu$ , the general pressure integral for radiation becomes

$$P = \frac{1}{3} \int_0^\infty h\nu n_\nu d\nu. \quad (301)$$

The problem of determining photon pressure now reduces to finding an appropriate expression for  $n_\nu d\nu$ . Since photons are bosons, the Bose-Einstein distribution function would apply. However, the problem may also be solved realizing that  $n_\nu d\nu$  represents the number density of photons having frequencies lying in the range  $\nu$  and  $\nu+d\nu$ . Multiplying by the energy of each photon in that range would then give the energy density over the frequency integral; that is  $u_\nu d\nu = h\nu n_\nu d\nu$ . But the energy density distribution is found from the Planck function for blackbody radiation, which integrated over frequency yields

$$P_{\text{rad}} = \frac{1}{3} a T^4, \quad (302)$$

where  $a \equiv 4\sigma/c$  is known as the *radiation constant* (Carroll & Ostlie 2007, pg. 295). Thus, for a system containing both gas and photons, the total pressure is

$$P_{\text{tot}} = P_{\text{gas}} + P_{\text{rad}} = \frac{\rho k T}{\mu m_H} + \frac{1}{3} a T^4. \quad (303)$$

As can be seen from equation (303), if the temperature is sufficiently high and the gas density is low enough, it is possible for radiation pressure to dominate over the gas pressure in certain regions of the star, a situation that can occur in the outer layers of very massive stars. As we have previously seen, the pressure gradient produced by a plane-parallel atmosphere of blackbody radiation is

$$\frac{dP}{dr} = -\frac{\kappa\rho}{c} \frac{L}{4\pi r^2}. \quad (304)$$

However, we know that hydrostatic equilibrium demands that the pressure gradient satisfy

$$\frac{dP}{dr} = -G \frac{M\rho}{r^2}, \quad (305)$$

where  $M$  is the star's mass (assuming that we are dealing with the envelope of the star). Combining equations (304) and (305) and solving for luminosity, we have

$$L_{\text{Ed}} = \frac{4\pi G c}{\kappa} M. \quad (306)$$

$L_{\text{Ed}}$  is the maximum radiative luminosity that a star can have and still remain in hydrostatic equilibrium. If the luminosity exceeds this value, mass loss must occur, driven by radiation pressure. This luminosity maximum, known as the **Eddington Luminosity**, appears in a number of areas of astrophysics, including late stages of stellar evolution, novae, and the structure of accretion disks (Carroll & Ostlie 2007, pg. 341).

Before discussing more, we can think a moment about the physical interpretation of equation (306). Firstly,  $L_{\text{Ed}}$  scales with  $M$  since more massive stars have a larger surface gravity and can therefore handle a larger radiation pressure force. In addition,  $L_{\text{Ed}}$  scales inversely with  $\kappa$  since it is the opacity that determines the coupling of radiation to gas as absorption or scattering is required to transfer some portion of the photon momentum to the surrounding medium; a larger opacity indicates stronger coupling and thus an enhanced radiation pressure force. Hence, the effect of radiation is to destabilize the atmosphere. If both the luminosity and opacity are high enough, nothing prevents the material from leaving the star and being accelerated to velocities well in excess of escape. We thus expect a strong outward radiation force produced in the atmospheres of high-luminosity stars when the stellar opacity is high, either through lines or continuum scattering (Shore 2003, pg. 245).

### STELLAR WINDS

We begin by describing the atmosphere of a star which may or may not be in hydrostatic equilibrium. If we suppose that the atmosphere is plane-parallel, thin layer with surface gravity  $g$  then the steady state momentum equation is

$$\rho v \frac{dv}{dz} = -\frac{dP}{dz} + \rho a, \quad (307)$$

where  $v$  is the gas velocity,  $a$  is any (generalized) acceleration (negative in the same sense as the surface gravity), and  $P = P_{\text{gas}} + P_{\text{rad}}$ . The radiation pressure gradient produces an outward acceleration  $g_{\text{rad}}$  which is the driving mechanism for making a stellar wind possible. If  $a$  differs only slightly from  $g$ , the radiative acceleration can still have a small effect on the atmosphere, but won't necessarily produce a net outflow. Radiative transfer of momentum to individual particles will induce an upward-directed Brownian drift<sup>58</sup>. However, when radiation pressure is sufficiently great,  $g$  must be replaced by  $a = g_{\text{eff}} = -g + g_{\text{rad}}$ . Then the hydrostatic equilibrium condition changes completely. Within the atmosphere the radiative acceleration is depth-dependent because of the opacity and change in the SED; the driving force will come to a maximum where the opacity and radiative flux are highest, and there may be some depth in the atmosphere where  $g = g_{\text{rad}}$ . At this depth the flow will become supersonic,  $v = c_s$ . Beyond that, since the radiative acceleration can continue to grow outward, the velocity gradient will be positive and the flow accelerates. In other words, once you reach the Eddington luminosity, nothing can stop the flow from reaching *at least* the escape velocity (Shore 2003, pg. 246).

Another potential driving mechanism for stellar winds involves thermal evaporation. The idea is simply that the high-energy tail of the Maxwell distribution causes a small fraction of the particles to have velocities exceeding that required for escape; if the mean thermal speed is approximately  $v_{\text{esc}}$  then a wind will result. For the low temperatures usually encountered in stellar atmospheres, there is no dynamical outflow caused by thermal pressure excesses. However, this process is attributed to the existence of the **solar wind**, an observed outflow reaching speeds on the order of  $v_{\text{esc}}$  with very low associated mass loss rate, of the order  $10^{-14} M_{\odot} \text{ yr}^{-1}$ . We know that this flow cannot be associated with the photosphere since its temperature is simply insufficient to drive evaporation, and radiation pressure fails to provide sufficient driving for a wind. Instead, the mass loss is associated with the **chromosphere**, an emission line forming region above the photosphere. The line cores coming from the chromosphere reflect an increase in the source function due to collisional populations of the upper states of the transition relative to the photosphere and thus indicates an enhanced temperature contrast. The most probable mechanism for heating the outer layers of the Sun and other cool stars is that in situ magnetic field annihilation in the complex fields of the coronal regions. This reconnection process releases the energy needed to create the extreme temperatures of  $T \sim 10^6 \text{ K}$  seen in the chromosphere and corona. Another possible mechanism is that acoustic (sound) waves generated by the turbulence of the upper convection zone steepen into shocks as they propagate into the upper atmosphere. These are strongly damped by viscosity and radiation and ultimately don't get very far, although that dissipation may be important in the chromospheric temperature rise. Another candidate that produces many of the same effects with longer propagation lengths is **Alfvén wave** heating, which may be important in late-type high-luminosity stars, where the connection with convection comes from the stirring required for the generation of the waves in the first place (Shore 2003, pg. 252).

Let's explore more about Alfvén waves. Simply stated, an Alfvén wave in a plasma is a low-frequency travelling oscillation of the ions and the magnetic field. The ion mass density provides the inertia and the magnetic field line tension provides the restoring force (Wikipedia). More specifically, this is a type of MHD wave generated by a magnetic pressure gradient. A magnetic field produces a pressure that is numerically equivalent to the magnetic energy density,  $P_m = B^2/2\mu_0$ . When a magnetic field line gets displaced by some amount perpendicular to the direction of the line, a pressure gradient is established; the pressure in the direction of displacement increases as induced by an increase in the number density of field lines, while at the same time the pressure in the opposite direction decreases. This pressure change then tends to push the line back again, restoring the original density of field lines. This process is analogous to the oscillations that occur in a guitar string when a portion of the string is displaced; it is the tension in the string that pulls it back when it is plucked. As with the traveling motion of a wave on a string, a disturbance in the magnetic field line can also propagate down the line, known as an Alfvén wave. The speed of propagation can be estimated in analogy to the sound speed of gas, so that  $v_A \sim \sqrt{P_m/\rho} \sim B/\sqrt{\mu_0\rho}$ . Since Alfvén waves propagate along magnetic field lines, they may also transport energy outward. Since a time-varying magnetic field induces an electric field, electrical currents in the highly conductive plasma will be drawn, implying that resistive heating may result; this is the heating source mentioned above (Carroll & Ostlie 2007, pg. 379).

When describing stellar wind from rotating stars it is important to consider magnetic fields. As the density of the stellar atmosphere drops, the magnetic field more effectively controls gas dynamics because its energy density drops more slowly, so in order for the wind to truly escape the star, its speed must not only exceed  $c_s$  but also  $v_A$ . The field has an additional effect on the wind if it is attached to a rotating underlying star. By the strong coupling when  $v < v_A$ , the field constrains the wind to (more or less) rigidly corotate with the surface. This means that the star loses angular momentum to the wind and ultimately spins down at the expense of the mass loss; this happens because generally  $v_A > v$  in the atmosphere. Even for comparatively weak fields like those present in the solar wind, the field plays the dominant role in torquing down the solar rotational velocity (Shore 2003, pg. 250).

<sup>58</sup> This process can create elemental separation within a star whereby differential radiative acceleration can act on a species depending on their spectral line distribution. Certain species will then receive kicks depending on their opacity and may move upward in the atmosphere. This driving eventually stops when the species reaches a point where increased temperature ionizes it and its coupling to radiation ceases (Shore 2003, pg. 248)

**QUESTION 12**

**State the central temperature of the Sun.**

### QUESTION 12

State the central temperature of the Sun.

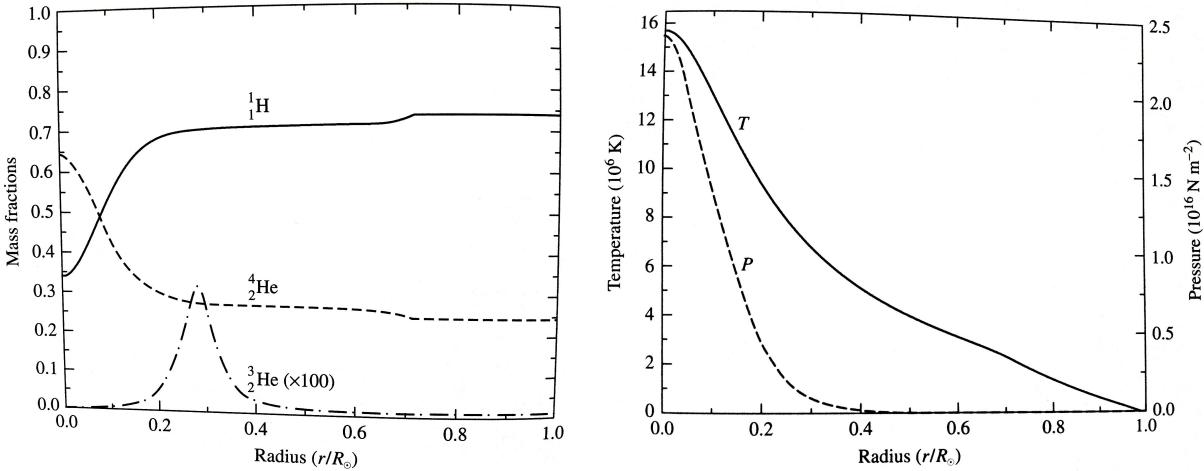


FIG. 86.— (left) The abundances of  ${}^1\text{H}$ ,  ${}^3\text{He}$ , and  ${}^4\text{He}$  as a function of radius for the Sun. Note that the abundance of  ${}^3\text{He}$  is multiplied by 100 for clarity. (right) The temperature and pressure profiles in the solar interior. Images taken from Carroll & Ostlie (2007).

### SOLAR INTERIOR

Based on its observed luminosity and effective temperature, the Sun is classified as a typical MS star of spectral class G2 with a surface composition of  $X = 0.74$ ,  $Y = 0.24$ , and  $Z = 0.02$  (the mass fractions of hydrogen, helium, and metas, respectively). The Sun has been converting hydrogen to helium via the pp chain during most of its lifetime, thereby changing its composition and internal structure. By comparing the results of radioactive dating tests of Moon rocks and meteorites (most specifically the CAIs) with stellar evolution calculations and the present-day observable Sun, the current age of the Sun is determined to be  $\sim 4.57$  Gyr. Furthermore, the Sun's luminosity has increased nearly 50% and its radius has increased 15% from their initial values; its effective temperature has also increased by about 150 K (Carroll & Ostlie 2007, pg. 350).

TABLE 9  
CENTRAL CONDITIONS IN THE SUN

Temperature	$1.57 \times 10^7 \text{ K}$
Pressure	$2.34 \times 10^{16} \text{ N m}^{-2}$ ( $2.3 \times 10^{11} \text{ atm}$ )
Density	$1.53 \times 10^5 \text{ kg m}^{-3}$ ( $153 \text{ g cm}^{-3}$ )
$X$	0.3397
$Y$	0.6405

Numerical models of the Sun can be constructed based on principles we will describe shortly. Table 9 gives the values of the central temperature, pressure, density and composition for one such solar model. According to the evolutionary sequence leading to this model, during its lifetime the mass fraction of hydrogen in the Sun's centre has decreased from its initial value of 0.71 to 0.34, while the central mass fraction of helium has increased from 0.27 to 0.64. In addition, due to diffusive settling of elements heavier than hydrogen, the mass fraction of hydrogen near the surface has increased by 0.03 while the mass fraction of helium has decreased by 0.03 (Carroll & Ostlie 2007, pg. 351).

Because of the Sun's past evolution, its composition is no longer homogeneous, but instead shows the influence of ongoing nucleosynthesis, surface convection, and elemental diffusion (settling of heavier elements). The composition structure of the Sun is shown in Figure 86 for  ${}^1\text{H}$ ,  ${}^3\text{He}$ , and  ${}^4\text{He}$ . Since the Sun's primary energy production mechanism is the pp chain,  ${}^3\text{He}$  is an intermediate species in the reaction sequence. During the conversion of hydrogen to helium,  ${}^3\text{He}$  is produced and then destroyed again. At the top of the hydrogen-burning region where the temperature is lower,  ${}^3\text{He}$  is relatively more abundant because it is produced more easily than it is destroyed<sup>59</sup>. At greater depths, the higher temperatures allow the  ${}^3\text{He}-{}^3\text{He}$  interaction to proceed more rapidly, and the  ${}^3\text{He}$  abundance again decreases (for comparison, the temperature and pressure profiles within the Sun are displayed in Figure 86). The slight ramp in the  ${}^1\text{H}$  and  ${}^4\text{He}$  curves near  $0.7 R_\odot$  reflects evolutionary changes in the position of the base of the surface convection zone. Within the convection zone, turbulence results in essentially complete mixing and a homogenous composition (Carroll & Ostlie 2007, pg. 352).

<sup>59</sup> This is because much higher temperatures are required for helium-helium interactions than proton-proton interactions.

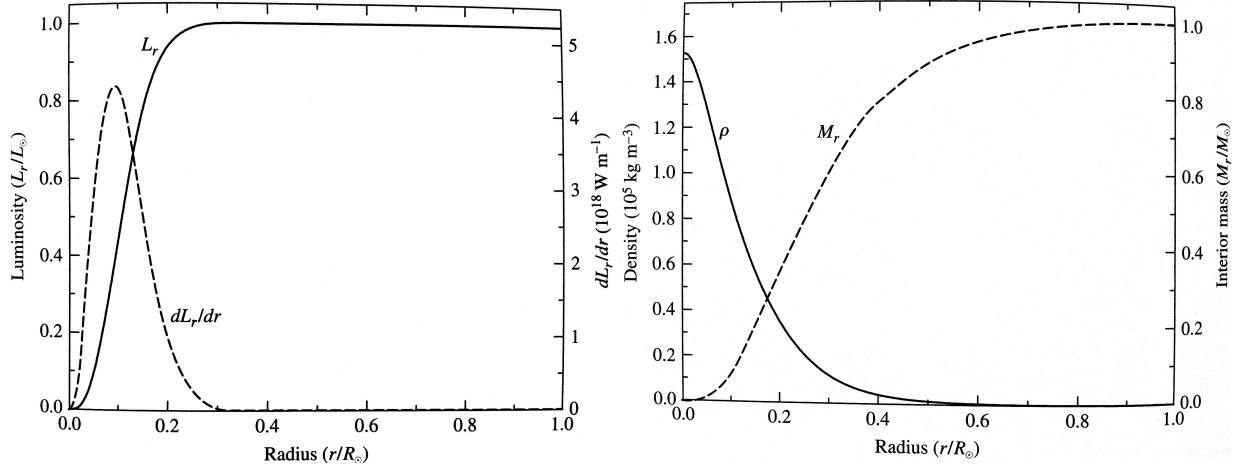


FIG. 87.— (left) The interior luminosity profile of the Sun and its derivative as a function of radius. (right) The density profile and the interior mass of the Sun as a function of radius. Images taken from Carroll & Ostlie (2007).

The largest contribution to the energy production in the Sun occurs at roughly one-tenth of the solar radius, as shown in Figure 87. If this result seems unexpected, consider the mass conservation equation,  $dM_r = 4\pi r^2 \rho dr = \rho dV$ , which indicates that the amount of mass within a certain radius interval increases with radius simply because the volume of a spherical shell increases with  $r$  for fixed choice of  $dr$ . Of course, the mass contained in the shell also depends on the density of the gas. Consequently, even if the amount of energy liberated per kg of material ( $\epsilon$ ) decreases steadily from the centre outward, the largest contribution to the total luminosity will occur, not at the centre, but in a shell that contains a significant amount of mass. In the case of the middle-aged Sun, the decrease in the amount of available hydrogen fuel at its centre will also influence the location of the peak in the energy production region (Carroll & Ostlie 2007, pg. 352).

Figures 86 and 87 show just how rapidly the pressure and density change with radius in the Sun. These variations are forced on the solar structure by the condition of hydrostatic equilibrium, the ideal gas law, and the composition structure of the star. Of course, boundary conditions applied to the stellar structure equations require that both  $\rho$  and  $P$  become negligible at the surface. Figure 87 also shows the interior mass as a function of radius. Notice that 90% of the mass of the star is located within roughly one-half of its radius. This should not come as a surprise since the density increases significantly as the centre of the Sun is approached (Carroll & Ostlie 2007, pg. 354).

#### ESTIMATE OF CENTRAL TEMPERATURE AND PRESSURE

The hydrostatic condition,

$$\frac{dP}{dr} = -\frac{Gm\rho}{r^2}, \quad (308)$$

along with an equation of state for an ideal gas,

$$P = \frac{\rho k T}{\mu m_H}, \quad (309)$$

enable us to estimate the pressure and temperature in the interior of a star of given mass and radius (Kippenhahn & Weigert 1990, pg. 8). Let us replace the left-hand side of equation (308) by an average pressure gradient,  $(P_S - P_C)/(R_S - R_C)$ , where  $R_C = 0$ ,  $R_S = R_*$  and we assume that the pressure at the surface of the star is zero,  $P_S = 0$ . Then by taking  $\rho = 3m/4\pi r^2$  on the right-hand side of equation (308) and replacing  $m$  and  $r$  by rough mean values  $M_*/2$  and  $R_*/2$ , we obtain

$$\frac{dP}{dr} \sim -\frac{P_C}{R_*} = -\frac{3Gm^2}{4\pi r^5} \rightarrow P_C \approx \frac{6GM_*^2}{\pi R_*^4}. \quad (310)$$

Evaluating this for the case of the Sun,  $M_* = M_\odot = 1.99 \times 10^{30}$  kg and  $R_* = R_\odot = 6.96 \times 10^8$  m, yields  $P_C = 2.15 \times 10^{15}$  N m<sup>-2</sup> =  $2.1 \times 10^{10}$  atm. From Table 9, we see that this is about an order of magnitude smaller than the actual value; the reason for this is the increased density profile near the centre of the Sun. To obtain a more accurate value, we would need to integrate equation (308) from the surface to the centre, taking into consideration the change in the interior mass  $m(r)$  at each point, together with the variation of the density with radius. Actually carrying out the integration requires functional forms of  $m(r)$  and  $\rho(r)$  where, of course, such explicit expressions are not available based on simple physical models (Carroll & Ostlie 2007, pg. 287).

The central temperature can now be estimated by combining equations (309) and (310), to give

$$T_C = \frac{P_C \mu m_H}{\rho C k} \approx \frac{6GM_*^2}{\pi R_*^4} \frac{\mu m_H}{\rho C k} = \frac{8GM_*}{R_*} \frac{3M_*}{4\pi R_*^3} \frac{\mu m_H}{\rho C k} \rightarrow T_C \approx 8 \frac{GM_*}{R_*} \frac{\bar{\rho}}{\rho C} \frac{\mu m_H}{k}. \quad (311)$$

Since in most stars the density increases monotonically from the surface to the centre, we have  $\bar{\rho}/\rho_C < 1$  (numerical simulations

show that  $\bar{\rho} \approx 0.03$ ) then we have that

$$T_C \lesssim 8 \frac{G\mu m_H}{k} \frac{M_*}{R_*}. \quad (312)$$

Evaluating this for the Sun with  $\mu = 0.62$  appropriate for complete ionization yields  $T_C = 1.1 \times 10^8$  K, about an order of magnitude larger than the actual value. So we can expect to encounter enormous pressures and very high temperatures in the central regions of stars. Note also that our assumption of an ideal gas turns out to be fully justified for such large values of  $P$  and  $T$  (Kippenhahn & Weigert 1990, pg. 9). In addition, our omission of radiation pressure in these estimates is justified since at  $T = 10^7$  K, radiation pressure is only about 0.07% that of gas pressure (Carroll & Ostlie 2007, pg. 296).

#### NUMERICAL SOLUTIONS

Using  $m$  as the independent variable, the basic differential equations for a spherically symmetric star are:

$$\begin{aligned} \frac{dr}{dm} &= \frac{1}{4\pi r^2 \rho}, \\ \frac{dP}{dm} &= -\frac{Gm}{4\pi r^4} - \frac{1}{4\pi r^2} \frac{d^2 r}{dt^2}, \\ \frac{dL}{dm} &= \epsilon_{\text{nuclear}} - \epsilon_{\text{neutrino}} + \epsilon_{\text{gravity}} \\ \frac{dT}{dm} &= -\frac{GmT}{4\pi r^4 P} \nabla, \end{aligned} \quad (313)$$

where  $\nabla \equiv d\ln T / d\ln P$ . If the energy transport is due to radiation (and conduction), then  $\nabla$  is to be replaced with

$$\nabla_{\text{rad}} = \frac{3}{16\pi acG} \frac{\kappa LP}{m T^4}. \quad (314)$$

On the other hand, if the energy is carried by convection, then  $\nabla$  has to be replaced by a value obtained from a proper theory of convection; this may be  $\nabla_{\text{ad}}$  in the deep interior, or obtained from a numerical solution of superadiabatic convection in outer layers. Note that the right-hand side of the  $dT/dm$  term assumes hydrostatic equilibrium. This does not matter in the case of radiative transport, since the local adjustment time of the radiation field is very short, and typical convection theory assumes hydrostatic equilibrium. Otherwise, another convection theory valid in rapidly changing regions would have to be used (Kippenhahn & Weigert 1990, pg. 64).

There are three types of time derivatives in equation (313); to each of them belongs a certain characteristic timescale. These timescales are  $\tau_{\text{hydro}}$ ,  $\tau_{\text{KH}}$ , and  $\tau_{\text{nuclear}}$  (which also regulates the timescale for chemical changes in the star). If the evolution of the star is governed by thermal adjustment or by nuclear reactions, the inertia term  $d^2r/dt^2$  can be ignored and the  $dP/dm$  term reduces to the ordinary hydrostatic equilibrium condition,

$$\frac{dP}{dm} = -\frac{Gm}{4\pi r^4}, \quad (315)$$

since  $\tau_{\text{KH}} \gg \tau_{\text{hydro}}$  and  $\tau_{\text{nuclear}} \gg \tau_{\text{hydro}}$ . Furthermore, if the star evolves on a timescale  $\tau_{\text{nuclear}} \gg \tau_{\text{KH}}$ , then the  $dL/dm$  term reduces to

$$\frac{dL}{dm} = \epsilon_{\text{nuclear}} - \epsilon_{\text{neutrino}}. \quad (316)$$

If both conditions are met, the star will evolve along a sequence of states in which it is not only in hydrostatic equilibrium but also thermally adjusted; we call this **complete equilibrium** (i.e. both mechanical and thermal equilibrium). In this case, the set of equations in (313) contain only spatial derivatives, which can be independently coupled to the chemical evolution of the star,  $dX_i/dt$ . Complete equilibrium is a good approximation for stars in many important evolutionary phases, for example stars on the MS (Kippenhahn & Weigert 1990, pg. 66).

The basic stellar equations in (313) require information concerning the physical properties of the matter from which the star is made. The required conditions are the equations of state of the material and are collectively referred to as **constitutive relations**. Specifically, we need relationships for the pressure, opacity, and the energy generation rate, in terms of fundamental characteristics of the material: the density, temperature, and composition. In general, we require  $P = P(\rho, T, X_i)$ ,  $\kappa = \kappa(\rho, T, X_i)$ ,  $\epsilon = \epsilon(\rho, T, X_i)$ . The pressure equation of state can be quite complex in the deep interiors of certain classes of stars, where the density and temperature can become extremely high. However, in most situations, the ideal gas law, combined with the expression for radiation pressure, is a good first approximation, particularly when the variation in the mean molecular weight with composition and ionization is properly calculated. The opacity of stellar material cannot be expressed exactly by a single formula. Instead, it is calculated explicitly for various compositions at specific densities and temperatures and interpolated from a lookup table. To calculate the nuclear energy generation rate, we can use analytic formulae for the pp chain and CNO cycle (Carroll & Ostlie 2007, pg. 331).

The actual solution of the stellar structure equations, including the constitutive relations, requires appropriate boundary conditions that specify physical constraints to the mathematical equations. The central boundary conditions are fairly obvious – namely that the interior mass and luminosity must approach zero at the centre of the star:  $M_r \rightarrow 0$ ,  $L_r \rightarrow 0$  as  $r \rightarrow 0$ . This simply

means that the star is physically realistic and does not contain a hole, a core of negative luminosity, or central points of infinite  $\rho$  or  $\epsilon$ . A second set of boundary conditions is required at the surface of the star. The simplest set of assumptions is that the temperature, pressure, and density all approach zero at some surface value for the star's radius:  $T \rightarrow 0$ ,  $P \rightarrow 0$ , and  $\rho \rightarrow 0$  as  $r \rightarrow R_*$ . Strictly, these surface conditions will never be obtained in a real star. Therefore, it is often necessary to use more sophisticated surface boundary conditions, such as when the star being modelled has an extended atmosphere or is losing mass, as most stars do (Carroll & Ostlie 2007, pg. 332).

Given the basic stellar structure equations, constitutive relations, and boundary conditions, we can now specify the type of star to be modelled. As can be seen from equations (313), the pressure gradient,  $dP/dr$ , at a given radius is dependent on the interior mass and the density. Similarly, the radiative temperature gradient,  $dT/dr$ , depends on the local temperature, density, opacity, and interior luminosity, while the luminosity gradient,  $dL/dr$ , is a function of the density and energy generation rate. The pressure, opacity, and energy generation rate in turn depend explicitly on the density, temperature, and composition at that location. If the interior mass at the surface of the star (i.e.  $M_*$ ) is specified, along with the composition, surface radius  $R_*$ , and luminosity  $L_*$ , application of the surface boundary conditions allows for a determination of the pressure, interior mass, temperature, and interior luminosity at an infinitesimal distance  $dr$  below the surface of the star. Continuing this numerical integration of the stellar structure equations to the centre of the star must result in agreement with the central boundary conditions (this type of numerical approach, which uses **difference equations** in place of differential equations, can also be carried out by moving from the centre to the surface). Since the values of the various gradients are directly related to the composition of the star, it is not possible to specify any arbitrary combination of surface radius and luminosity after the mass and composition have been selected. This set of conditions is known as the **Vogt-Russell theorem**, which states that the mass and composition structure throughout a star uniquely determine its radius, luminosity, and internal structure, as well as its subsequent evolution. In addition, the dependence of a star's evolution on mass and composition is a consequence of the change in composition due to nuclear burning. The statement of this theorem is somewhat misleading since there are other parameters that influence stellar interiors, such as magnetic fields and rotation, though these parameters are assumed to have little effect in most stars (Carroll & Ostlie 2007, pg. 333).

Obviously, it is not generally possible to solve the system of stellar structure equations and their associated constitutive relations analytically. However, under very special and restrictive situations, it is possible to find analytic solutions to a subset of the equations. To see this, note that the first two equations in (313) (the so-called *mechanical* equations) can be solved simultaneously without reference to the *energy* equations (the latter two in the set) if only a simple relationship existed between pressure and density. Of course, such a simple relationship does not generally exist; normally, temperature and composition must also enter into the pressure equation of state, often in a complicated way. However, under certain circumstances, such as for an adiabatic gas, the pressure can be written explicitly in terms of the density alone. Hypothetical stellar models in which the pressure depends on density in the form  $P = K\rho^\gamma$  are known as **polytropes**. Under this approximation, the **Lane-Emden equation** can be derived, which provides an analytic description of stellar interiors (Carroll & Ostlie 2007, pg. 335).

#### TAYLOR EXPANSIONS

It is often useful to know the behaviour of  $r$ ,  $L$ ,  $P$ , and  $T$  in the vicinity of the centre of the star,  $m \rightarrow 0$ . The equation of continuity can be manipulated to show how  $r$  varies with  $m$  around the centre:

$$d(r^3) = \frac{3}{4\pi\rho} dm \rightarrow r = \left( \frac{3}{4\pi\rho_C} \right)^{1/3} m^{1/3}, \quad (317)$$

where the second identity arises from integrating the continuity equation for constant  $\rho = \rho_C$  (i.e. for small enough values of  $r$  and  $m$  that  $\rho$  does not change much). This can be considered the first term in a series expansion of  $r$  around  $m = 0$ . A corresponding integration of the energy equation yields

$$L = (\epsilon_{\text{nuclear}} - \epsilon_{\text{neutrino}} + \epsilon_{\text{gravity}})m. \quad (318)$$

In both of these cases we have used the proper boundary conditions ( $m \rightarrow 0$  and  $L \rightarrow 0$  as  $r \rightarrow 0$ ) by taking the integration constants to be zero (Kippenhahn & Weigert 1990, pg. 68).

Eliminating  $r$  for small values of  $m$  by equation (317), we obtain from the hydrostatic equation

$$\frac{dP}{dm} = -\frac{G}{4\pi} \left( \frac{4\pi\rho_C}{3} \right)^{4/3} m^{-1/3} \rightarrow P - P_C = -\frac{3G}{8\pi} \left( \frac{4\pi\rho_C}{3} \right)^{4/3} m^{2/3}. \quad (319)$$

The pressure gradient must of course vanish at the centre, which can be seen by writing the hydrostatic equation in the form

$$\frac{dP}{dr} \sim \frac{m}{r^2} \sim \frac{r^3}{r^2} \rightarrow 0 \quad (320)$$

for  $r \rightarrow 0$  (Kippenhahn & Weigert 1990, pg. 69).

We can consider the variation of temperature for the radiative case, for which we have

$$\frac{dT}{dm} = -\frac{3}{64\pi^2 ac} \frac{\kappa L}{r^4 T^3}. \quad (321)$$

With  $P \rightarrow P_C$  and  $T \rightarrow T_C$ ,  $\kappa$  tends to some well-defined value  $\kappa_C$ . Replacing  $L$  with  $m$  from equation (318) and  $r$  with  $m^{1/3}$  from

equation (317), we can integrate equation (321) for small values  $m$  to obtain

$$T^4 - T_C^4 \propto m^{2/3}. \quad (322)$$

In the case of adiabatic convection we can use  $\nabla = \nabla_{\text{ad}}$  and our previous results to obtain

$$\ln T - \ln T_C \propto m^{2/3}. \quad (323)$$

The constants in front of these expressions can be found in Kippenhahn & Weigert (1990, pg. 69).

**QUESTION 13**

**Which have higher central pressure, high-mass or low-mass main-sequence stars? Roughly, what is their mass-radius relation? Derive this.**

### QUESTION 13

Which have higher central pressure, high-mass or low-mass main-sequence stars? Roughly, what is their mass-radius relation? Derive this.

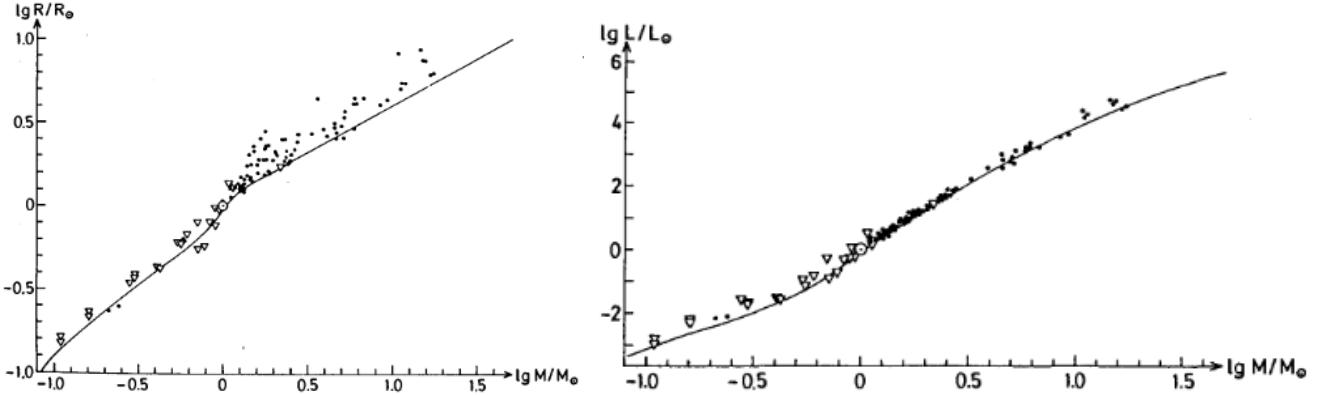


FIG. 88.— (left) The line shows the mass-radius relation for analytic models of the ZAMS. For comparison, the best measurements of MS components of detached (dots) and visual (triangles) binary systems are indicated. (right) The line gives the mass-luminosity relation for the same MS models, with points having the same meaning as before. Images taken from Kippenhahn & Weigert (1990).

When describing MS stars it is useful to model them as being chemically homogeneous and in complete (mechanical and thermal) equilibrium. Central hydrogen burning provides a long-lasting energy source, and consequently the stars change only on the very long nuclear timescale  $\tau_{\text{nuclear}}$ . Within the much shorter Kelvin-Helmholtz timescale the stars will *forget* the details of their thermal history long before the nuclear reactions have noticeably modified the composition. This is why one can reasonably treat them as homogeneous models in thermal equilibrium. The beginning evolution, in which hydrogen is slowly consumed in the stellar core, has such a long duration that most visible stars are presently found in this phase. This beginning phase is called the **zero-age main sequence (ZAMS)**, since one usually counts the age of a star from this point on. The ZAMS corresponds more or less with the lower border of the observed MS band (Kippenhahn & Weigert 1990, pg. 207).

As we have already seen, describing the interior of stars is a matter of solving the stellar evolution equations. For an order of magnitude approximation, these equations can be manipulated as follows:

$$\begin{aligned} \frac{dr}{dm} &= \frac{1}{4\pi r^2 \rho} \Rightarrow M \propto R^3 \rho, \\ \frac{dP}{dm} &= -\frac{Gm}{4\pi r^4} \Rightarrow P \propto \frac{M^2}{R^4}, \\ \frac{dL}{dm} &= \epsilon \Rightarrow L \propto M \epsilon, \\ \frac{dT}{dm} &= -\frac{3}{64\pi^2 ac} \frac{\kappa L}{r^4 T^3} \Rightarrow T^4 \propto \frac{ML}{R^4}. \end{aligned} \quad (324)$$

In the second identity we have assumed hydrostatic equilibrium and in the final identity we have assumed that radiation is the dominant mode of energy transport in the star. We can also make use of the ideal gas law,  $P \propto \rho T \propto MT/R^3$ , which is valid over most of the MS, except at the lowest masses (Charles). Combining the second identity,  $P \propto M^2/R^4$  with  $P \propto MT/R^3$  gives

$$T \propto \frac{M}{R}, \quad (325)$$

which is simply a reflection of the virial theorem where thermal energy  $\sim$  gravitational potential energy (Kippenhahn & Weigert 1990, pg. 196). Furthermore, the relation that  $T^4 \propto ML/R^4$  implies that

$$L \propto M^3. \quad (326)$$

This compares well with the empirical relation  $L \propto M^{3.2}$  (Kippenhahn & Weigert 1990, pg. 209). For temperatures around  $10^7$  K, nuclear energy production per unit mass follows the form  $\epsilon_{\text{pp}} \propto T_C^4$  and  $\epsilon_{\text{CNO}} \propto T_C^{20}$  for the pp chain and CNO cycle, respectively (Carroll & Ostlie 2007, pg. 311). From our relations that  $L \propto \epsilon M$  and  $L \propto M^3$ , this suggests that  $\epsilon \propto M^2$ , which tells us that  $T_C^{\text{pp}} \propto M^{1/2}$  and  $T_C^{\text{CNO}} \propto M^{1/10}$  for the pp chain and CNO cycle, respectively. The MS contains stars with masses ranging from  $0.08 M_\odot$  up to around  $100 M_\odot$  with the transition from pp to CNO burning occurring around  $1 M_\odot$ . Therefore, central temperature span less than an order of magnitude in value across the entire MS, at least according to our simple scaling relations.

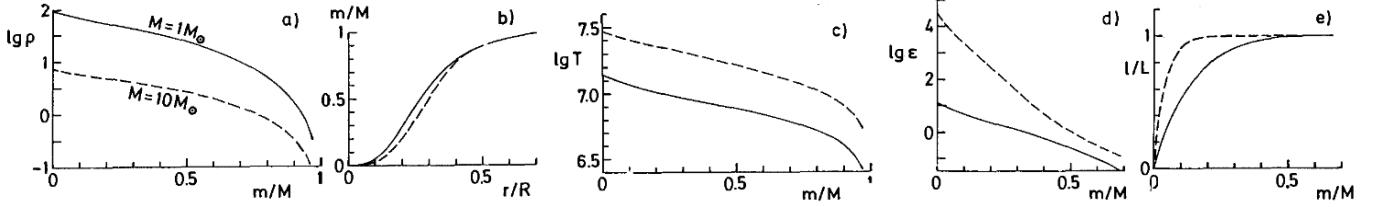


FIG. 89.— The run of some functions inside ZAMS models for  $M = 1 M_\odot$  (solid lines) and  $M = 10 M_\odot$  (dashed lines): (a)  $\rho$  in  $\text{g cm}^{-3}$ , (b) radial mass distribution  $m(r)$ , (c)  $T$  in K, (d) nuclear energy production in  $\text{erg g}^{-1} \text{s}^{-1}$ , (e) local luminosity. Image taken from Kippenhahn & Weigert (1990).

From equation (325) we have that  $M \propto T_C R$  which implies that

$$M \propto R, \quad (327)$$

for MS stars, assuming that  $T_C$  is roughly constant across the MS. Then, from the ideal gas law, we have that

$$P_C \propto M^{-2}. \quad (328)$$

Note that the same relation arises if we invoke the hydrostatic condition  $P \propto M^2/R^4$  instead. We therefore see that central pressure decreases with increasing mass along the MS, so that low-mass stars will have the largest central pressures. This should not be surprising since we know the the most massive stars have luminosities approaching the Eddington limit and thus barely hold themselves together, whereas low-mass stars like the Sun have no such problem (Charles).

Note a potential ambiguity arises in considering whether the temperature in equation (325) should refer to the central or surface temperature. If we instead use the surface temperature, then from the Stefan-Boltzmann equation we have that  $L \propto R^2 T^4$ , so that  $L \propto M^3$  and  $T \propto M/R$  imply that

$$M \propto R^2. \quad (329)$$

Coupling this result to the ideal gas law suggests that the central pressure varies as

$$P_C \propto M^{-1}. \quad (330)$$

This time, however, the same relation does not arise if we invoke the hydrostatic condition, which instead suggests that central pressure is independent of mass. The reason may be that the hydrostatic equation refers to some sort of average mass, rather than total mass of the system (Charles). In any event, we find that the radius scales either linearly or shallowly with mass, and pressure is inversely related to mass, so that we arrive at the same conclusions as before.

#### MORE ADVANCED MODELLING

When comparing different stellar models that are calculated under similar assumptions (concerning parameters or material functions) we might expect to find similarities in the solutions. Indeed, there is often a type of similarity between different solutions, called a **homology**, though the conditions for this are so severe that real stars will scarcely match them. Nevertheless, homology relations can still be useful in certain circumstances, such as in considering the MS, where homology relations can be used to transform solutions obtained at one mass to all other masses. For this case, when comparing different MS models (say of masses  $M$  and  $M'$  with radii  $R$  and  $R'$ ) we consider *homologous points* at which the relative radii are equal:  $r/R = r'/R'$ . We now speak of *homologous stars* if their mass shells ( $m/M = m'/M'$ ) are situated at homologous points. To be more precise, we can consider all radii as functions of the relative mass values  $\xi$ , which are the same for homologous masses:  $\xi \equiv m/M = m'/M'$ . We can then write the homology condition as

$$\frac{r(\xi)}{r'(\xi)} = \frac{R}{R'}, \quad (331)$$

for all  $\xi$ . In homologous stars the ratio of the radii  $r/r'$  for homologous mass shells is constant throughout the stars. Going from one homologous star to another, all homologous mass shells are compressed (or expanded) by the same factor  $R/R'$ . Note that therefore any two polytropic models of the same index  $n$  are homologous to each other (Kippenhahn & Weigert 1990, pg. 191).

For homologous models we can rewrite the stellar structure equations using derivatives with respect to  $d\xi$  instead of  $dm$ , and can transform the equations to other mass scales by supposing that quantities like  $P/P'$ ,  $T/T'$ ,  $L/L'$ , and  $\mu/\mu'$  are proportional among homology classes. Then, assuming a static, radiation-dominated star with a uniform composition, and power-law representations for  $\rho$ ,  $\epsilon$ , and  $\kappa$ , it is possible to show that  $(R/R') = (M/M')^{z_1}$ , where  $z_1$  ranges from 0.4 to 0.8, depending on whether the pp chain or CNO cycle is used. We also find that  $(P/P') = (M/M')^2/(R/R')^4$  so that  $P \propto M^{-0.4}$ , again stating that central pressure decreases with increasing mass (Charles).

The mass-radius relation can also be derived from polytropes. For any polytropic index  $0 \leq n \leq 5$  ( $n = 1/(\gamma - 1)$ ) it can be shown that  $R \propto \rho_C^{(1-n)/2n}$  while  $M \propto \rho_C^{(3-n)/2n}$  so that

$$R \propto M^{\frac{1-n}{3-n}}. \quad (332)$$

This is a useful solution in many cases (e.g., non-relativistic WDs), but in our case, the Sun is approximately an  $n = 3$  polytrope for which equation (332) fails (Charles).

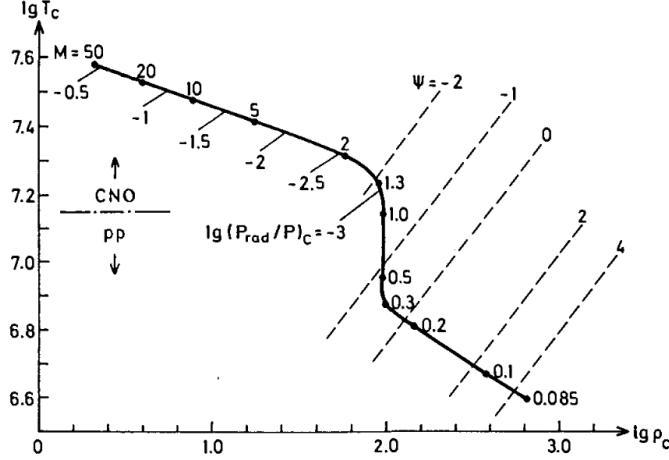


FIG. 90.— The heavy solid line gives the central temperature  $T_c$  (in K) over the central density  $\rho_c$  (in  $\text{g cm}^{-3}$ ) for ZAMS models. The dots give the positions of some models with masses between  $0.085$  and  $50 M_\odot$ . The labels below the curve indicate the fractional contribution of the radiation pressure  $P_{\text{rad}}$  to the total pressure in the centre. The dot-dashed line at the left gives roughly the border between dominating CNO-cycle and pp-chain reactions. The dashed lines give the constant degeneracy pressure  $\psi$  of the electron gas. Image taken from Kippenhahn & Weigert (1990).

#### EMPIRICAL RELATIONS

The scaling relations derived above are only useful for qualitative descriptions and order of magnitude approximations. If we wish to go further then we must rely on observations coupled to detailed theoretical treatments. In this case, we find that the mass-radius relation varies somewhat along the MS. If we define  $\xi$  such that  $R \propto M^\xi$  ( $\xi = 1$  above) then we find that  $\xi \approx 0.58$  and  $0.8$  in the upper and lower mass ranges of the MS, respectively. As shown in Figure 88, there is a pronounced maximum of the slope around  $1 M_\odot$ , indicating a remarkable deviation from homologous behaviour in this range. With decreasing effective temperature these models have outer convective zones of strongly increasing depth; this tends to decrease  $R$ . We will also define  $\eta$  such that  $L \propto M^\eta$  ( $\eta = 3$  above) and similarly plot the mass-luminosity relation in Figure 88. Over the mass range shown, the average of  $\eta$  is  $3.2$ ; its value in the low-mass ( $M \lesssim 10 M_\odot$ ) regime is  $\eta = 3.9$ , and is  $\eta = 3.3$  for larger masses. The decreasing slope towards larger mass is an effect of the increasing radiation pressure (Kippenhahn & Weigert 1990, pg. 209).

Let us consider the way in which the variation of the exponents  $\xi$  and  $\eta$  influences the slope of the MS in the HR diagram. The mass-radius and mass-luminosity relations can be combined as  $R \propto L^{\xi/\eta}$ . Introducing this into the Stefan-Boltzmann relation,  $L \propto R^2 T_{\text{eff}}^4$ , we obtain for the MS in the HR diagram the proportionality

$$L \propto T_{\text{eff}}^\zeta \quad \text{where } \zeta \equiv \frac{4}{1-2\xi/\eta}. \quad (333)$$

We have seen that for large stellar masses,  $\eta$  decreases and  $\xi$  remains about constant with further increasing  $M$ . Equation (333) then gives an increase of  $\zeta$ , which means that the MS must become gradually steeper towards high luminosities (Kippenhahn & Weigert 1990, pg. 209). With the simple scaling relations considered earlier, we derived  $\xi = 1$  and  $\eta = 3$ , suggesting that  $\zeta = 12$ . With the more realistic values of  $\xi = 0.69$  and  $\eta = 3.2$ , we have a more modest  $L \propto T_{\text{eff}}^7$ .

#### INTERIOR SOLUTIONS

The behaviour of stellar interiors can be illustrated by plotting characteristic variables as functions of  $m/M$ . This is done in Figure 89 for both  $1 M_\odot$  and  $10 M_\odot$  ZAMS models evaluated using homology relations. The density  $\rho$  increases appreciable towards the centre where we have  $\rho_c \sim 100 \text{ g cm}^{-3}$  for  $M = 1 M_\odot$ , roughly a factor of  $10^9$  larger than its value in the photosphere. For  $10 M_\odot$ , the central density is smaller by more than a factor of  $10$ . The inward increase in  $\rho$  indicates a very strong concentration of the mass elements towards the centre, as also illustrated in Figure 89. For  $1 M_\odot$ , the inner  $30\%$  of the radius (only  $3\%$  of the total volume) contains  $60\%$  of the mass; and in the outer  $50\%$  of its radius ( $88\%$  of the total volume) only about  $10\%$  of the mass can be found (Kippenhahn & Weigert 1990, pg. 210).

The temperature also increases towards the centre. For  $1 M_\odot$ , the central value of nearly  $1.5 \times 10^7 \text{ K}$  is a factor 2500 larger than the photospheric value. Values of  $T > 3 \times 10^6 \text{ K}$  extend to  $m \approx 0.95M$ , so that the average temperature (averaged over mass elements) is roughly  $8 \times 10^6 \text{ K}$ . In a  $10 M_\odot$  star,  $T$  has slightly more than twice the values of the corresponding mass elements for  $1 M_\odot$ . The behaviour of  $T$  is necessarily reflected by that of the rate of energy generation due to hydrogen burning. The dependence of  $\epsilon$  on  $T$ , together with the temperature gradient, yields a strong decrease of  $\epsilon$  from the centre outwards. In the  $1 M_\odot$  star,  $\epsilon$  has dropped by a factor of 100 from the centre to  $m = 0.6M$ , and still further outward it is quite negligible. This is particularly well seen by looking at the luminosity curve:  $90\%$  of  $L$  is generated in the inner  $30\%$  of  $M$  and it reaches  $99\%$  of  $L$  at  $m/M = 0.5$ . In the central part of the  $10 M_\odot$  star, where  $T_c = 3 \times 10^7 \text{ K}$ , the dominant energy source is the CNO cycle. The much larger  $T$  dependence of  $\epsilon$  gives an even more pronounced concentration of  $\epsilon$  towards the centre. Further outwards, where  $T$  is low enough for the pp chain to dominate, the slope of  $\epsilon$  becomes the same for both stars (Kippenhahn & Weigert 1990, pg. 211).

We have seen that in spite of all similarities there are characteristic differences between the interior solutions for different values of  $M$ . Some of these can be found in Figure 90 where we plot the central values of temperature and density. With increasing mass, there is a slight increase in  $T_C$  together with a substantial decrease in  $\rho_C$ . The striking change in the curve occurs around  $1 M_\odot$  where there is a clear deviation from homology. These deviations are connected partly with the changes of the central values and partly with those at the surface (especially  $T_{\text{eff}}$  and the depth of the outer convection zone). The extension of convection zones, for example, should certainly influence the centre, since they have a less pronounced mass concentration than radiative regions. Note that both flat parts of the  $T_C - \rho_C$  curve in Figure 90 belong to models in which the central part is convective (Kippenhahn & Weigert 1990, pg. 211).

In the upper range of mass, degeneracy is negligible, while it becomes increasingly important towards smaller  $M$  owing to the increasing density. Below  $0.5 M_\odot$ , other deviations from the ideal gas approximation also become important in the equation of state (e.g., electrostatic interaction between ions). On the other hand, radiation pressure must increase towards larger  $M$  owing to the increasing  $T$ , since  $P_{\text{rad}} \sim T^4$ . At  $1 M_\odot$ , radiation contributes only the negligible fraction of  $10^{-4}$  to the total central pressure. This fraction becomes about 1% at  $4 M_\odot$ , while in the centre of a  $50 M_\odot$  star,  $P_{\text{rad}}$  contributes about  $1/3$  of the total pressure. Another effect of the growing  $T_C$ , which also occurs around  $1 M_\odot$ , is the transition from the pp chain to the CNO cycle as the dominant energy source (Kippenhahn & Weigert 1990, pg. 212).

**QUESTION 14**

**Derive the scaling of luminosity with mass for a (mostly) radiative star. Do you need to know the source of energy for the star to derive this scaling?**

#### QUESTION 14

**Derive the scaling of luminosity with mass for a (mostly) radiative star. Do you need to know the source of energy for the star to derive this scaling?**

TABLE 10  
POWER-LAW INDICES

Energy Generation Mode	$\lambda$	$\nu$	Opacity Mode	$n$	$s$	Pressure Mode	$\chi_\rho$	$\chi_T$
pp-chain	1	4	Thomson Scattering	0	0	Ideal Gas	1	1
CNO cycle	1	20	Kramers' Law	1	3.5	Radiation	0	4
Triple- $\alpha$ Process	2	40						

*ENERGY GENERATION AND TRANSPORT*

To express the thermal balance occurring within stellar interiors quantitatively, consider a spherically symmetric shell of mass  $dM_r$  and thickness  $dr$ . Within that shell we denote the power generated per gram as  $\epsilon$  (erg g<sup>-1</sup> s<sup>-1</sup>), which we refer to as the *energy generation rate*. The total energy generated within the shell is thus  $4\pi r^2 \rho \epsilon = \epsilon dM_r$ . To balance the power generated, we must have a net flux of energy leaving the shell. If  $F(r)$  is the flux (in units of erg cm<sup>-2</sup> s<sup>-1</sup>), with positive values implying a radially directed outward flow, then  $L_r = 4\pi r^2 F(r)$  is the total power, or luminosity, in erg s<sup>-1</sup>, entering (or leaving) the shell's inner face, and  $L_{r+dr} = 4\pi r^2 F(r+dr)$  is the luminosity leaving through the outer face at  $r+dr$ . The difference of these two terms is the net loss or gain of power for the shell. For thermal balance that difference must equal the total power generated within the shell:

$$\frac{dL}{dr} = 4\pi r^2 \rho \epsilon \iff \frac{dL}{dm} = \epsilon. \quad (334)$$

Since, for now, we are considering only  $\epsilon \geq 0$ , then  $L_r$  must either be constant (in regions where  $\epsilon = 0$ ) or increase monotonically with  $r$  or  $m$ . For nuclear burning processes,  $\epsilon$  is a strong function of temperature and, because temperature is expected to decrease outward in a star,  $\epsilon$  should be largest in the inner stellar regions provided that fuel is present. Thus  $L_r$  should increase rapidly from the centre, starting from zero, and then level out to its surface value (Hansen et al. 2004, pg. 22).

Over sufficiently restricted ranges of temperature, density, and composition, the energy generation rate for nuclear fusion reactions can be written as a power-law of the form

$$\epsilon \propto \rho^\lambda T^\nu, \quad (335)$$

where the indices,  $\lambda$  and  $\nu$ , depend on the specific nuclear burning mechanism; these values are listed in Table 10. As important examples, consider briefly the two ways that stars burning hydrogen to helium: these are the proton-proton (pp) chains and the carbon-nitrogen-oxygen (CNO) cycles. The first is, for the most part, a simple sequence of nuclear reactions, starting with one involving two protons, that gradually add protons to intermediate reaction products to eventually produce helium. The second cycle uses, C, N, and O as catalysts to achieve the same ends. Table 10 lists the indices of the power-law relation of equation (335) for typical hydrogen-burning temperatures of  $T \sim 10^7$  K and densities of  $\rho \sim 100$  g cm<sup>-3</sup>. Also given are the values for the triple- $\alpha$  process which effectively combines three helium nuclei ( $\alpha$  particles) to make one carbon nucleus with a characteristic temperature of 10<sup>8</sup> K (Hansen et al. 2004, pg. 23).

Now that we have the right-hand side of equation (334) understood, we must investigate the other aspect of thermal balance; that is, what processes are determining  $L_r$ . As we know, there are three major modes of energy transport: radiation transfer, convection of hotter and cooler mass elements, and heat conduction, with the first two being most important for most stars (WDs depend heavily on the last mode, but those stars are in a class by themselves) (Hansen et al. 2004, pg. 23).

Fortunately, radiation transfer is easily described since, except for the very outermost stellar layer, the energy flux carried by photons obeys a Fick's Law of diffusion; that is, the flow is driven by a gradient of a quantity having something to do with the radiation field. The form is:

$$F(r) = -\mathcal{D} \frac{d(aT^4)}{dr}, \quad (336)$$

where  $aT^4$  is the radiation energy density and  $\mathcal{D} = c/3\kappa\rho$  is the diffusion coefficient. The opacity term  $\kappa$  in  $\mathcal{D}$  arises as it describes how the flow of radiation is hindered by the medium through which it passes. With the usual conversion from flux to luminosity, we can rewrite equation (336) as

$$L_r = \frac{16\pi acr^2}{3\kappa\rho} T^3 \frac{dT}{dr} \iff L_r = \frac{64\pi^2 acr^4}{3\kappa} T^3 \frac{dT}{dm}. \quad (337)$$

To simplify the use of equation (337) we will assume that the opacity follows a power-law of the form

$$\kappa \propto \rho^n T^{-s}, \quad (338)$$

where the indices,  $n$  and  $s$ , depend on the specific opacity mechanism. Important examples are electron Thomson scattering ( $n=s=0$ ), which is important for completely ionized stellar regions, and Kramer's opacity ( $n=1, s=3.5$ ), which is characteristic

of radiative processes involving atoms (Hansen et al. 2004, pg. 24). The power-law indices for these two processes are listed in Table 10.

#### STELLAR DIMENSIONAL ANALYSIS

Some texts on stellar evolution discuss the topics of homology and homologous stars. These terms describe sequences of simple spherical stellar models in complete equilibrium where one model is related to any of the others by a simple change in scale. More specifically, assume that the models all have the same constituent physics (equation of state, opacity, etc.), the same uniform composition, and that  $m$  and  $r$  are related as follows. If one of the stars in the homologous collection is chosen as a reference star – call it star 0 and refer to it by a subscript zero – then these relations must apply in order that the star be homologous to one another:

$$r = \frac{R}{R_0} r_0 \quad \text{and} \quad m = \frac{M}{M_0} m_0, \quad (339)$$

where those quantities not subscripted with a zero refer to any other star in the collection. These relations mean that the stars have the same relative mass distribution such that the radius and the mass interior to that radius are related by simple ratios to the corresponding quantities in the reference star (Hansen et al. 2004, pg. 24).

What follows is a simplified treatment of homologous stars using a form of dimensional analysis. We start by writing the Lagrangian versions (i.e. with respect to  $m$ ) of the stellar interior equations in a form that emphasizes the dependence of pressure, temperature, and luminosity on mass and radius. Fundamental constants, such as  $G$ , could be retained but, at the end, it would be apparent that they were not needed. As we have done in a previous question, the result is to construct the set of equations:

$$\begin{aligned} \frac{dr}{dm} &= \frac{1}{4\pi r^2 \rho} \Rightarrow M \propto R^3 \rho, \\ \frac{dP}{dm} &= -\frac{Gm}{4\pi r^4} \Rightarrow P \propto \frac{M^2}{R^4}, \\ \frac{dL}{dm} &= \epsilon \Rightarrow L \propto M \epsilon, \\ \frac{dT}{dm} &= -\frac{3}{64\pi^2 ac} \frac{\kappa L}{r^4 T^3} \Rightarrow T^4 \propto \frac{ML}{R^4}. \end{aligned} \quad (340)$$

Before continuing, we will assume that the pressure can be specified in power-law form as

$$P \propto \rho^{\chi_\rho} T^{\chi_T}, \quad (341)$$

where the proportionality constant and indices  $\chi_\rho$  and  $\chi_T$  are assumed to be the same for all stars in the collection (Hansen et al. 2004, pg. 25). Example cases are listed in Table 10

The aim here is to construct separate  $R$ ,  $\rho$ ,  $T$ , and  $L$  versus  $M$  relations as

$$\begin{aligned} R &\propto M^{\alpha_R}, \\ \rho &\propto M^{\alpha_\rho}, \\ T &\propto M^{\alpha_T}, \\ L &\propto M^{\alpha_L}, \end{aligned} \quad (342)$$

where the exponents  $\alpha$  are to be determined. We have the requisite number of equations to do this. For example, if we take the fourth identity in equation (340) and merge it with the power-law relation for opacity in equation (338), then we have that:

$$M \propto R^4 \rho^{-n} L^{-1} T^{4+s}, \quad (343)$$

which, after taking the logarithm of both sides with equation (342) in place and dividing out common  $\ln M$  terms, suggests that

$$4\alpha_R - n\alpha_\rho - \alpha_L + (4+s)\alpha_T = 1. \quad (344)$$

The same thing can be done for the remaining three identities in equation (340) using equations (335), (338), and (341) to obtain a set of equations involving the different  $\alpha$ 's. With ordinary algebra this system of four equations in four unknowns can be solved to express the  $\alpha$ 's in terms of the power-law indices  $\lambda$ ,  $\nu$ ,  $n$ ,  $s$ ,  $\chi_\rho$ , and  $\chi_T$  (Hansen et al. 2004, pg. 26).

We can test how well this analysis works by comparing to, say, the observed mass-luminosity relation of MS stars. We expect that stars on the upper (more massive and luminous part of) MS should have higher central temperatures just because they are more massive. The appropriate opacity law to use in this case is electron scattering for which  $n = s = 0$ . Similarly, the energy is generated primarily by the CNO cycle for which  $\lambda = 1$  and  $\nu = 15$ . Although the inner regions of these stars are convective, radiative transport of energy still dominates in the outer regions from which the power finally escapes. Finally, although radiation pressure is important, the pressure is mostly determined by the ideal gas law for which  $\chi_\rho = \chi_T = 1$ . Using these values, it turns out that  $\alpha_R = 0.83$  and  $\alpha_L = 3.0$ , which from equation (333) implies that  $L \propto T_{\text{eff}}^9$ . For comparison, empirical values based on high-mass stars on the MS gives  $\alpha_R = 0.75$  and  $\alpha_L = 3.5$ , so that  $L \propto T_{\text{eff}}^7$  (Hansen et al. 2004, pg. 28).

The lower (less massive) MS is more difficult to treat. The pp-chains ( $\lambda = 1$ ,  $\nu = 4$ ) dominate the energy generation rate and Kramers' opacity ( $n = 1$ ,  $s = 3.5$ ) operators through much of the star but, and especially for very low mass stars, convection is

important. However, being intrepid, let us see what happens if we assume radiation transport and try to duplicate stars around a solar mass; the result is  $\alpha_R = 0.08$  and  $\alpha_L = 5.5$ , so that  $L \propto T_{\text{eff}}^4$ . From empirical data, the latter value is closer to  $\alpha_L \approx 3.9$  (Hansen et al. 2004, pg. 29).

#### SIMPLER ESTIMATES

Our previous estimates based on homology arguments provided decent estimates of the mass-luminosity relation observed for MS stars. Since these arguments depended on specific power-law indices for the energy generation rate, they required knowledge of the source of energy for each star. However, as we have seen in a previous question, we can perform simpler estimates for these relations. In this case, we start with the set of equations in (340) and assume constant temperature and density throughout the star so that power-law indices are immaterial. In this case, we arrive at the following conclusions:

$$R \propto M \quad \text{and} \quad L \propto M^3. \quad (345)$$

The first result implicitly relied on knowledge of the energy source within the star as it assumed that central temperature is independent of stellar mass, based on the large temperature dependence of the pp-chain and CNO cycle. The second identity, however, did not rely on any knowledge of the energy source within the star. As we have already shown, this relation is derived by equating the hydrostatic condition,  $P \propto M^2/R^4$ , with the ideal gas law condition,  $P \propto MT/R^3$ , to arrive at  $T \propto M/R$ . Further equating this relation to the radiation transfer condition,  $T^4 \propto ML/R^4$ , gives the desired  $L \propto M^3$  (See Question 13 of Stars for details).

#### CONVECTIVE STARS

If energy transport is primarily by means of convection, then the above analysis must be modified. If we assume vigorous and efficient convection then the dependence of temperature on density as a function of radius is adiabatic. Specifically, that means

$$T_r \propto \rho_r^{\gamma-1}. \quad (346)$$

This relation replaces the radiative transfer equation in the previous analysis, so that equation (344) becomes

$$(\gamma - 1)\alpha_\rho - \alpha_T = 0, \quad (347)$$

and therefore the  $\alpha$ 's will have to be modified. With  $\gamma = 5/3$  we find that  $R \propto M^{-1/3}$ , equivalent to the mass-radius relation of nonrelativistic WDs that we have seen earlier, and  $\alpha_L = 8$  and 29 for the pp-chain and CNO-cycle, respectively. However, the problem of applying this relation is that the structure of convective stars is largely determined by what happens in the very outermost radiative surfaces (Hansen et al. 2004, pg. 29).

#### HAYASHI LINE

Convective stars are important in relation to the **Hayashi Line (HL)**, defined as the locus of points in the HR diagram of fully convective stars of given parameters. Note that for each set of parameters, such as mass or chemical composition, there is a separate HL. These lines are located far to the right in the HR diagram, typically at  $T_{\text{eff}} \sim 4000$  K, and they are very steep, in large parts almost vertical. The HL also represents a border line between an *allowed* region (to its left) and a *forbidden* region (to its right) in the HR diagram for all stars with these parameters, provided they are in hydrostatic equilibrium and have a fully adjusted convection. The latter means that, at any time, the convective elements have the properties required by the MLT. Changes in time of the large-scale quantities of the stars are supposed to be slow enough for the convection to have time to adjust to a new situation; otherwise, we would have to use a theory of time-dependent convection. Since hydrostatic and convective adjustments are rapid, stars could survive on the right-hand side of the HL only for a very short time (Kippenhahn & Weigert 1990, pg. 224).

It turns out that for convective transfer, the luminosity is not proportional to  $\nabla$  itself (like it is for radiative transfer), but is actually proportional to  $\nabla - \nabla_{\text{ad}}$ , which may be as small as  $10^{-7}$  for efficient convection. In this sense, the luminosity is effectively decoupled from the temperature-pressure stratification within the star. In order to calculate the luminosity of a fully convective star, we have to appeal to the only region where the gradient is sufficiently non-adiabatic. This is the radiative atmosphere and a layer immediately below where the convection is ineffective (i.e. strongly superadiabatic). By this argument, we arrive at the statement that the structure of the outermost layers determines the luminosity of a fully convective star. This means, on the other hand, that such stars are very sensitive to all influences and uncertainties near their outer boundary (Kippenhahn & Weigert 1990, pg. 226).

In order to derive some typical properties of the HL analytically, we shall use an extremely crude model for fully convective stars. We know that nearly all of the interior part of convective stars has an adiabatic stratification, such that  $d\ln T/d\ln P = \nabla_{\text{ad}}$ . We shall assume that this simple relation between  $P$  and  $T$  holds for the whole interior up to the photosphere (i.e. we neglect the superadiabaticity in the range immediately below the photosphere and the depression of  $\nabla_{\text{ad}}$  in those regions near the surface where H and He are partially ionized). We thus simply assume  $\nabla_{\text{ad}}$  to be constant throughout the star's interior, say  $\nabla_{\text{ad}} = 0.4$ , which is the value for a fully ionized ideal gas. We then have for the whole interior the simple  $P-T$  relation,  $P \propto T^{1+n}$ , where  $n = 1/\nabla_{\text{ad}} - 1 = 1.5$  is the polytrope index of the star. This is assumed to hold up to the photosphere where the optical depth  $\tau = 2/3$ ,  $P = P_S$ ,  $T = T_{\text{eff}}$ ,  $r = R$ , and  $m = M$ . Above this point we suppose to have a radiative atmosphere with a simple absorption law of the form  $\kappa \propto P^a T^b$ . Integration of the hydrostatic equation through the atmosphere yields the photospheric pressure

$$P_S = (\text{const.}) \left( \frac{M}{R^2} T_{\text{eff}}^{-b} \right)^{\frac{1}{1+a}}. \quad (348)$$

This can now be fit to the interior solution and can be mapped to the HR diagram for stars of different mass. Since fully convective stars have small photospheric temperatures  $T \lesssim 5000$  K,  $\kappa$  is dominated by H<sup>-</sup> absorption so that  $a = 1$  and  $b = 3$ . With these parameters set, the details find that for this set of fully convective stars,  $L \propto T_{\text{eff}}^{20}$  and  $T_{\text{eff}} \propto M^{0.2}$ . Hence, the HL is very steep on the HR diagram where it shifts slightly to the left with increasing mass. These two identities can then be combined to show that  $L \propto M^4$  (Kippenhahn & Weigert 1990, pg. 228).

**QUESTION 15**

**What source of pressure prevents gravitational collapse in a normal star, a white dwarf, and a neutron star? Why, physically, is there a maximum mass for the latter two? And why can we calculate this easily for a white dwarf, but not for a neutron star?**

### QUESTION 15

**What source of pressure prevents gravitational collapse in a normal star, a white dwarf, and a neutron star? Why, physically, is there a maximum mass for the latter two? And why can we calculate this easily for a white dwarf, but not for a neutron star?**

Simply stated, the sources of pressure that prevent collapse in a normal star, white dwarf (WD), and neutron star (NS) are thermal gas and radiation pressure, electron degeneracy pressure, and neutron degeneracy pressure, respectively. A maximum mass for the latter two occurs from the quantum-mechanical behaviour of electrons and neutrons in the relativistic limit. Determining the maximum mass for a neutron star becomes appreciably more difficult since general relativistic effects become important. To gain a better understanding of these points we will discuss each in detail below.

A star is a self-gravitating gaseous ball supported by thermal and radiation pressure. In particular, it is often stated that stars exist in hydrostatic equilibrium in which the pressure gradient within its interior counteracts the inward-pulling force of gravity:

$$\frac{dP}{dr} = -G \frac{M(r)\rho}{r^2} = -\rho g, \quad (349)$$

where  $g \equiv GM(r)/r^2$  is the local acceleration due to gravity at radius  $r$  within the star. Equation (349) is valid for a static and spherically symmetric star and is derived in Carroll & Ostlie (2007, pg. 287).

Within a normal star the pressure maintaining hydrostatic equilibrium comes from a combination of radiation pressure and thermal support from the gaseous interior. In general, the pressure exerted by a collection of particles occupying some distribution function,  $n_p dp$ , in momentum space, is given by

$$P = \frac{1}{3} \int_0^\infty n_p p v \, dp. \quad (350)$$

For the case of massive, non-relativistic particles, we may use the relation  $p = mv$ , to rewrite equation (350) in terms of  $n_v dv$ , the number density of particles having speeds between  $v$  and  $v+dv$ . In the case of an ideal gas this is simply the Maxwell-Boltzmann velocity distribution which, upon integration, leads to the famous ideal gas law

$$P_{\text{gas}} = nkT = \frac{\rho kT}{\mu m_H}. \quad (351)$$

The second relation makes use of the mean molecular weight  $\mu \equiv \bar{m}/m_H$ , where  $\bar{m}$  is the average mass of a gas particle (Carroll & Ostlie 2007, pg. 291). Since photons possess a momentum  $p = h\nu/c$ , we can derive a corresponding radiation pressure. Using this momentum relation, substituting  $v$  for  $c$ , making use of the Planck blackbody function, and plugging into equation (350) yields

$$P_{\text{rad}} = \frac{1}{3} a T^4, \quad (352)$$

where  $a = 4\sigma/c$  is a constant (Carroll & Ostlie 2007, pg. 295). In many astrophysical conditions the pressure due to photons can drastically exceed the pressure produced by the gas.

Combining the above results show that the pressure within a normal star follows

$$P = \frac{\rho kT}{\mu m_H} + \frac{1}{3} a T^4. \quad (353)$$

In our analysis we have made use of the ideal gas law, which is not strictly valid for all stellar environments. In particular, when relativistic effects are important, the upper integration limit in equation (350) must be modified. Moreover, when quantum-mechanical effects become important, the Maxwell-Boltzmann distribution function must be replaced by the appropriate Fermi-Dirac or Bose-Einstein distribution function, depending on the nature of the particles.

Eventually nuclear fusion within the core of a normal star will cease, at which point the star begins to cool and contract due to the reduced thermal support. As the star collapses its pressure will not go to zero because of degeneracy pressure. Since  $m_e \ll m_p$ , the electrons become degenerate first at a number density of roughly one electron in a cube of side the Compton wavelength.

As the star contracts it will eventually reach equilibrium where its total energy  $E = E_{\text{grav}} + E_{\text{kin}}$  is a minimum. For a star of mass  $M$  and radius  $R$  we have that

$$E \sim -\frac{GM^2}{R} + n R^3 \langle E \rangle, \quad (354)$$

where  $\langle E \rangle$  is the average kinetic energy of the atoms with number density  $n$  (Townsend 1997). From the uncertainty principle we know that (momentum of particles) · (separation of particles)  $\sim \hbar$ . Hence, as the electrons become degenerate we have that

$$\langle p_e \rangle n_e^{-1/3} \sim \hbar, \quad (355)$$

where  $\langle p_e \rangle$  and  $n_e$  are the average momentum and number density of electrons respectively.

If we assume that the electrons are non-relativistic then we have that  $\langle E \rangle \sim \langle p_e \rangle^2/m_e$ . Using this relation and equation (355) allows us to write the total gravitational plus kinetic energy of the star as

$$E \sim -\frac{GM^2}{R} + \frac{\hbar^2}{m_e} n_e^{5/3} R^2 = -\frac{GM^2}{R} + \frac{\hbar^2}{m_e} (n_e R^3)^{5/3} \frac{1}{R^2} \sim -\frac{GM^2}{R} + \frac{\hbar^2}{m_e} \left( \frac{M}{m_p} \right)^{5/3} \frac{1}{R^2}. \quad (356)$$

The last identity arises by noting that  $n_e R^3 \sim M/m_p$ , assuming that the mass of the white dwarf arises from protons only. We then compute the equilibrium conditions by setting  $dE/dR = 0$ , yielding

$$R \sim \frac{\hbar^2}{Gm_e} \left( \frac{1}{Mm_p^5} \right)^{1/3}. \quad (357)$$

This relation shows that the volume of a WD decreases with increasing mass. This arises since electrons must be more closely spaced to generate the larger degeneracy pressure required to support more massive stars.

However, our non-relativistic approximation requires that  $\langle p_e \rangle \ll m_e c$ , which from equation (357), leads to the equivalent condition that

$$M \ll \frac{1}{m_p} \left( \frac{\hbar c}{G} \right)^{3/2}. \quad (358)$$

Hence, for sufficiently large mass the electrons must be relativistic, implying that  $\langle E \rangle \sim \langle p_e \rangle c$ . This modifies the total energy to

$$E \sim -\frac{GM^2}{R} + \hbar c \left( \frac{M}{m_p} \right)^{4/3} \frac{1}{R}, \quad (359)$$

and it is easy to see that the only equilibrium condition is

$$M_C \sim \frac{1}{m_p^2} \left( \frac{\hbar c}{G} \right)^{3/2}. \quad (360)$$

For smaller  $M$ ,  $R$  must increase until the electrons become non-relativistic, in which case the star is supported by electron degeneracy pressure, as above. For larger  $M$ ,  $R$  must continue to decrease, so electron degeneracy pressure cannot support the star.  $M_C$  therefore represents a critical mass, above which a WD cannot exist. This is known as the Chandrasekhar mass and more detailed calculations place it at  $M_C \sim 1.4 M_\odot$  (Townsend 1997). A corresponding Chandrasekhar radius is obtained by substituting equation (360) into equation (357).

The electron energies available within a WD are of the order of the Fermi energy,  $\epsilon_F$ . Necessarily  $\epsilon_F \ll m_e c^2$ , since the electrons would otherwise be relativistic and unable to support the star. Hence, a WD is stable against inverse  $\beta$ -decay,



since this reaction needs energy of at least  $(\Delta m_{np})c^2$  where  $\Delta m_{np}$  is the neutron-proton mass difference. Clearly  $\Delta m_{np} > m_e$  since otherwise  $\beta$ -decay would be impossible, and indeed  $\Delta m_{np} \sim 3m_e$ . Only within stars with  $M > M_C$  can the star contract enough until  $\epsilon_F \sim (\Delta m_{np})c^2$  so that inverse  $\beta$ -decay can occur. While the reaction occurs it cannot come into equilibrium with its inverse reaction due to the escape of neutrinos from the star. In addition,  $\beta$ -decay,



cannot occur because all electron energy levels below  $\Delta m_{np}c^2$  will be filled by degenerate electrons. Since inverse  $\beta$ -decay removes the electron degeneracy pressure the star will undergo a catastrophic collapse to nuclear matter density, after which point it becomes supported by neutron degeneracy pressure (Townsend 1997).

As the star collapses to incredibly large pressures its interior becomes a mixture of free neutrons, protons, and electrons with a the ratio of neutrons:protons:electrons approaching a limiting value of 8:1:1, as determined by the balance between the competing processes of electron capture and  $\beta$ -decay inhibited by the presence of degenerate electrons. In addition, both the neutrons and protons pair with their own species to produce bosonic superfluids. The properties of NS material at densities beyond the nuclear density are poorly understood (Carroll & Ostlie 2007, pg. 581). Its description becomes complicated since general relativistic effects become important at the scales associated with NSs.

We can easily see that general relativity becomes important when describing NSs by applying the same procedure we used for WDs, but with  $m_e \rightarrow m_n$ . The critical mass is independent of  $m_e$  and will therefore remain the same, but the critical radius will becomes

$$R_C \sim \frac{GM_C}{c^2}, \quad (363)$$

which is the Schwarzschild radius, so the neglect of general relativity is not justified (Townsend 1997). Nevertheless, detailed numerical models find that the mass of a NS cannot exceed  $\sim 2.2 M_\odot$  if it is static, and  $\sim 2.9 M_\odot$  if it is rotating rapidly (Carroll & Ostlie 2007, pg. 583). The increased mass of the latter arises from the additional support provided by centrifugal effects.

#### HYDROSTATIC EQUILIBRIUM

The mass-radius relation for WDs and NSs call for a simple explanation, since they contradict the everyday experience that spheres of given material (say iron) become larger with increasing mass. This experience is not only obtained by handling small iron spheres, but also by measurements of planets. Let us consider rough averages (taken over the whole star) of the basis equation of hydrostatic equilibrium, in which case we have

$$\frac{P}{M} \approx \frac{GM}{4\pi R^4}, \quad (364)$$

where  $P$  is some average value. We replace it by the average density,  $\rho \sim M/R^3$ , using a degenerate equation of state,

$$P \sim \rho^\gamma \sim \left( \frac{M}{R^3} \right)^\gamma. \quad (365)$$

The pressure term,  $f_P$ , on the left-hand side of equation (364), and the gravity term,  $f_G$ , on the right-hand side are then

$$f_P \sim \frac{M^{\gamma-1}}{R^{3\gamma}} \quad \text{and} \quad f_G \sim \frac{M}{R^4}. \quad (366)$$

Their ratio must be unity for hydrostatic equilibrium:  $f \equiv f_G/f_P = M^{2-\gamma} R^{3\gamma-4} = M^{1/3} R$  for  $\gamma = 5/3$ , and  $= M^{2/3}$  for  $\gamma = 4/3$ . Suppose we have a given stellar mass with  $M < M_{\text{Ch}}$  and nonrelativistic electrons with  $\gamma = 5/3$ . Then the star can easily find an equilibrium by adjusted  $R$  such that  $f = 1$ . If we now slightly increase  $M$ , then  $f > 1$  (gravity exceeds the pressure force), and  $R$  must decrease in order to regain equilibrium at  $f = 1$ . However, if the electrons are relativistic so that  $\gamma = 4/3$  then  $f$  is independent of  $R$ . Equilibrium can only be achieved by adjusting  $M$  to a certain value  $M_{\text{Ch}}$ . If  $M < M_{\text{Ch}}$ , then  $f < 1$  and the dominant pressure term makes the star expand until the star becomes nonrelativistic. For  $M > M_{\text{Ch}}$ ,  $f > 1$ , and the dominant gravity term makes the star contract; but this does not help either, and the star must collapse without finding an equilibrium. So  $M_{\text{Ch}}$  is quite obviously a mass limit for these stable configurations (Kippenhahn & Weigert 1990, pg. 369).

**QUESTION 16**

**Sketch the SED of an O, A, G, M, and T star. Give defining spectral characteristics, and describe physically.**

### QUESTION 16

**Sketch the SED of an O, A, G, M, and T star. Give defining spectral characteristics, and describe physically.**

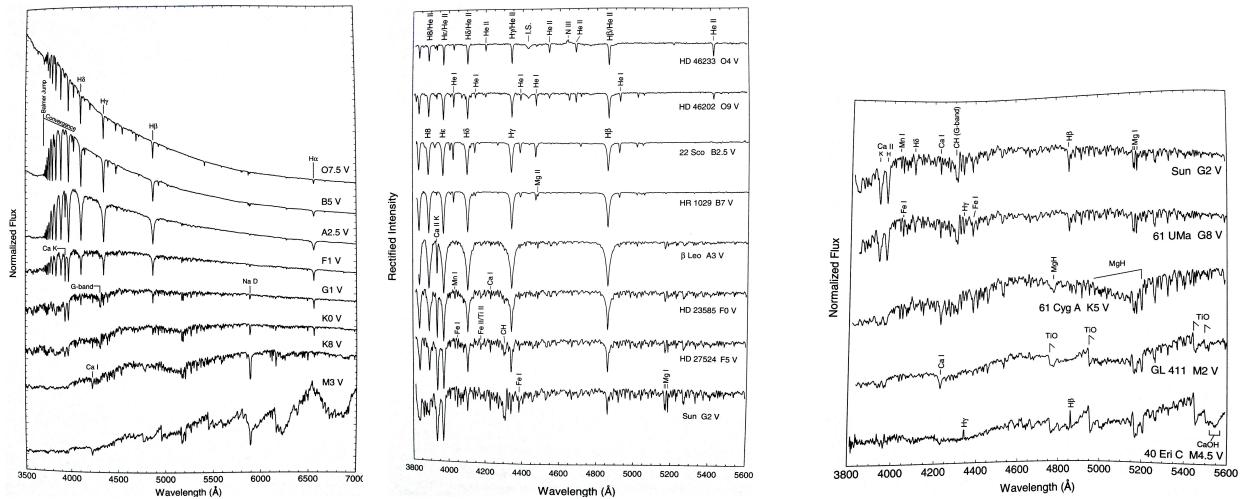


FIG. 91.— (left) The OBAFGK spectral sequence for MS stars illustrating that the spectral sequence is ordered in terms of temperature. Here, the normalized stellar flux (the *energy distribution*) is plotted against wavelength. Some of the more prominent features are marked, including the Balmer jump and convergence. The spectra have been normalized at a common wavelength, and separated by one continuum unit for clarity, except the bottom spectrum, which is offset two units. (centre) The MS from O4 to G2, with the spectra offset by 0.7 continuum units for clarity. (right) The MS from G2 to M4.5, where the flux-calibrated spectra have been normalized at 5445 Å and given integer vertical offsets for clarity. Images taken from Gray & Corbally (2009).

The surprising diversity of spectral forms is a reflection of the wide range of physical phenomena that go into the formation of stellar spectra. It is remarkable, therefore, that the vast majority of stellar spectra can be comprehended on the basis of only two physical parameters – temperature and gas pressure (or its proxies, surface gravity and density) (Gray & Corbally 2009, pg. 32). For this reason, the Harvard classification (OBAFGKMLT) has been ordered in such a way that it reflects a temperature sequence. In terms of classification, important lines in this group are the hydrogen Balmer lines, the lines of neutral helium, the iron lines, the H and K doublet of ionized calcium at 396.8 and 393.3 nm, the G band due to the CH molecule and some metals around 431 nm, the neutral calcium line at 422.7 nm and the lines of titanium oxide (TiO) (Karttunen et al. 2006, pg. 209).

The fact that the Harvard ordering of spectral types is a temperature sequence is made clear in Figure 91, which shows how the optical SEDs of MS stars vary with spectral type. Note that the centre and right panels use different spectral formats. In the centre panel, the *early-type* stars have spectra presented in “rectified” format, in which the intensity of the continuum points (i.e. the points not affected by line absorption) have been normalized to unity, whereas the right panel, for *late-type* stars, use “normalized” format where the stellar fluxes have been normalized to unity at a common point (this is done since the density of spectral lines is so great that no true continuum of points exist, and so rectification is not a good representation) (Gray & Corbally 2009, pg. 34).

A glance at Figure 91 indicates the salient feature of the sequence of early-type spectra is the behaviour of the hydrogen Balmer lines (the H $\beta$ , H $\gamma$ , H $\delta$ , He I, H $\zeta$ , and H $\eta$  lines are visible in these spectra). Note that in O stars, the hottest normal stars, the Balmer lines are quite weak. With decreasing temperature, the Balmer lines increase in strength, coming to a maximum in the early-type A stars, at a spectral type of about A2. They then fade rapidly with decreasing temperature, and cease to dominate the blue-violet spectrum in K-type and later stars. Lines of other species show a similar behaviour. For instance, lines of neutral helium (He I) are very weak in O stars, but grow in strength with decreasing temperature, coming to a maximum at B2. They then fade and essentially disappear from classification-resolution by A0. In the lines of O stars, lines of singly ionized helium (He II) are already declining in strength; their peak would be attained only in stars substantially hotter than the hottest known O-type star (Gray & Corbally 2009, pg. 34).

Another outstanding feature of the spectral sequence is the appearance and rapidly growing strength of lines due to metals. While lines of doubly and triply ionized metals appear in the spectra of O stars, and singly and doubly ionized metals in B stars, lines of metals begin to dominate only in A stars. The strongest metal line in this spectral region, the Ca II K-line, first appears at a spectral type of B9, and grows rapidly with decreasing temperature. Notice that most of the metal lines, especially the lines due to neutral species, grow in strength with decreasing temperature until K5, after which many begin to decline in strength. While this behaviour appears superficially like that of the hydrogen and helium lines, the physics governing these line strengths is considerably more complex, as we will see shortly (Gray & Corbally 2009, pg. 36).

Finally, spectral features due to molecules make their first appearance in F stars. The feature marked “CH” in the centre panel of Figure 91 is due to the diatomic molecule CH and is called the **G-band**. This molecular band grows rapidly in strength and comes to a maximum at K2, after which it fades away. In K stars, molecular bands due to CN and MgH are prominent in the blue-violet region. The M spectra are dominated by strong bands of TiO (Gray & Corbally 2009, pg. 37).

The T-dwarf class is the latest-type spectral class currently studied, encompassing the lowest luminosity and lowest effective temperature BDs. These sources are distinguished from L dwarfs (and indeed all other stellar classes) by the presence of CH<sub>4</sub>

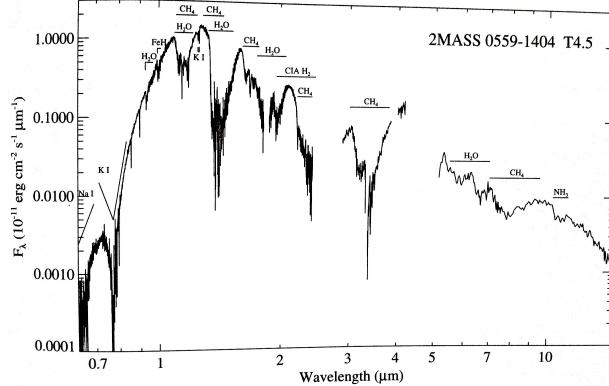


FIG. 92.— Observed spectrum of a T dwarf (MS) with prominent atomic and molecular features labelled. Image taken from Gray & Corbally (2009).

absorption in the NIR spectra, in addition to strong  $\text{H}_2\text{O}$  and  $\text{NH}_3$  bands, and SEDs that are increasingly peaked at NIR and MIR wavelengths. Their spectra are thought to be similar in appearance to the so-called “hot Jupiters”, making them important templates for exoplanet studies. Figure 92 shows a typical SED of a T dwarf star. The primary distinguishing spectral traits are the NIR  $\text{CH}_4$  absorption bands along with the strong  $\text{H}_2\text{O}$  bands, which are present (but weaker) in the NIR spectra of M and L dwarfs. An additional molecular opacity source present in the NIR spectra is collision-induced  $\text{H}_2$  absorption (CIA  $\text{H}_2$ ), arising from an induced quadrupolar moment of the symmetric  $\text{H}_2$  molecule from gas collisions in the photosphere. Atomic line absorption at NIR wavelengths is largely limited to K I lines, although these disappear in the spectra of the latest-type T dwarfs (Gray & Corbally 2009, pg. 391).

In summary, the defining spectral characteristics of O, A, G, M, and T stars are:

- **O Stars:** Spectra dominated by lines from multiply ionized atoms. For example, these stars have the strongest He II absorption lines that sometimes appear as emission lines. Balmer lines are relatively weak since hydrogen is ionized and He I lines are somewhat weaker than in B stars.
- **A Stars:** The H I Balmer lines are the strongest here and dominate the whole spectrum. He I is no longer visible here whereas neutral metal lines begin to appear. At this point, the Ca II H and K lines are also beginning to emerge in strength.
- **G Stars:** Ca II lines are continuing to become stronger as well as Fe I and other neutral metal lines. Conversely, the hydrogen Balmer lines are still becoming weaker. The G-band is also clearly visible here.
- **M Stars:** Spectra dominated by molecular absorption bands, especially TiO. Neutral metal absorption lines remain strong.
- **T Stars:** Very strong molecular absorption bands of  $\text{CH}_4$  and  $\text{H}_2\text{O}$ .

#### PHYSICAL CONDITIONS IN STELLAR PHOTOSPHERES

We have seen that one of the outstanding features of the spectral sequence is the behaviour of the hydrogen Balmer lines. The behaviour of the Balmer lines in the MS stars may be understood qualitatively as follows. In the late-type stars the photospheric temperatures are low enough that only a very small proportion of the hydrogen atoms have been excited to the  $n = 2$  state; most are in the ground  $n = 1$  state. As we move toward earlier types, the photospheric temperatures rise, and greater proportion of hydrogen atoms are in the  $n = 2$  state. This leads to an increase in the line opacity so that the Balmer lines increase in strength. At yet higher temperatures, hydrogen begins to become ionized, and so while the proportion of neutral hydrogen atoms in the  $n = 2$  state is still increasing, the actual number of atoms in that state is actually declining. This leads to a decline in the line opacity, and a weakening of the Balmer lines. This behaviour is quantitatively described through the use of the Boltzmann and Saha equations, both of which are only strictly valid under the conditions of thermodynamic equilibrium. The former relates the number densities of a given species (atom or ion) in two different excitation states, while the latter relates the number densities of ions in two adjacent ionization states. The description of Balmer lines is made easy since continuous opacity plays only a minor role here. In particular, the continuous opacity in the vicinity of the Balmer lines peaks strongly at a temperature of about  $10^4$  K. However, it turns out that the variation in the hydrogen line opacity with temperature is many orders of magnitude greater than the variation in the continuous opacity (Gray & Corbally 2009, pg. 54).

The behaviour of the spectral lines due to metals obey the same general principles as the hydrogen lines. However, the spectra in Figure 91 show features that are not immediately understandable from this standpoint. For example, the enormous strength of the Ca II K and H lines compared with lines of Fe II is somewhat puzzling considering that iron is nearly 10 times as abundant as calcium in a typical stellar photosphere. The answer is that while most spectral lines of metals follow the same pattern of behaviour as the hydrogen lines, that is, they come to a maximum at some intermediate temperature depending upon the excitation and ionization energies, permitted (i.e. not forbidden) lines that arise from the ground state of the atom or ion

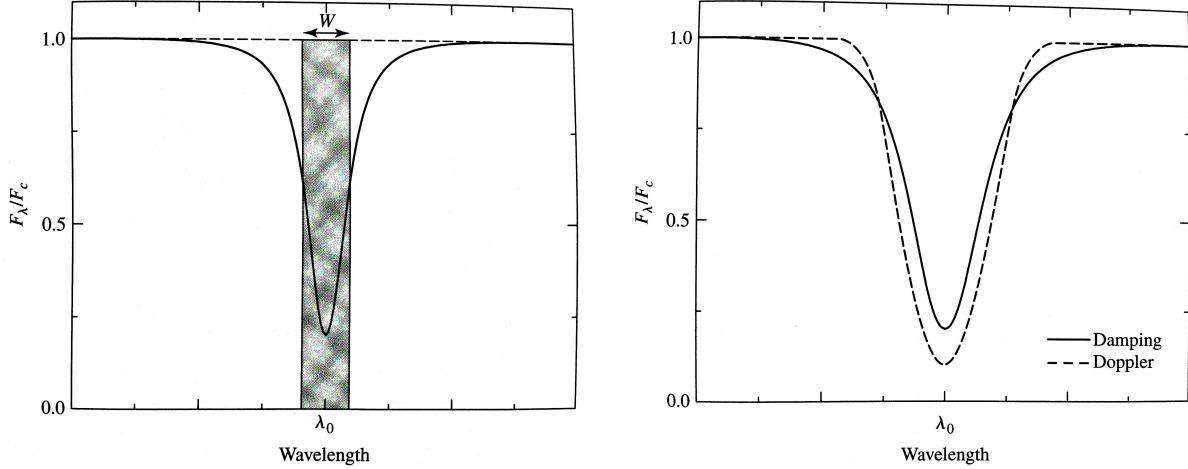


FIG. 93.— (left) The profile of a typical absorption line. (right) Schematic damping and Doppler line profiles, scaled so that they have the same equivalent width. Images taken from Carroll & Ostlie (2007).

(called **resonance lines**) are an exception<sup>60</sup>. At low temperatures, atoms and ions are preferentially found in the ground state, and hence the strength of these lines, in particular the resonance lines of neutral species, grows dramatically with declining temperature. This is exactly the behaviour for the 4226 Å resonance line of Ca I. The resonance lines of ions also tend to grow fairly dramatically with decreasing temperature, but this growth is reversed at temperatures low enough that the ion in question is no longer the dominant ionization state. For instance, the Ca II K and H lines grow with decreasing temperature down to K5 on the MS, and then fade rapidly after that. The reason for this is that Ca I is transitioning to the dominant ionization state in the spectral line-forming region in stellar photospheres with  $T_{\text{eff}} \lesssim 4500$  K. This fact also explains the explosive growth of the Ca I resonance line in spectra later than K5 (Gray & Corbally 2009, pg. 58).

#### SUPERGIANTS

Superficially the OBAFGKM sequence for Ib supergiant stars appears similar to MS dwarfs stars discussed above. However, there are a number of important differences. For instance, in supergiants we see that same general behaviour for Balmer lines as we saw earlier; they are weak in O stars, come to a maximum in A stars, and then fade for later spectral types. The details of this, however, are different for the two luminosity classes. For early-type stars the Balmer lines are broader in the dwarfs than in the supergiants, and also the maximum occurs at later spectral type for the supergiants (late A or F0 as compared to A2 for the dwarfs). Later than F0 the Balmer lines in dwarfs and supergiants have essentially the same strengths. In the A, F, and G supergiants, the metal lines are considerably stronger than the same lines in dwarfs; this is particularly true for lines of ionized metals. In addition, a strong CN band emerges in G supergiants while being absent in G dwarfs, and conversely molecular bands due to MgH are prominent in K dwarfs, though are absent in supergiants (Gray & Corbally 2009, pg. 37). The difference between the spectra of dwarf and supergiant stars resides primarily in the differences between their surface densities. For example, if we use a reduced electron number density in the Saha equation, appropriate for supergiants, we find that the peak strength in Balmer absorption will occur at a lower temperature, just as we have described (Gray & Corbally 2009, pg. 56). At a given temperature, a reduced electron number density implies that more hydrogen atoms will be in the ionized state since less free electron are available for recombination. Consequently, the suppression in Balmer line absorption due to ionization will occur at a smaller temperature for supergiant stars.

#### MEASURING SPECTRA

The most important methods of forming a spectrum are by means of an objective prism or a slit spectrograph. In the former case one obtains a photograph, where each stellar image has been spread into a spectrum. Up to several hundred spectra can be photographed on a single plate and used for spectral classification. The amount of detail that can be seen in a spectrum depends on its **dispersion**, defined as the range of wavelengths per mm on the plate (or per pixel on a CCD). The dispersion of a prism is  $\sim 10$  nm/mm; more detailed observations require a slit spectrograph, which can reach a dispersion of  $\sim 0.01$  nm/mm. The detailed shape of individual spectral lines can then be studied (Karttunen et al. 2006, pg. 208).

The detailed shape of a spectral line is called the **line profile**. The true shape of the line reflects the properties of the stellar atmosphere, but the observed profile is also spread out by the measuring instrument. However, the total absorption in the line, usually expressed in terms of the equivalent width, is less sensitive to observational effects (Karttunen et al. 2006, pg. 209). Figure 93 shows a graph of radiant flux,  $F_\lambda$ , as a function of wavelength for a typical absorption line. Here  $F_\lambda$  is expressed as a fraction of  $F_C$ , the value of the flux from the continuous spectrum outside the spectral line. Near the central wavelength,  $\lambda_0$ , is the **core** of the line, and the sides sweeping upward to the continuum are the line's **wings**. The quantity  $(F_C - F_\lambda)/F_C$  is referred to as the **depth** of the line. The strength of a spectral line is measured in terms of its **equivalent width**, defined as the width of a

<sup>60</sup> The resonance lines of hydrogen – the lines of the Lyman series – lie in the UV and thus outside of the traditional blue-violet classification region

box (shaded in Figure 93) reaching up to the continuum that has the same area as the spectral line:

$$W \equiv \int \frac{F_C - F_\lambda}{F_C} d\lambda, \quad (367)$$

where the integral is taken from one side of the line to the other. The plotted spectral line is termed *optically thin* because there is no wavelength at which the radiant flux has been completely blocked. The opacity  $\kappa_\lambda$  of the stellar material is greatest at the wavelength  $\lambda_0$  at the line's centre and decreases moving into the wings. From a previous question, we know that this means that the centre of the line is formed at higher (and cooler) regions of the stellar atmosphere. Moving into the wings from  $\lambda_0$ , the line formation occurs at progressively deeper (and hotter) layers in the atmosphere, until it merges with the continuum-producing region at an optical depth of 2/3 (Carroll & Ostlie 2007, pg. 267).

Three main mechanisms are responsible for broadening of spectral lines, with each of these producing its own distinctive line profile. The first process is **natural broadening**. Spectral lines cannot be infinitely sharp, even for motionless, isolated atoms. According to Heisenberg's uncertainty principle, as the time available for an energy measurement decreases, the inherent uncertainty of the result increases. Because an electron in an excited state occupies its orbitals for only a brief instant,  $\Delta t$ , the orbital's energy must have a range of energies  $\Delta E \approx \hbar/\Delta t$ <sup>61</sup>. Electrons can make transitions from and to anywhere within these "fuzzy" energy levels, producing an uncertainty in the wavelength of the photon absorbed or emitted in a transition; natural broadening produces a FWHM of the line profile on the order of  $(\Delta\lambda)_{1/2} \simeq 10^{-5}$  nm. The second process is **Doppler broadening**. In thermal equilibrium, the atoms in a gas are moving randomly about with a distribution of speeds described by the Maxwell-Boltzmann distribution function, with the most probable speed given by  $v_{mp} = \sqrt{2kT/m}$ . The wavelengths of the light absorbed or emitted by the atoms in the gas are Doppler-shifted according to  $\Delta\lambda/\lambda = \pm|v_r|/c$ , and so we can expect broadening on the order of  $2\lambda v_{mp}/c \simeq 10^{-2}$  nm; about 10<sup>3</sup> times greater than for natural broadening. Although the line profile for Doppler broadening is much wider than for natural broadening, the line depth decreases exponentially as the wavelength moves away from the central  $\lambda_0$ . This rapid decline is due to the high-speed exponential tail of the Maxwell-Boltzmann distribution. Doppler shifts can also be caused by the large-scale turbulent motions of large masses of gas, as in the atmospheres of giant and supergiant stars, as well as things like stellar rotation, pulsation, and mass loss. The final mechanism is **pressure broadening** whereby orbitals of an atom are perturbed by collisions and electrostatic interactions with other atoms. The general shape of the line produced by this effete is similar to that of natural broadening, with a line profile sometimes known as a **damping profile** (or Lorentz profile). The values of the FWHM for natural and pressure broadening are usually comparable, though the latter can at a times be more than an order of magnitude wider. For this process, the width of the line is proportional to the number density of atoms, which explains why narrower lines are observed for the more luminous giant and supergiant stars that have more tenuous atmospheres (Carroll & Ostlie 2007, pg. 268).

The total line profile, called a **Voigt profile**, is due to the contributions of both the Doppler and damping profiles. The wider line profile for Doppler broadening dominates near the central wavelength  $\lambda_0$ . Farther from  $\lambda_0$ , however, the exponential decrease in the line depth for Doppler broadening means that there is a transition to a damping profile in the wings. Thus, lines tend to have *Doppler cores* and *damping wings*; with these two components shown in Figure 93. The calculation of a spectral line depends not only on the abundance of the element forming the line but also on the quantum-mechanical details of how atoms absorb photons. To find the column density of absorbing atoms,  $N_a$ , that have electrons in the proper orbital for absorbing a photon at the wavelength of the spectral line, the temperature and density are used in the Boltzmann and Saha equations to calculate the atomic states of excitation and ionization. For this purpose, the **curve of growth** is frequently used. This simply gives the equivalent width of a line as a function of column density (Carroll & Ostlie 2007, pg. 271).

<sup>61</sup> The electron's lifetime in the ground state can be taken as infinite, so in that case  $\Delta E = 0$ .

**QUESTION 17**

**What physical and orbital parameters of an extra-solar planet can be determined (a) from radial velocity (Doppler) measurements alone, (b) from transit observations alone, and (c) from the combination of both Doppler and transit observations?**

### QUESTION 17

**What physical and orbital parameters of an extra-solar planet can be determined (a) from radial velocity (Doppler) measurements alone, (b) from transit observations alone, and (c) from the combination of both Doppler and transit observations?**

The generally accepted picture of stellar formation teaches us that a planetary system is a natural byproduct of the stellar formation process. When a cloud of gas and dust contracts to give origin to a star, conservation of angular momentum leads to the formation of a flat disk of gas and dust around the central newborn protostar. As time passes, in a process still not completely understood, dust particles and ice grains in the disk are gathered to form the first planetary seeds. In the outer regions of the disk, where ices can condense, these planetesimals are thought to grow in a few Myr. When such a planetesimal achieves enough mass ( $\sim 10 M_E$ ), its gravitational pull enables it to accrete gas in a runaway process that gives origin to a giant gaseous planet similar to the outer planets in our own Solar System. Later on, in the inner part of the disk, where temperatures are too high and volatiles cannot condense, silicate particles are gathered to form the terrestrial planets like our Earth (Santos 2008).

HST images have revealed a multitude of such protoplanetary disks in the Orion stellar nursery. Together with the number of NIR detections of disks around T-Tauri stars, these findings show that disks are indeed very common around young solar-type stars. This supports the idea that **exoplanets** should be common. However, such systems have escaped detection until very recently. The reason for this has to do with the difficulty to detect such systems. Planets are cold bodies, and their visible spectra results basically from reflected light of the parent star. As a result, in optical wavelengths the planet/stellar luminosity ratio is on the order of  $10^{-9}$ . Seen from a distance of a few pc, a planet is no more than a small undetectable speckle embedded in the diffraction and/or aberration of the stellar image. Only the development of adaptive optics imaging will enable the direct detection of exoplanets (Santos 2008).

But a planet also induces dynamical perturbations into its parent star, giving the possibility to detect its presence by indirect means. Indeed, any star in a binary or multiple system will present a periodic motion about the centre of mass of the system. This effect gives the possibility to indirectly detect a planet orbiting another star, by measuring this dynamical effect. As we will explain below, for solar-type stars this can be used to try to detect planets using two different techniques: **astrometry** and **radial-velocities**. Along with the radial-velocity method, the **transit technique** has proven to be very important in the discovery of exoplanets. Other possible methods, which are more difficult, include **gravitational microlensing**, **pulsar timing**, and **direct imaging**, though we will not discuss those here (Santos 2008).

#### ASTROMETRY

The astrometric technique is based on direct measurements of stellar motion on the night sky. This type of exoplanet detection can be described using simple physics. The semi-major axis of the orbital motion of a star around the centre of mass of a two-body system can be described by

$$M_S a_S = M_P a_P, \quad (368)$$

where  $M_S$  and  $M_P$  are the masses of the star and planet with corresponding semi-major axes of their respective orbits,  $a_S$  and  $a_P$ . Defining  $a = a_S + a_P$  as the semi-major axis of the relative orbit (i.e. it denotes the separation of the two bodies), we can write

$$a_S = \frac{M_P}{M_S + M_P} a \quad (369)$$

This equation relates the expected astrometric displacement  $a_S$  of a star of mass  $M_S$  due to the presence of a planet with mass  $M_P$ , separated by the distance  $a$ . The distance  $a$  is also related to the orbital period through Kepler's Third Law:

$$P^2 = \frac{4\pi^2}{G(M_S + M_P)} a^3. \quad (370)$$

In principle, if we can measure  $a_S$ ,  $a_P$ , and  $P$  then we are left with the three equations above with three unknowns ( $a$ ,  $M_1$ , and  $M_2$ ), for which we can solve exactly. Since large  $a$  is desired for this method, it is most sensitive to systems where either the planet is massive, the star is not massive, or the period is long. This process, for example, can be used to determine the masses of visual stellar binary systems (Santos 2008).

In practice, the measurement of the astrometric motion of the primary star in a star-planet system is far more complex. First, we can only hope to measure  $a_S$  and the period  $P$ , since we are not able to directly observe the planet. To solve the above system we need, for example, to estimate the mass of the star,  $M_S$  using stellar evolution models. Secondly, since we are observing an inherently three dimensional process in two dimensions, additional information (e.g., radial-velocity measurements) is needed to construct the true orbit. Thirdly, the actual movement of a star on the night sky is very small ( $\sim 1 \mu\text{arcsecond}$  for the best cases) and thus requires sensitive instruments. Finally, limitations to this technique may also be induced by the stars themselves, and in particular for the most active young stars. The existence of stellar spots may induce variations in the photocentre of the image, causing the measurements of spurious astrometric motions (this can be avoided by making measurements in the NIR where the contrast between spotted and non-spotted regions in the stellar photosphere is smaller) (Santos 2008).

#### RADIAL-VELOCITY

Another technique used to search for the stellar motion induced by an orbiting planet is based on the measurement the star's radial-velocity (RV) (i.e. motion along the line of sight). Using Newtonian mechanics it is straightforward to show (derivation

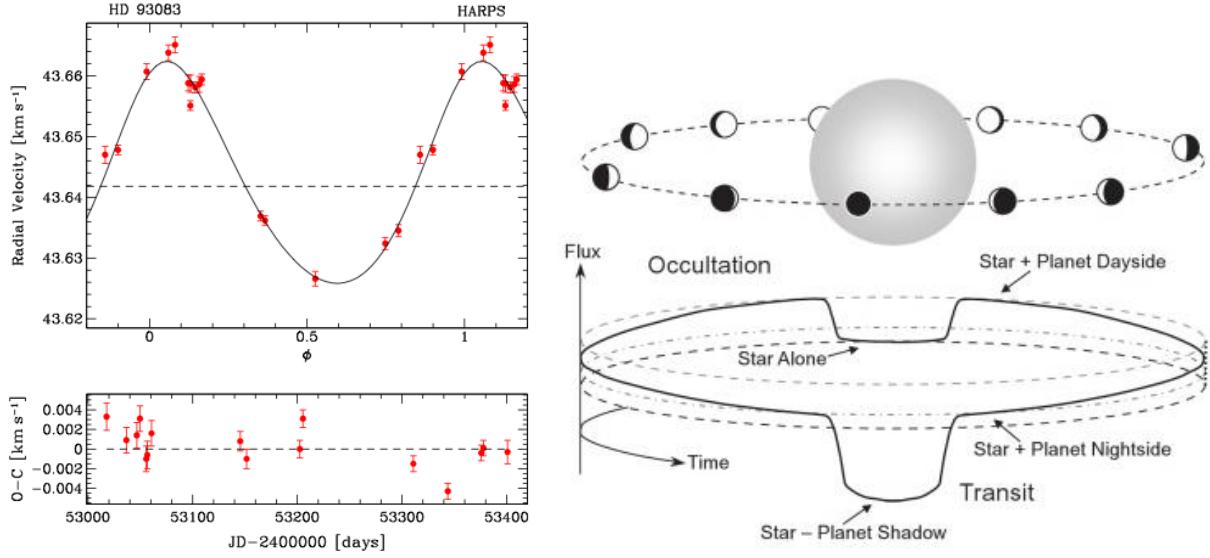


FIG. 94.— (left) Periodic radial-velocity signal induced by the presence of a planet with a mass similar to that of Saturn in a 143-day period, nearly-circular orbit. Image taken from Santos (2008). (right) Illustration of transits and occultations. Only the combined flux of the star and planet is observed. During a transit, the flux drops because the planet blocks a fraction of the starlight. Then the flux rises as the planet's dayside comes into view. The flux drops again when the planet is occulted by the star. Image taken from (Seager 2011).

below) that the RV wobble expected for a star of mass  $M_S$  orbited by a planet of mass  $M_P$  is

$$K_S = 2\pi G \left( \frac{M_S}{P} \right)^{1/3} \frac{q}{(1+q)^{2/3}} \frac{\sin i}{\sqrt{1-e^2}}, \quad (371)$$

where  $q \equiv M_P/M_S$ ,  $i$  is the inclination of the orbital axis with respect to the line of sight, and  $e$  is the eccentricity of the orbit. The RV of the star can be measured from the Doppler shift using high-resolution spectroscopic measurements. The biggest challenge of this technique is that one needs to measure the stellar velocity with a very high-precision. From equation (371),  $K_S$  for Jupiter-like planet (mass and distance) is  $\sim 10$  m s<sup>-1</sup>, while for an Earth-like planet  $K_S \sim 10$  cm s<sup>-1</sup>. The nonrelativistic Doppler shift has the form

$$\frac{\Delta\lambda}{\lambda} = \frac{v}{c}, \quad (372)$$

where  $\lambda$  is the reference wavelength (at zero velocity – typically this corresponds to an absorption spectral line). In the optical regime the typical  $K_S$ 's translate to  $\Delta\lambda \sim 10^{-4}$  Å. For comparison, a typical high-resolution spectrograph is able to resolve two adjacent wavelengths separated by  $\sim 0.1$  Å. To circumvent this problem, the thousands of well-defined absorption lines in a typical stellar spectrum can be combined in a statistical way to achieve the necessary precision (Santos 2008).

One immediate limitation of the RV technique is that we are only able to measure the projected RV. This implies that we can only estimate the *projected mass* of the companion responsible for the observed stellar wobble; what is called the **minimum mass**  $M_{\min} = M_P \sin i$ . Fortunately, it can be shown that for orbits randomly oriented in space it is much more likely to have  $\sin i$  close to unity, meaning that the minimum masses obtained are statistically very close to the true values<sup>62</sup>. The unambiguous determination of the true mass is however only possible if a value for the orbital inclination is obtained (e.g., through an astrometric detection, a transit measurement or, in the case of very young planetary systems, if the disk inclination is measured) (Santos 2008).

In Figure 94 we show a typical radial-velocity curve of a star induced by the presence of a planetary companion. In practice, the orbital parameters of the system (semi-amplitude  $K_S$ , orbital period  $P$ , eccentricity  $e$ ) can be obtained from a fit of the observed points. From these, the planetary minimum mass can be obtained directly from the so-called **mass function**:

$$f(M) \equiv \frac{PK_S^3}{2\pi G} = M_P \frac{\sin^3 i}{(1+q)^2}, \quad (373)$$

with a derivation of this shown below (Santos 2008). Obviously, the stellar mass  $M_S$  must be known in order to use equation (373); this can be obtained indirectly via spectroscopic analysis, photometry, parallax measurements, and comparison with stellar evolution models (Seager 2011, pg. 29). Since this technique requires that we observe several cycles of this motion, it is generally most applicable to short-period, massive planets.

As for the astrometric technique, the RV method has its own limitations. It is well known that intrinsic stellar features, like non-radial pulsation, inhomogeneous convection, or spots may be responsible for the detection of radial-velocity variations. These situations can prevent us from finding planets, if the perturbation is larger than the orbital radial-velocity variation, or even give

<sup>62</sup> Specifically, the distribution function for  $i$  is given by  $\sin i di$  and  $\sin i$  has an average value of  $\pi/4 = 45^\circ$ . As a result, the probability that  $\sin i$  is larger than 0.5 is 87% (Seager 2011, pg. 29).

us false candidates, if they produce a periodic and stable signal over a few rotational periods. The presence of unknown stellar blends can also induce spurious radial-velocity signals, which can simulate the presence of a planetary companion in the case of triple systems. Finally, the acoustic modes of solar-type stars as well as atmospheric granulation and turbulence motions can also cause significant noise in the measurements, with RV amplitudes on the order of  $1 \text{ m s}^{-1}$ . To circumvent this effect, long exposures ( $\sim 15 \text{ min}$ ) are usually taken to average out the solar-type acoustic modes which have typical timescales of  $\sim 5 \text{ min}$  (Santos 2008).

#### PHOTOMETRIC TRANSITS

When a planet crosses the stellar disk as seen from us, it will block part of the star's light. This phenomenon, called a **transit**, can be observed if the orbital axis of the planet is closely perpendicular to our line of sight. For a given system, we can compute that the geometric probability ( $P$ ) that a full transit will occur can be expressed by

$$P = \frac{R_{\text{star}}}{a}, \quad (374)$$

where  $R_{\text{star}}$  and  $a$  are the stellar and orbital radius, respectively (Santos 2008). This formula is valid for a circular orbit and intuitively arises since the geometric area through which a transit can be observed increases with the size of the star and decreases with the distance between the star and planet (e.g., think in relation to a solar eclipse). For a circular orbit, transits and occultations always go together, but for an eccentric orbit it is possible to see transits without occultations or vice versa. An **occultation** is the opposite of a transit whereby the planet passes behind and becomes completely concealed by the larger star. Transits and occultations are examples of **eclipses** which are generically defined to occur whenever one celestial body is obscured by another (Seager 2011, pg. 57).

If a transit event is observed, the expected luminosity variation can be derived to be of the order of:

$$\frac{\Delta L}{L} = \left( \frac{R_{\text{planet}}}{R_{\text{star}}} \right)^2, \quad (375)$$

so that a Jupiter orbiting the Sun would induce a photometric variation on the order of 1%. Obviously, this method will be biased to finding massive planets orbiting close to their parent stars. Furthermore, in the case of an equatorial transit (best case scenario) the transit duration goes like

$$t \propto R_{\text{star}} \sqrt{\frac{a}{M_{\text{star}}}}, \quad (376)$$

since more distance is traversed at large separation and planetary speed increases with  $M_{\text{star}}$ . Usual transit times are of a few hours for short period planets (Santos 2008).

Figure 94 shows a schematic representation of an idealized transit and occultation. During a transit, the observed flux (star plus planet) decreases since the planet blocks part of the starlight. On the other hand, the flux drops during an occultation when no reflected starlight is observed from the daytime side of the planet. Within the intervening time the flux steadily increases after the end of a transit and before the start of an occultation as the amount of starlight reflecting from the planet rises; the opposite occurs after the end of an occultation and before the start of a transit. As the planet initially moves in front of the star during a transit, there will be a period of time when only part of the star is blocked by the planet. Similarly, as it moves away from the star there will be a period of time where it only partially obscures the star. These periods are known as *ingress* and *egress*, respectively, and their values,  $\tau_{\text{ingress}}$  and  $\tau_{\text{egress}}$ , are almost always the same. These times, along with the total transit time, may be used to determine the impact parameter,  $b = R_{\text{star}}/a$ , which can then be used to determine the inclination angle  $i$ . In reality, the star is not a uniform disk and a model of limb darkening (i.e. the star will be brighter in the middle and fainter at the edge) must be used to accurately determine ingress and egress times (Seager 2011, pg. 60).

#### DETERMINING PLANETARY PARAMETERS

With the RV technique we are only directly measuring the line-of-sight velocity of the star as a function of time. Plotting this variation, as in Figure 94, allows us to determine the orbital period  $P$  (simply the oscillatory period), semi-major axis  $a$  (variation in absolute value of RV), and eccentricity  $e$  (variation in frequency of oscillatory motion)<sup>63</sup>. Through the use of equation (373) we can calculate the minimum mass of the planet,  $M_{\text{min}} = M_{\text{Psini}}$ , assuming that the stellar mass can be measured by some external means. With the transit technique we can measure the planetary radius  $R_{\text{planet}}$  (assuming that  $R_{\text{star}}$  is known), the inclination (measurements of  $\tau_{\text{ingress}}$  and  $\tau_{\text{egress}}$ ;  $\text{sini} \approx 1$  for all transits), and period  $P$ . If additional information is known we can measure the semi-major axis  $a$  (if  $M_{\text{star}}$  is known in equation (376)) and eccentricity (if secondary transits are available). Hence, no information can be derived about the planet other than its radius, and at masses below about 100 Jupiter masses, the mass-radius relation is relatively flat. However, if we combine both RV and transit methods we could obtain all orbital properties ( $a, P, e, i$ ) and all planetary properties ( $M_{\text{planet}}, R_{\text{planet}}$ , and  $\bar{\rho}_{\text{planet}}$ ). If no information about the star is included, planetary masses and radii would be known only up to scaling constants for the stellar radius and mass, though we would still be able to determine period and eccentricity (Seager 2011, pg. 60).

<sup>63</sup> See <http://astro.unl.edu/naap/esp/animations/radialVelocitySimulator.html>

### PLANETARY MIGRATION

With the discovery of **hot Jupiters** it became realized that planets must be able to migrate inward while they are forming, and Jupiter is no exception. In fact, computer simulations suggest that Jupiter likely formed 0.5 AU farther out in the solar nebula than its current position. One mechanism by which inward migration of Jupiter (and exoplanets) could occur involves gravitational torques between the planet and the disk. In this mechanism, initial deviations from axial symmetry produce density waves in the disk. The gravitational interaction between a growing planet and density waves results in the simultaneous transfer of angular momentum outward and mass inward. This so-called **Type I migration** mechanism can be shown to be proportional to mass implying that as the planet accretes more material, it moves more rapidly toward its parent star. This process may be able to cause some planets to collide with the star on a timescale of 1 – 10 Myr. However, it initially appeared that the timescale for this migration process was too short compared with the runaway accretion of gases onto the growing Jupiter; in other words, Jupiter would crash into the Sun before it could fully form. In addition, it appeared that Jupiter couldn't grow rapidly enough to reach its present size before the nebula was dissipated by the T-Tauri wind. The solution to these problems may rest with the migration process itself. As the growing planet moves through the solar nebula, it continually encounters fresh material to feed on. If the planet remained fixed in orbit, it would quickly consume all of the available gas and grow only slowly after that. Migration allows it to move through the disk without creating a significant gap in the nebula (Carroll & Ostlie 2007, pg. 867).

It has also been shown that viscosity within the disk can cause objects to migrate inward. This **Type II migration** mechanism causes slowly orbiting particles farther out to speed up because of collisions with higher-velocity particles occupying slightly smaller orbits. The loss of kinetic energy by the inner particles causes them to spiral inward. This type of migration can become the more significant, if slower, migration process when a gap is opened up in the disk. Outward migration is also possible. In this case, the scattering of planetesimals inward results in migration outward. Whether inward or outward depends on the density of the nebula and the abundance of planetesimals (Carroll & Ostlie 2007, pg. 868).

Applying the mechanisms of migration to the evolution of our own solar system, it appears that Jupiter not only influenced objects interior to its present-day location, but also was influential in causing Saturn, Uranus, and Neptune to migrate outward. It seems that Uranus and Neptune initially formed their cores in a region of the nebula with greater density, just as Jupiter and Saturn did. However, because of outward migration, they were able to put only a small amount of extra gas and remain today as ice giants, rather than gas giants (Carroll & Ostlie 2007, pg. 868).

Assuming that Jupiter originally formed at 5.7 AU from the Sun (as simulations suggest), and that Saturn formed about 1 AU closer to the Sun than its current position, the two gas giants would have moved through a critical resonance as Jupiter migrated inward and Saturn migrated outward. When the orbital periods of the two planets reached a 2:1 resonance (i.e. the orbital period of Saturn was exactly twice that of Jupiter), their gravitational influences on other objects in the solar system would have periodically combined at the same points in their orbits, causing significant perturbations to orbits of objects in the asteroid and Kuiper belts. Computer simulations suggest that this resonance effect may have occurred some 700 Myr after the formation of the inner planets, making it plausible that this caused the episode of late heavy bombardment (Carroll & Ostlie 2007, pg. 868).

As a consequence of Neptune's outward migration, Neptune swept up some of the remaining planetesimals, trapping them in 3-to-2 orbital resonances with the planet as it moved outward. The orbits of the scattered Kuiper belt objects (KBOs) were also likely to have been perturbed by the migration of Neptune. The classical KBOs were probably far enough from Neptune not to be as drastically affected by its migration. In fact, the Kuiper belt may be the solar system's analog to debris disks seen around other stars. Similarly, the Oort cloud cometary nuclei are likely to be planetesimals that were scattered more severely by Uranus and Neptune. Once sufficiently far from the Sun, scattered comets had their orbits randomized by passing stars and interstellar clouds (Carroll & Ostlie 2007, pg. 868).

### DERIVATIONS

Here we present a derivation of equations (371) and (373) for the case where a planet of mass  $M_P$  orbits a star of mass  $M_S$  in a circular orbit. In this case, both bodies have constant orbital speeds which are related to the orbital period,  $P$ , via

$$v_S = \frac{2\pi}{P} a_S \quad \text{and} \quad v_P = \frac{2\pi}{P} a_P, \quad (377)$$

where  $a_S$  and  $a_P$  are the semi-major axes for the star and planet (i.e. radii from common centre of mass), respectively. Using  $a = a_S + a_P$  we can rewrite Kepler's Third Law in terms of the orbital speeds:

$$P^2 = \frac{4\pi^2}{G(M_S + M_P)} (a_S + a_P)^3 = \frac{4\pi^2}{G(M_S + M_P)} \left( \frac{P}{2\pi} \right)^3 (v_S + v_P)^3 \Rightarrow (M_S + M_P) = \frac{P}{2\pi G} (v_S + v_P)^3. \quad (378)$$

We now wish to rewrite this expression in terms of the orbital speeds of the two bodies that we measure on Earth. For an inclination angle  $i$  ( $i = 0$  indicates a face-on orientation) we measure the speeds of the two masses to be  $K_S = v_S \sin i$  and  $K_P = v_P \sin i$ . Thus we have

$$(M_S + M_P) = \frac{P}{2\pi G} K_S^3 \frac{\left(1 + \frac{K_P}{K_S}\right)^3}{\sin^3 i}. \quad (379)$$

Note that the centre of mass of the system is defined so that  $M_S a_S = M_P a_P$  and therefore

$$\frac{M_S}{M_P} = \frac{a_P}{a_S} = \frac{v_P}{v_S} = \frac{K_P}{K_S}. \quad (380)$$

Using this relation we rewrite equation (379) as

$$M_S \left( 1 + \frac{M_P}{M_S} \right) = \frac{P}{2\pi G} K_S^3 \frac{\left( \frac{M_S}{M_P} \right)^3 \left( 1 + \frac{M_P}{M_S} \right)^3}{\sin^3 i} \Rightarrow K_S = 2\pi G \left( \frac{M_S}{P} \right)^{1/3} \frac{q}{(1+q)^{2/3}} \sin i, \quad (381)$$

where the mass-ratio  $q \equiv M_P/M_S$ . Equation (381) can alternatively be written in the form

$$\frac{PK_S^3}{2\pi G} = M_P \frac{\sin^3 i}{(1+q)^2}, \quad (382)$$

which is defined to be the mass-function,  $f(M)$ .

**QUESTION 18**

**What spectroscopic signatures help distinguish a young (pre-main sequence) star from a main sequence star of the same spectral type?**

### QUESTION 18

**What spectroscopic signatures help distinguish a young (pre-main sequence) star from a main sequence star of the same spectral type?**

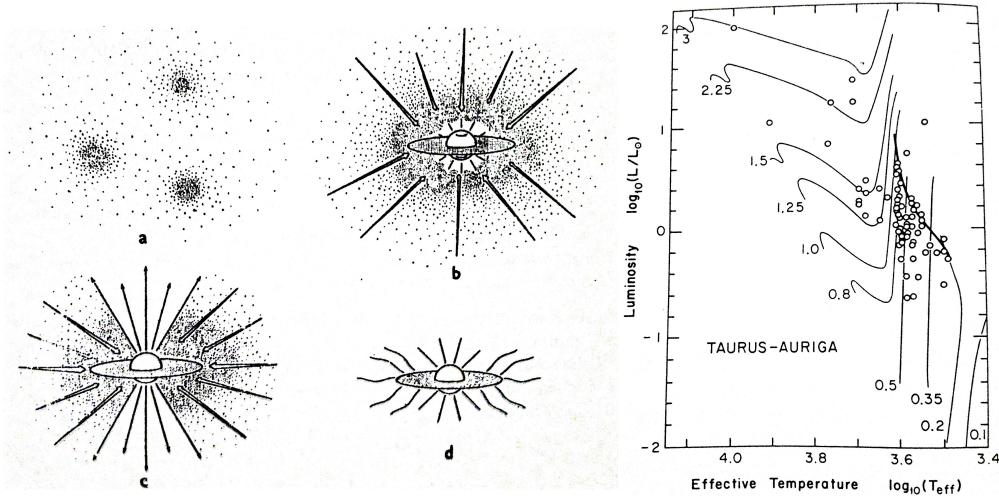


FIG. 95.— (left) This cartoon illustrates the four stages of star formation. (a) First protostar cores form within molecular clouds. Then, in (b), the protostar builds up from the inside out while the surrounding nebular disk rotates around it. (c) Bipolar flows break out along the rotation axis of the system. Finally, in (d), the surrounding nebular material is swept away, and the newly formed star, with disk, is revealed. (right) Shown are pre-MS evolutionary tracks for a variety of ZAMS masses. The evolution starts in the upper right (luminous but cool) and proceeds to the point where hydrogen is ignited on the MS. Also shown are the observed locations of a number of T Tauri stars. Images taken from (Hansen et al. 2004).

### YOUNG STELLAR OBJECTS

A protostar becomes a star when the energy released by thermonuclear fusion (hydrogen to helium) exceeds that released by contraction from the supply of gravitational potential energy. This is not something we can directly observe. Thus protostars and young stars are put into classes 0, 1, and 2 based on things we can see – ratio of IR to visible light, amount of molecular gas around, how the gas is moving, and so forth. The class 0's are still contracting and very few members are known, primarily because of the short time scales involved. The 1's and 2's are already living on nuclear energy and, typically, blowing material off their surfaces in bipolar or jet-like outflows. These jets gradually clear away surrounding afterbirth, opening out from narrow beams to wide cones, until visible light can find its way from the stellar surface (photosphere) to us without being absorbed and reemitted as IR. A cartoon of this sequence is shown in Figure 95 (Hansen et al. 2004, pg. 44).

Signatures of the **young stellar object (YSO)** phase include the following:

- **Variability:** We see variability in the visible light, because material is still falling down onto the surface of the star from a residual disk, so that both the stellar surface and the disk have temperature irregularities that change in time from hours to days and longer.
- **Emission:** Emission lines are observed in their spectra, from the disk, or the bipolar outflow, or both. X-ray emission is also seen from the hot corona. There is also radio emission, but it is too faint to see except from very nearby, very active stars.
- **Infrared:** YSOs have more IR luminosity than older stars of the same mass because there is more dust around.
- **Activity:** A high level of activity is seen, meaning flares, star spots, emission from a hot corona, and so forth, all of which are found at a low level in the Sun and other MS stars. The reason seems to be two-fold: young stars are often rapid rotators (rotation periods from hours to days, versus a month for the Sun) and, because they are cooler than they will be when settled onto the MS, they have surface convection that extends deeper (protostars are actually convective throughout their interior). The combination results in a strong dipole magnetic field, which, in turn, drives the activity.

Due to these features, YSOs were first recognized from the combination of variability, emission lines, location on the HR diagram, and location in space near clouds of gas and dust (Hansen et al. 2004, pg. 45).

The masses, luminosities, and radii of YSOs are not terribly different from solar values, though their lifetimes are much shorter. Since the energy source for these objects is gravitational potential energy, their lifetimes are on the order of  $\tau_{\text{KH}}$ . We can consider the pre-MS evolution of protostars by supposing that they have no interior thermonuclear energy sources (although deuterium burning may occur), so that contraction from a protostellar cloud will eventually yield high luminosities at large radii, and large luminosities usually require convection. If accretion of matter onto the forming star may be neglected (which is not true), the

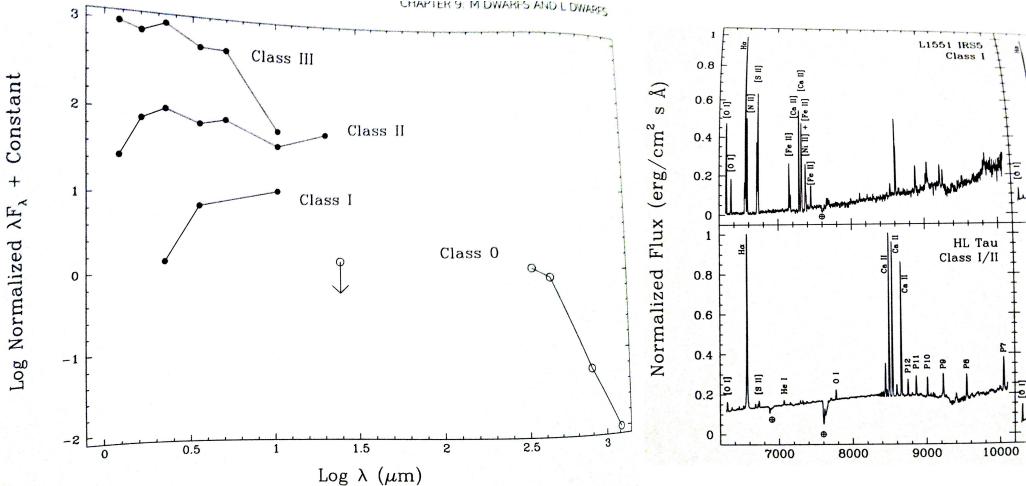


FIG. 96.— (left) Typical SEDs of Class 0, I, II, and III sources. Each spectrum has been normalized to its peak flux and integral offsets added to separate the spectra vertically. Shown by open circles are broad-band photometry for some Class 0 source, while the solid circles are photometric measures of Class I, II, and III sources. The Class 0 source is undetected at the shorter wavelengths where the Class I, II, and III are easily seen. (right) Optical spectra of the Class I and I/II sources, where the flux is normalized to unity at H $\alpha$  emission. The only absorption lines seen are from the Earth's atmosphere. In Class II sources, intrinsic absorption by TiO in the star's photosphere begins to be seen (not shown here). The horizontal axis is wavelength in Å. Images taken from (Gray & Corbally 2009).

object follows the **Hayashi track**, which are the downward tracks in Figure 95. As the star continues to contract, however, its luminosity may decrease to the point where the deep interior ultimately becomes radiative, in which case we can assume the opacity follows Kramers' law and apply the homology arguments of a previous question. For an ideal gas with no nuclear generation ( $\epsilon = 0$ ) we find the mass-luminosity-effective temperature relation:

$$L \propto M^{4.5} T_{\text{eff}}^{0.8}. \quad (383)$$

The implication is that when the luminosity of the contracting star falls below a critical value, evolution proceeds along a track given by equation (383) to higher effective temperatures until, finally, interior temperatures reach the point of hydrogen ignition and the MS stage of evolution begins. Note that if the mass is *too* low, then the track given by equation (383) may lie below the intersection of the MS and the Hayashi track. If so, then hydrogen burning will commence at that intersection but the star will remain completely convective on the MS. If the protostar is even less massive than this, then hydrogen burning may never begin and the result is a BD (Hansen et al. 2004, pg. 374).

Note that the heavy line in Figure 95 defines an upper envelope to where T-Tauri stars appear on the HR diagram. The reason for this *birthline* has to do with the actual hydrodynamical processes of star formation from interstellar clouds. Among these processes is accretion of gas onto the forming star. This provides a high luminosity at the accretion surface but this is obscured by dust and gas. It is only after the accretion ends that the star is fully revealed below the birthline (Hansen et al. 2004, pg. 374).

#### SPECTRAL CLASSIFICATION

The classification of pre-MS objects has followed a different path from that of other low-mass stars. This is primarily because these objects are in the early stages of their evolution and are still surrounded by dust envelopes or disks that make them extremely faint or invisible at the optical wavelengths traditionally used for spectroscopic classification. Therefore, fluxes at longer wavelengths have been used to set up a typing scheme for these objects (Gray & Corbally 2009, pg. 378).

Sources are generally classified using the parameter

$$\alpha \equiv \frac{d \ln \lambda F_\lambda}{d \ln \lambda}, \quad (384)$$

over the wavelength interval  $2\mu\text{m} \lesssim \lambda \lesssim 25\mu\text{m}$ . Three classes of objects have been established based on the value of  $\alpha$ : Class I has a steeply rising continuum with  $0 < \alpha < 3$ , Class II has a relatively flat spectrum with  $-2 < \alpha < 0$ , and Class III has a steeply declining spectrum with  $-3 < \alpha < -2$ . There are also Class 0 sources that are defined as having very little emission shortward of  $10\mu\text{m}$ . Examples of each class are shown in Figure 97 (Gray & Corbally 2009, pg. 378).

Alternatively, the classes can be defined by a **bolometric temperature**,  $T_{\text{bol}}$ , which is the temperature of a blackbody having the same mean frequency as the source's observed SED. When using this parameter, the classes roughly fall as follows: Class 0 has  $T_{\text{bol}} < 70\text{ K}$ , Class I have  $75\text{ K} < T_{\text{bol}} < 650\text{ K}$ , Class II have  $650\text{ K} < T_{\text{bol}} < 2880\text{ K}$ , and Class III have  $T_{\text{bol}} > 2880\text{ K}$ . It should be noted that Class I sources are sometimes referred to as embedded protostars, Class II sources are known as classical T-Tauri stars, and Class III sources are known as weak-lined T-Tauri stars. These classes were established in an attempt to approximate an evolutionary sequence. The scenario as envisioned runs as follows: Class 0 sources are cloud cores just beginning their collapse into stars. Class I sources are stars lying in a cocoon of circumstellar material that is assembling itself into a disk and being funnelled onto the star. Class II sources are more mature stars undergoing normal disk accretion but having little residual

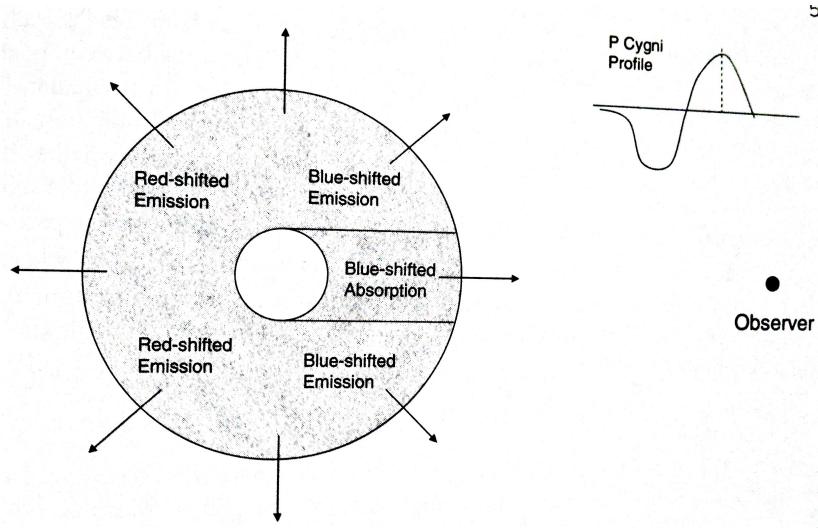


FIG. 97.—(A star surrounded by an expanding envelope will show P Cygni profiles for the stronger spectral lines. The blueshifted absorption trough is formed in the column of gas between the observer and the star, whereas the emission component is formed in the remainder of the envelope. The form of the P Cygni profile is shown in the upper-right-hand corner; the dotted line shows the rest wavelength of the spectral line. The exact shape of the profile depends on conditions in the envelope. Image taken from (Gray & Corbally 2009).

material in the envelope. Class III sources are stars that have moved beyond the accretion stage but have yet to reach the MS (Gray & Corbally 2009, pg. 379).

The most unambiguous determination of a pre-MS star's atmospheric properties and accretion signature is to acquire spectroscopy in the optical or NIR where the emitted photospheric flux is maximum. For Class III and II sources, this is relatively easy with modes spectrographs, but Class I sources are very difficult, with Class 0 sources still beyond reach. Figure 97 shows optical spectra for representative Class I and Class I/II sources. In the Class I source the emission is probably dominated by a jet of collimated material associated with the young protostar. In the Class I/II source the spectrum is dominated by emission from the accretion disk or scattering off the circumstellar envelope. The Class II sources (not shown here) show less obscuration and veiling so that true photospheric features such as the familiar TiO bands of an M dwarf begin to appear (Gray & Corbally 2009, pg. 382).

#### T TAURI STARS

The prototype low-mass YSO is a T Tauri star (TTS), which is a name often used in reference to the entire class of YSOs. Historically, TTSs were recognized as a distinct class of emission-line variable stars. Traditionally, TTSs are divided into *weak-line T Tauri stars* (WTTSs) and those with stronger Balmer-line emission (as determined by its equivalent width), the *classical T Tauri stars* (CTTSs). As well as the hydrogen lines, the emission can also be seen in Ca II, Fe II, and Na I, and sometimes in [O I] and [S II] lines; hence, the emission lines resemble those from the solar chromosphere. Underlying these is an absorption spectrum, generally veiled, which can be anything from G to M. Abnormally strong Li absorption at  $\lambda 6708$  and P Cygni-type line profiles, especially for the H $\alpha$  and the Ca II H and K lines, are also characteristic of TTSs. The former indicates that these are young objects, and the latter that they have mass outflow. The H $\alpha$  and Ca II H and K emission lines are most readily used in the classification of TTSs (Gray & Corbally 2009, pg. 275).

#### P CYGNI PROFILE

A P Cygni profile is a line profile characterized by a blueshifted absorption trough, coupled with an emission component (see Figure ??). A P Cygni profile originates in an expanding stellar atmosphere, for instance, a stellar wind, and therefore is a spectral diagnostic of the presence of a wind, especially in early-type stars. Some stars show lines with an *inverse P Cygni profile*, which consists of a redshifted absorption trough combined with an emission component. Such profiles are indicative of mass infall, and are commonly found in pre-MS objects such as the Herbig AeBe stars. Occasionally, only the P Cygni absorption troughs are visible in the spectrum, without the appearance of an emission component (Gray & Corbally 2009, pg. 549).

**QUESTION 19**

Sketch the spectral energy distribution (SED) of a T Tauri star surrounded by a protoplanetary disk. How would the SED change (a) if the disk develops a large inner hole, (b) if the dust grains in the disk grow in size?

### QUESTION 19

**Sketch the spectral energy distribution (SED) of a T Tauri star surrounded by a protoplanetary disk. How would the SED change (a) if the disk develops a large inner hole, (b) if the dust grains in the disk grow in size?**

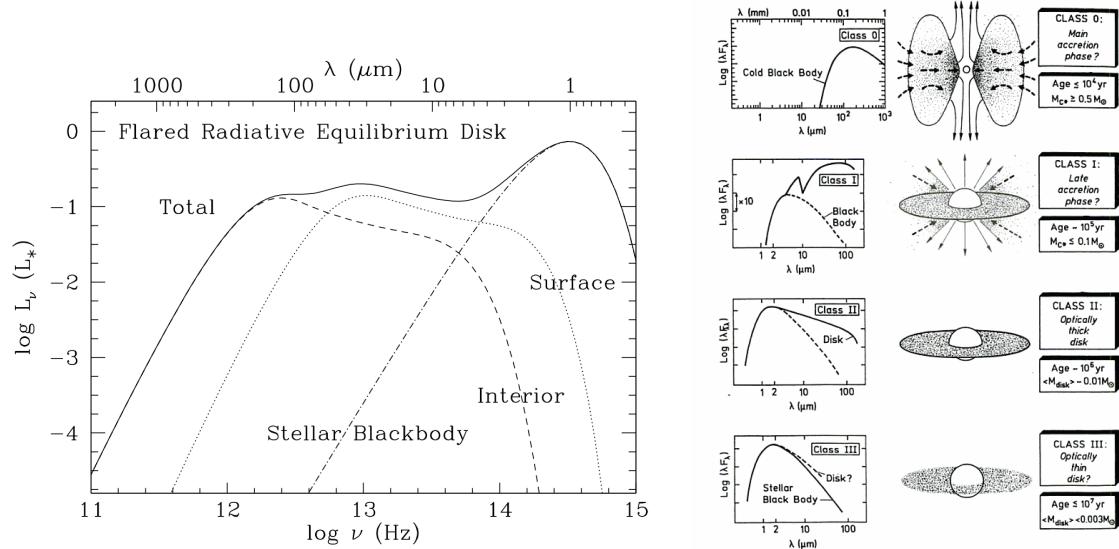


FIG. 98.— (left) SED for a three component TTS system considered by Chiang & Goldreich (1997). In their work they model the central star as a spherical blackbody of temperature  $T = 4000$  K and mass  $M = 0.5 M_\odot$ . The disk contains an interior and surface layer with an overall uniform composition of dust and gas, with the dust comprising 1% of the total mass. They show that this model can be fit well to observational data of TTSs, and show this explicitly for the system GM Aur. (right) The classification scheme based on protostellar and TTS SEDs. This displays the evolutionary progression from a collapsing protostar (Class 0) to a PMS star (Class III). Classical TTSs are generally classified as Class II. Image taken from Montmerle (1994, pg. 189).

T Tauri stars (TTSs) are an important class of low-mass pre-main-sequence (PMS) objects that represent a transition between stars that are still shrouded in dust and main-sequence stars. Consequently, they represent a class of very young ( $\sim 1 - 10$  Myr), low mass ( $0.5$  to  $\sim 2 M_\odot$ ), PMS stars that radiate most of their energy in the NIR regime. TTSs fall on the fully convective or the beginning of the partially radiative branches of the hydrostatic PMS evolutionary tracks on the HR diagram (Appenzeller & Mundt 1989).

TTSs are commonly characterized by their unusual spectral features. For instance, they display a variety of optical features, most notably in the form of strong hydrogen Balmer and Ca II (the H and K lines) emission lines. Also present are lithium absorption features and [O I] and [S II] forbidden lines. Interestingly, the H $\alpha$  line often exhibits a P Cygni profile; a spectral feature that arises from an expanding mass shell during a period of significant mass loss (Carroll & Ostlie 2007, pg. 436). Moreover, TTSs are known for displaying strong IR excess emission that often exceeds the optical emission by a large factor. Many are also identified as strong sources of FIR and submm radiation which appear to arise from dust concentrated in thin circumstellar disks. The presence of dust is further confirmed by the occurrence of  $10 \mu\text{m}$  SiO dust emission. Based on observational values of the FIR-submm flux, disk masses are estimated to range from  $10^{-3}$  to  $1 M_\odot$  (Appenzeller & Mundt 1989).

Another common feature between TTSs are the rather large and fairly rapid variations they portray in their luminosity and spectral features. Irregular flux variations have been observed in all wavelength bands except the FIR-submm range. The most conspicuous are at X-ray and UV energies with variability on the timescale of hours. The observed amplitude distribution for these variations follows well with the corresponding relation found for solar flare events. Similar stellar flare events seem to be responsible for the variations in emission line spectra that can occur on timescales as short as 10 minutes (Appenzeller & Mundt 1989).

An explanation for the observational traits of TTSs is to consider them as very active, but otherwise more or less normal late-type stars, surrounded by accretion disks. The enhanced activity of the central star can be understood as a natural consequence of more rapid rotation (Appenzeller & Mundt 1989).

To determine the SED we would expect for such a system we can follow Chiang & Goldreich (1997) and consider a simplified model of a flared protoplanetary disk in radiative equilibrium with the central star. The body of the disk flares outward with increasing distance from the star as a consequence of vertical hydrostatic equilibrium. A simplified description of the disk distinguishes two regions: (1) a surface layer containing grains that are directly exposed to light from the central star and (2) the shielded interior of the disk. Stellar radiation strikes the surface layer and penetrates to a visible optical depth of unity. This superheats dust in the surface layer causing it to radiate strongly in IR emission. About half of this emission escapes to space as blackbody radiation while the remaining is emitted toward the mid plane and regulates the temperature of the cooler disk interior.

The combined SED that emerges from this three component model of a T Tauri system is displayed in Figure 98. The final shape results from the combination of the blackbody continuum of the central star and the integrated emission from the surface and interior regions of the disk. The flatness of the disk components arises from the combination of blackbody emission coupled

to a temperature gradient within the disk. Over most of the IR, the SED is dominated by emission from the optically thick interior. The surface layer radiates more than the interior shortward of  $10\ \mu\text{m}$ , although its contribution in this region is hidden by the central star. Most of the radiation longward of 1 mm comes from the outer, optically thin part of the disk.

Obviously, if the disk were to develop a large inner hole, there would be a reduction in radiation absorbed by the disk (see Chen et al. 2009). This would result in a decrease in the temperature of the disk, with a corresponding reduction in its blackbody emission. The disk components in Figure 98 would therefore shift to lower values. For this reason some TTSs observed with low flux in the  $10\ \mu\text{m}$  range are identified as containing an inner hole. However, Chiang & Goldreich (1999) warn that the same effect can be mimicked by observing the system with a nonzero disk inclination. In this case, the line-of-sight radiation from the inner disk can be absorbed by the flared outer disk, thereby producing a low  $10\ \mu\text{m}$  flux, without the need for an inner hole.

We can expect the opposite result to occur if the dust grains in the disk were to grow in size. In this case, the absorption cross section of the grains would increase causing the disk to become more optically thick to the incident radiation (see, for example, Draine 2011). The overall temperature and blackbody emission of the disk would increase, shifting the disk components in Figure 98 upwards.

We will conclude by noting that the SEDs of TTSs have been classified into three groups as Class I, Class II, and Class III. The classification scheme is based on the relative strength of the radiation from the disk to the stellar blackbody radiation<sup>64</sup>. The classification has been further extended to protostars, which are the predecessors of TTSs, and are assigned the Class 0 designation. The classification system is believed to represent an evolutionary progression as depicted in Figure 98.

<sup>64</sup> For a detailed discussion of the different groups, see [http://th.nao.ac.jp/~tomisaka/Lecture\\_Notes/StarFormation/3/node7.html](http://th.nao.ac.jp/~tomisaka/Lecture_Notes/StarFormation/3/node7.html) and also the previous question.

**QUESTION 20**

**What are the primary origins of the heat lost to space by Jupiter, Earth, and Io?**

## QUESTION 20

### What are the primary origins of the heat lost to space by Jupiter, Earth, and Io?

We begin by noting that geochemical evidence suggests that planets (or at least their surface layers) were partly molten by the time they finished forming. Lunar rock chemistries and ages seem to require that the Moon was initially (4.5 Gyr ago) covered by a **magma ocean** – a molten layer at least several hundred km deep – whereas meteorite parent bodies apparently also melted and resolidified some 4.5 Gyr. Recent dating of 4.4 Gyr old Australian zircons indicates continental crust formed on Earth by that time (Hartmann 2005, pg. 205). Several heat sources are possible for this initial melting, and may also be important in the future evolution of the thermal history of planets:

- **Gravitational Contraction:** The planets formed out of material that had to collapse from a much larger size, releasing significant amounts of gravitational potential energy as heat.
- **Planetesimal Impact:** The impact of minor bodies onto planets is highly inelastic, and therefore adds heat to planets.
- **Radioactive Decay:** Radioactive isotopes within planets eject gamma radiation and particles, and the scattering of these radioactive products off other atoms produces heat. Long-lived isotopes found today like uranium, potassium, and thorium produced more heat at the beginning, but not enough to melt planets in the first few Myr. Short-lived isotopes, such as aluminum-26 and iodine-129, were so abundant that they produced enough heat to melt planets a few Myr after they formed.
- **Differentiation:** Within a planet, lighter material tends to rise, and heavier material tends to sink, which releases gravitational potential energy through friction (the total energy released is the total potential energy of the system before and after differentiation). This process is based on the same principle as gravitational contraction, but is not identical, since the radius of the entire planet does not decrease.
- **Rotational Spin-down:** Tides between a planet and a large moon can spin the planet down (if the planet's spin angular velocity exceeds the orbital angular velocity of the moon).
- **Solar Radiation:** Photons are absorbed by the surfaces of planets, adding to their thermal energy budget. Solar wind may also contribute through electromagnetic induction heating (a process much more important in asteroids than planets, and during the early Solar System, with its significantly stronger T Tauri wind).

Heat produced in a planet's interior is transported outward through convection in molten regions of the interior, and through conduction in solid regions. At the surface, heat is radiated back into space.

### COOLING TIME

At the time of formation, planets and moons were hot. The internal heat takes time to escape into space, and as the form of heat escape is in blackbody radiation into space, the rate of energy loss is proportional to the surface area ( $R^2$ ). The total amount of heat is proportional to the volume ( $R^3$ ). Therefore the cooling time is roughly proportional to  $R$  meaning that larger bodies should retain their internal heat for longer (Yevgeni).

As a rough estimate of the initial internal energy of the Earth, we can use its current central temperature,  $T_C$ , and apply it to the whole planet. Of course, the rest of the planet is cooler than its centre, but at the same time, the planet would have been much hotter at the time of formation. We thus have that

$$E_{\text{int}} \approx \frac{3}{2} NkT \approx \frac{3}{2} \frac{M}{\mu m_H} kT_C, \quad (385)$$

where we can take  $\mu = 60$  as appropriate for  $\text{SiO}_2$  being the most abundant crust compound. Since energy is lost through blackbody radiation, the rate of heat escape is  $\dot{E} = RA$ , where  $A$  is the surface area of the planet and  $R$  is its flux of heat escape. For an average surface temperature of  $T_S = 287$  K, we would expect  $R \sim 10^3 \text{ W m}^{-2}$ . The cooling timescale is thus

$$\tau_{\text{cool}} \approx \frac{E_{\text{int}}}{\dot{E}} \sim 10^5 \text{ yr.} \quad (386)$$

Obviously, this timescale is much too short since we do not live in a frozen world. Of course, we expect this crude approximation to underestimate the actual timescale since it neglects the influx of internal energy due to solar radiation and ongoing radioactive decay as well as processes such as global warming that tend to suppress the amount of energy lost to space. Using the measured radiance of  $R \sim 10^{-1} \text{ W m}^{-2}$  we arrive at a more realistic timescale of  $\tau_{\text{cool}} \sim 10^{10} \text{ yr}$  (Yevgeni).

### EARTH

Early in the life of the solar system, gravitational contraction, accretion of planetesimals and rapid radioactive decay of isotopes such as  $^{26}\text{Al}$  and  $^{129}\text{I}$  created molten interiors inside terrestrial planets, which enables differentiation to occur. Over time, terrestrial planets release their heat via radiation until they come into radiative balance with the solar radiation impinging on their surfaces – the flux of heat escaping the Earth is  $0.078 \text{ W m}^{-2}$  (Carroll & Ostlie 2007, pg. 752). Giant planets obtain much of their initial heat from gravitational contraction, and likewise slowly radiate it away over time.

The solar flux on Earth is  $1365 \text{ W m}^{-2}$ , meaning the heat radiated from the Earth's interior is  $10^{-4}$  the reflected/absorbed-then-reemitted light from the Sun. The Earth's albedo is 0.3, meaning that  $\sim 1000 \text{ W m}^{-2}$  of sunlight is absorbed. Thus, most of the heat lost to space by Earth is actually absorbed sunlight. In addition, since the heat received by the surface from the Sun is some four orders of magnitude greater than the escaping heat flux, the equilibrium temperature of the Earth is determined by solar radiation and *not* by internal heat (Cole & Woolfson 2002, pg. 46). The majority of internal heat lost to space comes from a combination of remnant internal heat from accretion/gravitational-contraction/early radioactive decay (20% of thermal energy lost), and modern radioactive decay of heavy elements like  $^{238}\text{U}$ ,  $^{235}\text{U}$ ,  $^{232}\text{Th}$  and  $^{40}\text{K}$ , all with half-lives in the Gyr range (80% of thermal energy lost). Radioactive elements heat about  $10^{-11} \text{ W kg}^{-1}$  of mantle (which makes up the majority of mass on Earth), giving a total energy generation rate of  $\sim 10^{13} \text{ W}$ , translating to  $\sim 0.1 \text{ W m}^{-2}$ .

### *JUPITER*

From the albedo of Jupiter, measured as 0.343, it is possible to estimate how much energy it absorbs from the Sun and from an analysis of the spectrum of Jupiter it is possible to estimate its effective temperate as a radiator of energy; the average temperature is 125 K. Taking this average temperature it can be shown that Jupiter emits about twice as much radiation as it receives from the Sun. The mechanism that gives rise to this excess radiated energy is not known but two main mechanisms have been advanced. Firstly, gravitational contraction of Jupiter may still be occurring, with a contraction rate of the radius of less than 0.05 m per century releasing enough gravitational energy to explain the excess. Note that this rate of collapse over the lifetime of the planet (4.5 Gyr) would have reduced its radius by only 3%, so the explanation is feasible. Secondly, if helium separates out from hydrogen and sinks toward the centre of the planet then this will release gravitational energy. The required rate of separation to explain the excess heat is tiny compared with the total amount of helium in the planet, and the change in the distribution of helium since the planet formed would be negligible (Cole & Woolfson 2002, pg. 70).

### *IO*

Because of its proximity to Jupiter, Io experiences strong tidal forces; the resultant internal stress is the main source for Io's internal heat supply. Even though the moon's rotation period is the same as its orbital period, small deviations from a perfectly circular orbit means that its orbital velocity is not constant. Consequently, the moon tends to wobble, not quite keeping one side locked in place toward Jupiter. This effect is due to the curious resonance that exists among the orbits of Io, Europa, and Ganymede. Their orbital periods form ratios that are approximately 1:2:4, meaning that both Europa and Ganymede perturb Io's orbit at about the same location each time Io orbits Jupiter. This forces Io's orbit to remain slightly elliptical (Carroll & Ostlie 2007, pg. 795).

Io is also heated by Jupiter's magnetic field. Since Jupiter rotates in just under 10 hrs, whereas Io orbits the planet in about 2 days, Jupiter's magnetic field sweeps past Io at a speed of  $\sim 60 \text{ km s}^{-1}$ . This motion through the magnetic field sets up an electrical potential difference across the moon. The potential difference acts much like a battery, causing a large current to flow back and forth along magnetic field lines between Io and Jupiter. This current flow of charged particles also generates Joule (i.e., resistive) heating within the moon, analogous to a resistor in a circuit. Roughly  $10^{12} \text{ W}$  of power is generated in this way. However, this contribution to the total internal heating of the moon is only a small fraction of the total energy liberated from the surface per second, which is roughly  $10^{14} \text{ W}$  (Carroll & Ostlie 2007, pg. 795).

Io receives about  $50 \text{ W m}^{-2}$  on its surface, and has a radius of about 2000 km, meaning that it receives about  $10^{14} \text{ W}$  of Solar energy. Its main source of internal heat is tidal stress from Jupiter which generates some  $10^{13} \text{ W}$ . While most of Io's heat radiated into space comes from the Sun, Io's internal heat keeps it volcanically active.

### *OTHER PLANETS*

Saturn radiates 2.3 times the heat it receives from the Sun, likely due to differentiation. Cole & Woolfson (2002, pg. 78) suggests Uranus actually may be emitting less energy than it receives from the Sun, but Hartmann (2005) suggests this is not the case, and Uranus radiates about the same heat it receives from the Sun. Neptune also radiates more heat than it receives from the Sun.

**QUESTION 21**

**Consider a small test-mass orbiting around a compact object. What are the essential differences in the properties of its orbits between Newtonian gravity and General Relativity?**

### QUESTION 21

**Consider a small test-mass orbiting around a compact object. What are the essential differences in the properties of its orbits between Newtonian gravity and General Relativity?**

The information for this question comes from a combination of Charles Dyers's lecture notes, Harald Pfeiffer's lecture notes, and (Blau 2010, pgs. 120 -138).

#### SCHWARZSCHILD SPACETIME

The paths of photons and time-like particles in the Schwarzschild spacetime differ from those of similar particles in Newtonian gravitation. We begin by writing the Schwarzschild line element in the usual spherical coordinate system:

$$ds^2 = \left(1 - \frac{2m}{r}\right) dt^2 - \left(1 - \frac{2m}{r}\right)^{-1} dr^2 - r^2 (d\theta^2 + \sin^2\theta d\phi^2), \quad (387)$$

where  $m \equiv GM/c^2$ . The geodesics in this spacetime can be obtained by solving the Euler-Lagrange equations for the Lagrangian function based on this line element, which we take to be:

$$L = \left(1 - \frac{2m}{r}\right) t^2 - \left(1 - \frac{2m}{r}\right)^{-1} r^2 - r^2 (\dot{\theta}^2 + \sin^2\theta \dot{\phi}^2), \quad (388)$$

where overdots are with respect to the affine parameter  $\tau$ . Since this situation corresponds to the central-force situation in Newtonian gravity (i.e., isotropic gravitational field), it is reasonable to expect that each non-radial geodesic will be planar. In this case, we can arbitrarily choose to orient our coordinate system so that this plane is defined by  $\theta = \pi/2$ , so that  $\dot{\theta} = 0$  and the penultimate term in equation (388) drops out.

The time-translation and rotational Killing vectors (i.e.  $\partial L/\partial\phi = \partial L/\partial t = 0$ ) result in conservation of energy and angular momentum:

$$\begin{aligned} E &= \left(1 - \frac{2m}{r}\right) \dot{t}, \\ L &= r^2 \dot{\phi}. \end{aligned} \quad (389)$$

Using these results, we can now rewrite equation (387) as

$$\frac{1}{2} \dot{r}^2 + V(r) = \frac{1}{2} E^2 \quad \text{where} \quad V(r) \equiv \frac{1}{2} \kappa - \kappa \frac{m}{r} + \frac{L^2}{2r^2} - \frac{mL^2}{r^3}. \quad (390)$$

Here  $\kappa = 1$  for time-like massive particles and  $\kappa = 0$  for massless photons. The form of equation (390) guarantees that particles with energy  $E$  are restricted to radii for which  $V(r) < E^2/2$ .

In contrast, the Newtonian potential that would fit into equation (390) is

$$V(r) = -\frac{m}{r} + \frac{L^2}{2r^2}, \quad (391)$$

where the last term,  $L^2/2r^2 = r^2\dot{\phi}^2$  arises from switching to polar coordinates (i.e., centrifugal force term). We thus see that the general relativistic correction is the last term in equation (390),  $mL^2/r^3$ , and note that as  $r \rightarrow \infty$ , this term goes to zero.

#### MASSIVE PARTICLES

To understand the case of massive particles, let's first determine the extrema of the potential  $V(r)$ . For a given  $L$ , the orbits of massive particles have extreme at the values of  $r_c$  satisfying

$$0 = \frac{\partial V}{\partial r}(r_c) \Rightarrow r_c = \frac{L^2}{2m} \pm \sqrt{\frac{L^4}{4m^2} - 3L^2}. \quad (392)$$

Existence of roots implies that  $L > L_{\text{ISCO}} = \sqrt{12}m$ .

Let's first consider the case for which  $L < L_{\text{ISCO}}$ . In this case, there are no real turning points and the potential looks approximately like that in the left panel of Figure 99. From this picture we can read off that there are no bounded orbits for these values of parameters. Any inward bound particle with  $L < 2m$  will continue to fall inwards (provided that it moves on a geodesic). This should be contrasted with the Newtonian situation in which for any  $L \neq 0$  there is always the centrifugal barrier reflecting incoming particles since the repulsive term  $L^2/2r^2$  will dominate over the attractive  $-m/r$  for small values of  $r$ . In GR, on the other hand, it is the attractive term  $-mL^2/r^3$  that dominates for small  $r$ .

Fortunately for the stability of the solar system, the situation is qualitatively quite different for sufficiently large values of the angular momentum, namely  $L > L_{\text{ISCO}} = \sqrt{12}m$ ; see the potential landscape in the right panel of Figure 99. In this case, there is a minimum and a maximum of the potential. The critical radii correspond to exactly circular orbits, unstable at  $r_-$  (on top of the potential) and stable at  $r_+$  (the minimum of the potential). For  $L \rightarrow L_{\text{ISCO}}$ , these two orbits approach each other, the critical radius

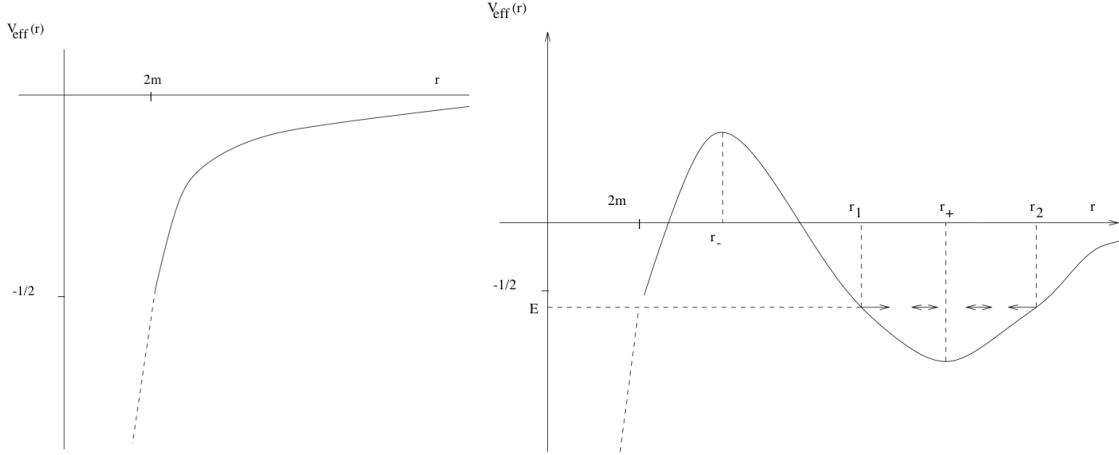


FIG. 99.— (left) Effective potential for a massive particle with  $L < \sqrt{12}m$ . The extrapolation to values of  $r < 2m$ , which is done with caution, is represented by the dashed line. (right) Effective potential for a massive particle with  $L > \sqrt{12}m$ . Shown are the maximum of the potential at  $r_-$  (an unstable circular orbit), the minimum at  $r_+$  (a stable circular orbit), and the orbit of a particle with  $E < 0$  with turning points  $r_1$  and  $r_2$ . Images taken from Blau (2010).

tending to  $r_{\text{ISCO}} = 6m$ , called the **innermost stable circular orbit**. For other values of  $L$ , and with sufficiently large values of  $E$ , a particle will fall all the way down the potential. For  $E < 0$  there are bound orbits which are not circular and which range between the radii  $r_1$  and  $r_2$ , the turning points at which  $E = V(r)$ .

Because of the general relativistic correction  $\sim 1/r^3$ , the bound orbits will not be closed (elliptical). In particular, the position of the perihelion, at the radial separation  $r_1$  will not remain constant. However, because  $r_1$  is constant, and the planetary orbit is planar, this point will move on a circle of radius  $r_1$  around the sun. One can easily find this **perihelion precession** by considering elliptical orbits that vary slightly from being circular. For example, if we write  $r(t) = r_c + \delta(t)$ , take a time derivative of equation (390), and then Taylor expand, we find

$$\ddot{\delta} + V''(r_c)\delta = 0, \quad (393)$$

that is, harmonic oscillation with frequency  $\omega_r = \sqrt{V''(r_c)}$ . In the case of Newtonian gravity,  $\omega_r = \dot{\phi}$  so the radial oscillation period and orbital periods agree; the orbit is thus closed. For general relativity, on the other hand, we find that

$$\omega_r \approx \dot{\phi} \left( 1 - \frac{3m}{r_c} \right). \quad (394)$$

Hence, the radial oscillations have *lower* frequency than the orbital frequency, indicating that **perihelion precession** will ensue. This effect is observable in the solar system for the planet Mercury, whose perihelion was known to advance about  $43''$  per century over and above already known Newtonian gravitational effects caused by perturbations from other planets in the solar system.

An interesting result from this analysis is that accretion disks of compact objects will end at the location  $r_{\text{ISCO}}$ ; at closer regions the particles are destined to fall into the object. Particles at the inner edge of the disk will have an associated energy per unit rest-mass of

$$E_{\text{ISCO}} = \sqrt{2V(r_{\text{ISCO}})} = \sqrt{8/9} = 0.943. \quad (395)$$

The difference to 1 (i.e., about 5.7% of the rest mass) represents gravitational binding energy, which was emitted during the slow inspiral of the particle to the innermost stable circular orbit.

As an aside, Harald's notes state that the smallest radii achievable for eccentric bound orbits is  $r_{\text{mb}} = 4m$ ... don't know how to get this though.

#### PHOTON PATHS

For null geodesics ( $\kappa = 0$ ), equation (390) can be rewritten as

$$\dot{r}^2 + \frac{1}{r^2} - \frac{2m}{r^3} = \frac{E^2}{L^2}, \quad (396)$$

so that null geodesics depend only on the **impact parameter**  $b \equiv L/E$ . As shown in Figure 100, this potential has no local minima, so *no stable bound orbits exist*. Its maximum is obtained at  $r = 3M$ , known as the **photon sphere** where photons can circle the centre of symmetry indefinitely. For light rays approaching from  $r = \infty$ , their orbits will depend on their impact parameter:

- $b < \sqrt{27}m$ : photons are captured by the star and will spiral into it.
- $b > \sqrt{27}m$ : there will be a turning point, so that the light ray will be deflected by the source.

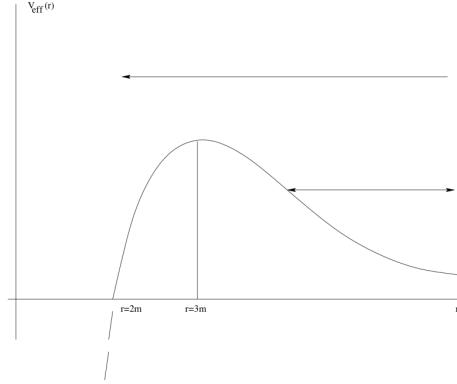


FIG. 100.— (left) Effective potential for a massless particle. Displayed is the location of the unstable circular orbit at  $r = 3m$ . A photon with an energy  $E^2 < L^2/27m^2$  will be deflected (lower arrow), photons with  $E^2 > L^2/27m^2$  will be captured by the star. Images taken from Blau (2010).

This may sound counterintuitive in that we may suspect that a higher energy photon would be more likely to zoom by the compact object without being forced to spiral into it. We can think of this in the following way:  $L = 0$  corresponds to a photon falling radially towards the star, small  $L$  corresponds to a slight deviation from radial motion, while large  $L$  (thus large  $\dot{\phi}$ ) means that the photon is travelling along a trajectory that will not bring it very close to the star at all. It is then not surprising that photons with small  $L$  are more likely to be captured by the star than photons with large  $L$  that will only be deflected in their path. Note also that the above description implies that the cross-section for light-capture into a compact object is actually  $27\pi m^2$ , and not  $4\pi m^2$ , as one might suspect.

It can be shown that for a light ray with impact parameter  $b$ , its trajectory will be bent through the angle

$$\delta \approx \frac{4m}{b}. \quad (397)$$

For light just passing the sun the predicted value is  $\delta \sim 2''$ . This effect is difficult to observe, since observation of stars when they lie very close to the bright disk of the sun is difficult. A total solar eclipse of the sun by the moon does provide a situation where the observation is possible, since the moon obscures the bright visible disk of the sun. This observation was performed in 1919 by Sir Arthur Eddington and collaborators. Their results provided reasonable confirmation of the Einstein predicted deflection angle. This was important because von Soldner had concluded in 1801 that based on Newtonian gravitation, and treating the light ray as that of a very small particle moving at the speed of light, there should be a deflection angle of  $2m/a$ , that is half the deflection predicted by general relativity (although the Newtonian prediction is flawed since light has no mass). Eddington's result provided sufficient precision to show that the Einstein result was the more correct result, thus providing early experimental support for the then new theory of general relativity.

**QUESTION 22**

**Write out the p-p cycle. Summarize the CNO cycle.**

**QUESTION 22**  
**Write out the p-p cycle. Summarize the CNO cycle.**

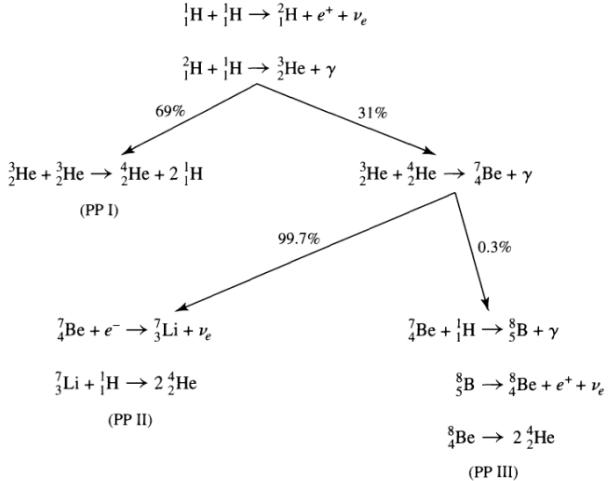


FIG. 101.— The three branches of the pp-chain, along with the branching ratios appropriate for conditions in the core of the Sun. Image taken from (Carroll & Ostlie 2007).

*STELLAR NUCLEOSYNTHESIS*

In any nuclear reaction process, baryon number, lepton<sup>65</sup> number, and charge must all be conserved. Note that antiparticles subtract from these numbers, so that particle-particle pair results in zero change in any of these numbers. For instance, when counting the number of leptons involved in a nuclear reaction, we treat matter and antimatter differently. Specifically, the total number of matter leptons *minus* the total number of antimatter leptons must remain constant. To assist in counting the number of nucleons and total electric charge, nuclei are commonly represented like  ${}^A_Z\text{X}$  where X is the chemical symbol of the element, Z is the number of protons, and A is the mass number (i.e. the total number of nucleons, protons plus neutrons) (Carroll & Ostlie 2007, pg. 309). The energy liberated or captured in a nuclear reaction can be determined by the difference in net rest mass between the reactants and products ( $\Delta E = \Delta mc^2$ ) - this energy difference is the difference in nuclear strong force binding energy between the two sets of particles. Binding energy is simply the energy required to separate nucleons against their mutual attraction through the strong, but short-range nuclear forces; equivalently it is the energy gained if the particles are brought in from infinity (Kippenhahn & Weigert 1990, pg. 147).

*THE PROTON-PROTON CHAIN*

Applying the conservation laws, one chain of reactions that can convert hydrogen into helium is the **proton-proton chain (pp-chain)**. It involves a reaction sequence that ultimately results in



through the intermediate production of deuterium ( $^2\text{H}$ ) and helium-3 ( $^3\text{He}$ ); strictly speaking, this reaction is called the **first pp chain (PPI)**. The entire PPI reaction chain is:



Note that 0.4% of the time, the first reaction step is accomplished by the so-called pep reaction:  ${}^1\text{H} + e^- + {}^1\text{H} \rightarrow {}^2\text{H} + \nu_e$  (Carroll & Ostlie 2007, pg. 309). The net energy gain through reaction (398) is 26.731 MeV, ten times the energy liberated from any other nuclear fusion process (2–30% of this energy is carried away in neutrinos). The rate at which  $4^1\text{H}$  atoms can come together and tunnel to form an  ${}^4\text{He}$  atom is tiny compared to the rate of a more energetically favourable chain known as the proton-proton chain (Charles). Each step in reaction (399) has its own reaction rate, since different Coulomb carries and cross sections are involved. The slowest in the sequence is the initial one, because it involves the  $\beta$ -decay of a proton into a neutron via  $p^+ \rightarrow n + e^+ + \nu_e$ . Also, to be exacting two  ${}^1\text{H}$  atoms that entered through reaction (399) did so as catalysts, meaning that PPI can be called a pp-cycle (Carroll & Ostlie 2007, pg. 309).

<sup>65</sup> Lepton means “light thing” and includes electrons, positrons, neutrinos, and antineutrinos.

The production of helium-3 nuclei in the PPI chain also provides for the possibility of their interaction directly with helium-4 nuclei, resulting in a second branch of the pp-chain. In an environment characteristic of the centre of the Sun, 69% of the time a helium-3 interacts with another helium-3 nuclei in the PPI chain, whereas 31% of the time the **PPII** chain occurs:



Yet another branch, the **PPIII** chain, is possible because the capture of an electron by the beryllium-7 nucleus in the PP II chain competes with the capture of a proton (a proton is captured only 0.3% of the time in the centre of the Sun):



The branching ratios of the three different chains (i.e. the percentage of p-p reactions that utilize each chain) depends on the exact temperature, density and chemical composition of the material undergoing fusion. Figure 101 gives values for the interior of the Sun (Carroll & Ostlie 2007, pg. 310).

The nuclear energy generation rate is given by the rate-limiting first step of the chain. This rate, when translated to emitted power per unit mass, yields

$$\epsilon_{\text{pp}} = 0.241 \rho X^2 f_{\text{pp}} \psi_{\text{pp}} C_{\text{pp}} T_6^{-2/3} e^{-33.8 T_6^{-1/3}} \text{ W kg}^{-1}, \quad (402)$$

where  $f_{\text{pp}} \simeq 1$  is the pp-chain electron screening factor,  $\psi_{\text{pp}} \simeq 1$  is a correction factor that accounts of the simultaneous occurrence of the three branches, and  $C_{\text{pp}} \simeq 1$  involves higher-order correction terms. When written as a power law near  $T \sim 10^7$  K, the energy generation rate has the form

$$\epsilon_{\text{pp}} \approx (10^{-12} \text{ W kg}^{-1}) \rho X^2 f_{\text{pp}} \psi_{\text{pp}} C_{\text{pp}} T_6^4 \propto \rho T^4. \quad (403)$$

Hence, the power law form of the energy generation rate demonstrates a relatively modest temperature dependence of  $T^4$  near  $T_6 = 10$  (Carroll & Ostlie 2007, pg. 311).

#### THE CNO CYCLE

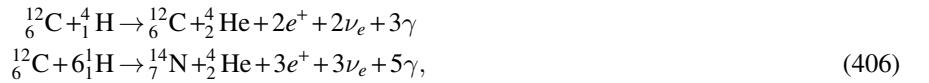
A second, independent cycle also exists for the production of helium-4 from hydrogen. In the so called **carbon-oxygen-nitrogen cycle** (CNO cycle), carbon, nitrogen, and oxygen are used as catalysts, being consumed and then regenerated during the process. Just as with the pp-chain, the CNO cycle has competing branches. The first branch culminates with the production of carbon-12 and helium-4:



The second branch occurs only about 0.04% of the time and arises when the last step above produces oxygen-16 and a photon, rather than carbon-12 and helium-4:



More succinctly, the two cycles can be summarized as



where the catalysts have been left in the reaction and the notation  $3\gamma$  is used to represent three photon releases, not necessarily of the same energy (Carroll & Ostlie 2007, pg. 311).

The energy generation rate for the CNO cycle is given by

$$\epsilon_{\text{CNO}} = 10^{21} \rho X X_{\text{CNO}} C_{\text{CNO}} T_6^{-2/3} e^{-153 T_6^{-1/3}} \text{ W kg}^{-1}, \quad (407)$$

where  $X_{\text{CNO}}$  is the total mass fraction of carbon, nitrogen, and oxygen, and  $C_{\text{CNO}}$  is a higher-order correction term. When written as a power law centred around  $T \sim 10^7$  K, the energy generation becomes

$$\epsilon_{\text{CNO}} \approx (10^{-30} \text{ W kg}^{-1}) \rho X X_{\text{CNO}} T_6^{19.9} \propto \rho T^{20}. \quad (408)$$

We see that the CNO cycle is much more strongly temperature-dependent than is the pp chain. This property implies that low-mass stars, which have smaller central temperatures, are dominated by pp chains during their “hydrogen burning” evolution, whereas more massive stars, with their higher central temperatures, convert hydrogen to helium by the CNO cycle (Carroll & Ostlie 2007, pg. 312).

When hydrogen is converted into helium by either the pp chain or CNO cycle, the mean molecular weight  $\mu$  of the gas increases. If neither the temperature nor the density of the gas changes, the ideal gas law predicts that the central pressure will necessarily decrease. As a result, the star would no longer be in hydrostatic equilibrium and would begin to collapse. This collapse has the effect of actually raising both the temperature and the density to compensate for the increase in  $\mu$  (i.e. through the virial theorem). When the temperature and density become sufficiently high, helium nuclei can overcome their Coulomb repulsion and begin to burn (Carroll & Ostlie 2007, pg. 312).

#### TRIPLE ALPHA PROCESS

The reaction sequence by which helium is converted into carbon is known as the **triple alpha process**, given by



In the triple alpha process, the first step produces an unstable beryllium nucleus (the right hand side of the first reaction is 100 keV more massive than the left side, so this reaction is highly endothermic) that will rapidly decay back into two separate helium nuclei if not immediately struck by another alpha particle. As a result, the reaction may be thought of as a three-body interaction, and therefore, the reaction rate depends on  $(\rho Y)^3$ . As a power law written around  $10^8$  K (around the temperature where the first step can be achieved), we find a very large temperature dependence of

$$\epsilon_{3\alpha} \propto \rho^2 T^{41}. \quad (410)$$

With such a strong dependence, even a small increase in temperature will produce a large increase in the amount of energy generated per second (Carroll & Ostlie 2007, pg. 313).

#### CARBON AND OXYGEN BURNING

In the high-temperature environment of helium burning, other competing processes are also at work. After sufficient carbon has been generated by the triple alpha process, it becomes possible for carbon nuclei to capture alpha particles, producing oxygen. Some of the oxygen in turn can capture alpha particles to produce neon:



At helium burning temperatures, the continued capture of alpha particles leading to progressively more massive nuclei quickly becomes prohibitive due to the even higher Coulomb barrier. If a star is sufficiently massive, still higher central temperatures can be obtained and many other nuclear products become possible. In particular, carbon burning reactions occur at  $T \approx 6 \times 10^8$  with the general form  ${}_6^{12}\text{C} + {}_6^{12}\text{C} \rightarrow X$  where  $X$  includes things like O, Ne, Na, and Mg. At even higher temperatures of  $T \approx 10^9$  K, oxygen burning reactions of the form  ${}_8^{16}\text{O} + {}_8^{16}\text{O} \rightarrow X$  can occur, with  $X$  including Mg, Si, P, and S. Some of these latter carbon and oxygen burning reactions are highly endothermic and therefore have to be powered by gravitational collapse or other exothermic reactions (Carroll & Ostlie 2007, pg. 314). Fusion alone may continue to be energetically favourable until  ${}_{56}^{56}\text{Fe}$  is fused. Past this point, fusion is energetically unfavourable, and therefore must come at the expense of loss of gravitational potential energy - in stars this is what eventually causes supernovae (Charles).

#### NUCLEAR REACTION RATES

The potential energy landscape for interacting nuclei consists of two parts: a Coulomb repulsion term and a strong nuclear binding force term, as shown in Figure 102. Classically, in order for two nuclei to bind sufficient thermal energy must be provided for them to overcome the height of the Coulomb potential barrier occurring at a separation on the order of  $r \sim 1$  femtometer(fm) =  $10^{-15}$  m. The required thermal energy can be estimated as

$$\frac{1}{2} \mu v^2 = \frac{3}{2} k T_{\text{classical}} = \frac{1}{4\pi\epsilon_0} \frac{Z_1 Z_2 e^2}{r} \Rightarrow T_{\text{classical}} = \frac{Z_1 Z_2 e^2}{6\pi\epsilon_0 k r}, \quad (412)$$

which yields  $T_{\text{classical}} \sim 10^{10}$  K for a collision between two protons ( $Z_1 = Z_2 = 1$ ). The central temperature of the Sun is only  $1.57 \times 10^7$  K, so this cannot explain the observed nuclear fusion rates, even if the high-end of the Maxwell Boltzmann distribution is considered. Clearly, quantum mechanical tunnelling induced by the uncertainty principle is required for fusion to happen at an appreciable rate. As a crude estimate of the effect of tunnelling on the temperature necessary to sustain nuclear reactions, assume

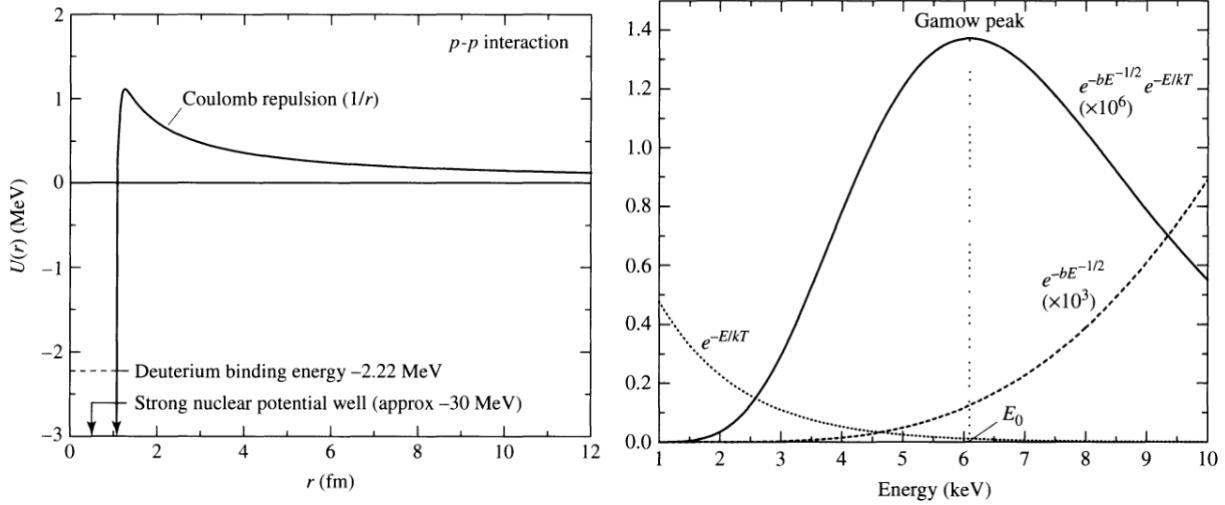


FIG. 102.— (left) The potential energy curve characteristic of nuclear reactions. The Coulomb repulsion between positive nuclei results in a barrier that is inversely proportional to the separation between nuclei and proportional to the product of their charges. The nuclear potential well inside the nucleus is due to the attractive strong nuclear force. (right) The likelihood that a nuclear reaction will occur is a function of the kinetic energy of the collision. The Gamow peak arises from the contribution of the  $e^{-E/kT}$  Maxwell-Boltzmann high-energy tail and the  $e^{-bE^{-1/2}}$  Coulomb barrier penetration term. This particular example represents the collision of two protons at the central temperature of the Sun (Note that the labelled terms have been multiplied by  $10^3$  and  $10^6$  for clarity). Images taken from (Carroll & Ostlie 2007).

that a proton must be within approximately one de Broglie wavelength of its target in order to tunnel through the Coulomb barrier. Recalling that the wavelength of a massive particle is  $\lambda = h/p$ , we have that

$$\frac{1}{2}\mu\vec{v}^2 = \frac{(h/\lambda)^2}{2\mu} = \frac{1}{4\pi\epsilon_0} \frac{Z_1 Z_2 e^2}{\lambda} \Rightarrow T_{\text{quantum}} = \frac{Z_1^2 Z_2^2 e^4 \mu}{12\pi^2 \epsilon_0^2 h^2 k}, \quad (413)$$

giving  $T_{\text{quantum}} \sim 10^7$  K for two protons ( $Z_1 = Z_2 = 1$ ;  $\mu = m_p/2$ ); in much better agreement with typical stellar values (Carroll & Ostlie 2007, pg. 301).

To obtain a more detailed analysis of nuclear reactions we will assume that we have a collection of non-relativistic gas particles obeying the Maxwell Boltzmann distribution which allows us to compute the function  $n_E dE$ , that is the number of particles with kinetic energies between  $E$  and  $E + dE$ . To describe the probability of interaction between particles we define the function

$$\sigma(E) = \frac{\text{number of reactions/nucleus/time}}{\text{number of incident particles/area/tme}}, \quad (414)$$

which can be thought of as the cross-sectional area of the target particle; any incoming particle with sufficient energy that strikes within that area, centred on the target, will result in a nuclear reaction. Let  $x$  denote a target particle and  $i$  denote an incident particle. If the number of incident particles per unit volume having emerges between  $E$  and  $E + dE$  is  $n_{iE} dE$ , then the number of reactions  $dN_E$ , is the number of particles that can strike  $x$  in a time interval  $dt$  with a velocity  $v(E) = \sqrt{2E/\mu}$ ; hence

$$dN_E = \sigma(E)v(E)n_{iE}dEdt. \quad (415)$$

Integrated over all energies, and assuming a density  $n_x$  of targets per unit volume, we obtain the reaction rate

$$r_{ix} = \int_0^\infty n_x n_i \sigma(E) v(E) \frac{n_E}{n} dE. \quad (416)$$

To evaluate this expression we must know the functional form of  $\sigma(E)$ , which is difficult since it changes rapidly with energy (Carroll & Ostlie 2007, pg. 303).

To obtain  $\sigma(E)$  we first note that it should be roughly proportional to the size of the target nucleus, and the tunnelling probability through the potential well. From the former we have that

$$\sigma(E) \propto \pi r^2 \propto \pi(\hbar/p)^2 \propto 1/E. \quad (417)$$

Hence,  $\sigma(E)$  should be inversely proportional to energy. The ability to tunnel through the Coulomb barrier is related to the ratio of the barrier height to the initial kinetic energy of the incoming nucleus. Through the use of the WKB approximation it can be shown that this process demands  $\sigma(E) \propto \exp[-2\pi^2 U_C/E]$ , where  $U_C$  is the barrier height. Estimating the barrier height from the Coulomb force and equating  $E$  with thermal energy of some speed  $v$ , it can be shown that

$$\sigma(E) \propto e^{-bE^{-1/2}} \text{ where } b \equiv \frac{\pi\mu^{1/2} Z_1 Z_2 e^2}{2^{1/2} \epsilon_0 h}. \quad (418)$$

Clearly,  $b$  depends on the masses and electric charges of the two nuclei involved in the interaction (Carroll & Ostlie 2007, pg. 304).

Combining the previous results and defining  $S(E)$  to be some (we hope) slowly varying function of energy, we may now express the cross section as

$$\sigma(E) = \frac{S(E)}{E} e^{-bE^{-1/2}}, \quad (419)$$

which through equation (416), yields the rate

$$r_{ix} = \left( \frac{2}{kT} \right)^{3/2} \frac{n_i n_x}{(\mu\pi)^{1/2}} \int_0^\infty S(E) e^{-bE^{-1/2}} e^{-E/kT} dE. \quad (420)$$

In this equation, the term  $e^{-E/kT}$  represents the high-energy wing of the Maxwell Boltzmann distribution, and the term  $e^{-bE^{-1/2}}$  comes from the penetration probability. As shown in Figure 102, the product of these two factors produces a strongly peaked curve, known as the **Gamow peak**. The top of the curve occurs at an energy of

$$E_0 = \left( \frac{bkT}{2} \right)^{2/3}. \quad (421)$$

As a consequence of the Gamow peak, the greatest contribution to the reaction rate integral comes in a fairly narrow energy band that depends on the temperature of the gas, together with the charges and masses of the constituents of the reaction (Carroll & Ostlie 2007, pg. 306).

Two additional factors affecting the reaction rate are **resonances** and **electron shielding**. The former allows  $S(E)$  to peak at specific energies corresponding to energy levels within the nucleus, analogous to the orbital energy levels of electrons. It is a resonance between the energy of the incoming particle and differences in energy levels within the nucleus that accounts for these strong peaks. On the other hand, electron shielding is a process that reduces the effective charge of the nucleus. On average, the electrons liberated when atoms are ionized at the high temperatures of stellar interiors produce a sea of negative charges that partially hides the target nucleus, reducing its effective positive charge. The result of this is a lower Coulomb barrier to the incoming nucleus and an enhanced reaction rate (Carroll & Ostlie 2007, pg. 307).

---

MATH AND GENERAL PHYSICS (INCLUDES RADIATION PROCESSES, RELATIVITY, STATISTICS)

---

**QUESTION 1**

**Draw the geometry of gravitational microlensing of one star by another, and estimate the angular displacement of the background star's image.**

### QUESTION 1

**Draw the geometry of gravitational microlensing of one star by another, and estimate the angular displacement of the background star's image.**

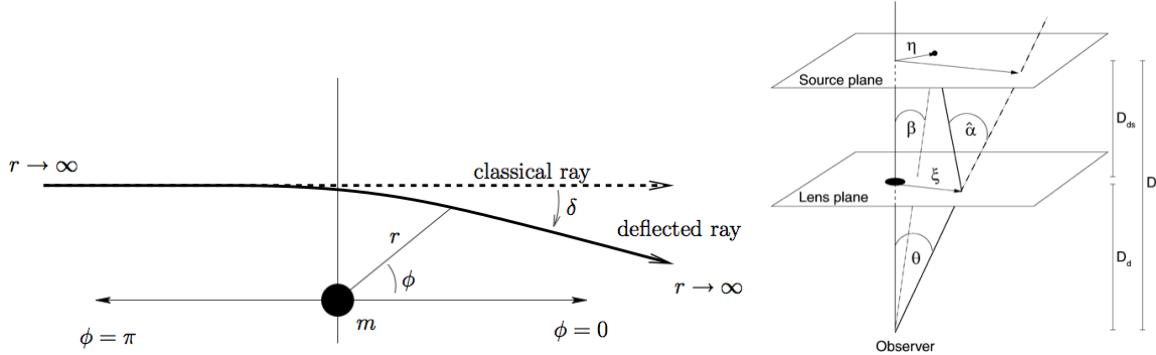


FIG. 103.— (left) Depiction of microlensing geometry. The system has been set up so that classically the photon comes in from infinity at  $\phi = \pi$  and travels back out to infinity when  $\phi = 0$ . Image taken from Charles Dyer's AST2060 Lecture Notes. (right) Geometry of a gravitational lens system. Consider a source to be located at a distance  $D_s$  from us and a mass concentration at distance  $D_d$ . An optical axis is defined that connects the observer and the centre of the mass concentration; its extension will intersect the so-called source plane, a plane perpendicular to the optical axis at the distance of the source. Accordingly, the lens plane is the plane perpendicular to the line-of-sight to the mass concentration at distance  $D_d$  from us. The intersections of the optical axis and the planes are chosen as the origins of the respective coordinate systems. Let the source be at the point  $\eta$  in the source plane; a light beam that encloses an angle  $\theta$  to the optical axis intersects the lens plane at the point  $\chi$  and is deflected by an angle  $\alpha(\chi)$ . All these quantities are two-dimensional vectors. The condition that the source is observed in the direction  $\theta$  is given by the lens equation (425) which follows from the theorem of intersecting lines. Image taken from Schneider (2002).

GR predicts that light, just like massive particles, is deflected in a gravitational field. We have seen in a previous question that a light ray passing by an object of mass  $M$  with impact parameter  $b$  is deflected through an angle

$$\delta = \frac{4GM}{c^2 b}. \quad (422)$$

This is valid so long as  $\delta \ll 1$  and is shown schematically in the left panel of Figure 103. **Classical way of estimating this result?** Inserting solar values into equation (422) we find that light deflection at the limb of the Sun is  $\delta_{\odot} \sim 2$  arcseconds. However, the Sun is not able to cause multiple images of distant sources as observed on Earth. The maximum deflection angle  $\delta_{\odot}$  is much smaller than the angular radius of the Sun, so that two beams of light that pass the Sun to the left and to the right cannot converge by light deflection at the position of the Earth. Given its radius, the Sun is too close to produce multiple images, since its angular radius is (far) larger than the deflection angle  $\delta_{\odot}$ . However, the light deflection by more distant stars (or other massive celestial bodies) can produce multiple images of sources located behind them (Schneider 2002, pg. 65).

The geometry of a gravitational lens system is depicted in the right panel of Figure 103; where the notation  $\alpha$  is used in place of  $\delta$ , and  $\xi$  in place of  $b$ . We consider light rays from a source at distance  $D_s$  from us that pass a mass concentration (called a lens or deflector) at a separation  $\xi$ . The deflector is at a distance  $D_d$  from us. In Figure 103  $\eta$  denotes the true, two-dimensional position of the source in the source plane, and  $\beta$  is the true angular position of the source, that is the angular position at which it would be observed in the absence of light deflection,

$$\beta = \frac{\eta}{D_s}. \quad (423)$$

The position of the light ray in the lens plane is denoted by  $\xi$ , and  $\theta$  is the corresponding angular position,

$$\theta = \frac{\xi}{D_d}. \quad (424)$$

Hence,  $\theta$  is the observed position of the source on the sphere relative to the position of the “centre of the lens” which we have chosen as the origin of the coordinate system,  $\xi = 0$ .  $D_{ds}$  is the distance of the source plane from the lens plane. As long as the relevant distances are much smaller than the “radius of the Universe”  $c/H_0$ , which is certainly the case within the MW and in the Local Group, we have  $D_{ds} = D_s - D_d$  (Schneider 2002, pg. 65).

From Figure 103 we can deduce the condition that a light ray from the source will reach us from the direction  $\theta$  (or  $\xi$ ),

$$\eta = \frac{D_s}{D_d} \xi - D_{ds} \alpha(\xi), \quad (425)$$

or, after dividing by  $D_s$  and using equations (423) and (424):

$$\beta = \theta - \frac{D_{ds}}{D_s} \alpha(D_d \theta) \equiv \theta - \alpha_{\text{red}}, \quad (426)$$

where the *reduced deflection angle* is defined as

$$\alpha_{\text{red}} \equiv \frac{D_{ds}}{D_s} \alpha(D_d \theta) = \frac{D_{ds}}{D_s} \alpha(\xi). \quad (427)$$

The deflection angle  $\alpha_{\text{red}}$  depends on the mass distribution of the deflector. However, if we assume that the lensing object is a point mass (at large distances all objects appear as point masses), then we can use equation (422) for which we obtain

$$\alpha_{\text{red}}(\theta) = \frac{4GM}{c^2 b} \frac{D_{ds}}{D_d D_s} \frac{\theta}{|\theta|^2} \quad (428)$$

Multiple images of a source occur if the lens equation (426) has multiple solutions  $\theta_i$  for a (true) source position  $\beta$  – in this case, the source is observed at the locations  $\theta_i$  on the sphere (Schneider 2002, pg. 66).

The lens equation for a point mass is simple enough to be solved analytically which means that for each source position  $\beta$  the respective image positions  $\theta_i$  can be determined. If we define the so-called **Einstein angle** of the lens,

$$\theta_E \equiv \sqrt{\frac{4GM}{c^2} \frac{D_{ds}}{D_d D_s}}, \quad (429)$$

then the lens equation (426) for the point mass with deflection angle (428) can be written as

$$\beta = \theta - \theta_E^2 \frac{\theta}{|\theta|^2}. \quad (430)$$

Obviously,  $\theta_E$  is a characteristic angle in this equation, so that for practical reasons we will use the scaling  $\mathbf{y} \equiv \beta/\theta_E$  and  $\mathbf{x} \equiv \theta/\theta_E$  so that the lens equation reduces to

$$\mathbf{y} = \mathbf{x} - \frac{\mathbf{x}}{|\mathbf{x}|^2} \Rightarrow \mathbf{x} = \frac{1}{2} \left( |\mathbf{y}| \pm \sqrt{4 + |\mathbf{y}|^2} \right) \frac{\mathbf{y}}{|\mathbf{y}|}. \quad (431)$$

The latter shows the solution of the lens equation (Schneider 2002, pg. 67). From this we can immediately draw a number of conclusions:

- For each source position  $\mathbf{y}$ , the lens equation for a point-mass lens has two solutions – any source is (formally, at least) imaged twice. The reason for this is the divergence of the deflection angle for  $\theta \rightarrow 0$ . This divergence does not occur in reality because of the finite geometric extent of the lens (e.g., the radius of the star), as the solutions are of course physically relevant only if  $\xi = D_d \theta_E |\mathbf{x}|$  is larger than the radius of the star. We need to point out again that we explicitly exclude the case of strong gravitational fields such as the light deflection near a BH or a NS, for which the equation for the deflection angle has to be modified.
- The two images  $\mathbf{x}_i$  are collinear with the lens and the source. In other words, the observer, lens, and source define a plane, and light rays from the source that reach the observer are located in this plane as well. One of the two images is located on the same side of the lens as the source, the second image is located on the other side.
- If  $\mathbf{y} = 0$ , so that the source is positioned exactly behind the lens, the full circle  $|\mathbf{x}| = 1$ , or  $|\theta| = \theta_E$ , is a solution of the lens equation (431) – the source is seen as a circular image. In this case, the source, lens, and observer no longer define a plane, and the problem becomes axially symmetric. Such a circular image is called an **Einstein ring**.
- The angular diameter of this ring is then  $2\theta_E$ . From the solution (431), one can easily see that the distance between the two images is about  $\Delta x \gtrsim 2$ , hence  $\Delta\theta \gtrsim 2\theta_E$ ; the Einstein angle thus specifies the characteristic image separation.

In the case of microlensing where we take the lens to be a galactic star, we find that

$$\theta_E = 0.902 \text{ mas} \left( \frac{M}{1 M_\odot} \right)^{1/2} \left( \frac{D_d}{10 \text{ kpc}} \right)^{-1/2} \left( 1 - \frac{D_d}{D_s} \right)^{1/2}. \quad (432)$$

Since the angular separation  $\Delta\theta$  of the two images is about  $2\theta_E$ , the typical image splittings are about a milliarcsecond (mas) for lens systems including galactic stars; such angular separations are as yet not observable with optical telescopes (Schneider 2002, pg. 69).

#### MAGNIFICATION

Light beams are not only deflected as a whole, but they are also subject to differential deflection. For instance, those rays of a light beam that are closer to the lens are deflected more than rays at the other side of the beam. The differential deflection is an effect of the tidal component of the deflection angle. By differential deflection, the solid angle which the image of the source subtends on the sky changes. Let  $\omega_s$  be the solid angle the source would subtend if no lens were present, and  $\omega$  the observed solid angle of the image of the source in the presence of a deflector. Since gravitational light deflection is not linked to emission or absorption of radiation, the surface brightness (or specific intensity) is preserved. The flux of a source is given as the product of surface brightness and solid angle. Since the former of the two factors is unchanged by light deflection, but the solid angle

changes, the observed flux of the source is modified. If  $S_0$  is the flux of the unlensed source and  $S$  the flux of an image of the source, then

$$\mu \equiv \frac{S}{S_0} = \frac{\omega}{\omega_s}, \quad (433)$$

describes the change in flux that is caused by a magnification (or a diminution) of the image of a source. Obviously, the magnification is a purely geometrical effect (Schneider 2002, pg. 68).

Although image splitting for galactic microlensing is small, the magnification by the lens should nevertheless be measurable. To do this, we have to realize that the absolute magnification is observable only if the unlensed flux of the source is known – which is not the case, of course (for nearly all sources). However, the magnification, and therefore also the observed flux, changes with time by the relative motion of source, lens, and ourselves. Therefore, the flux is a function of time, caused by the time-dependent magnification (Schneider 2002, pg. 69).

**QUESTION 2**

A two-element interferometer consists of two telescopes whose light is combined and interfered. Sketch the response of such an interferometer to a nearby red giant star, as a function of the (projected) separation between the two telescopes. The red giant subtends one-fiftieth of an arc second on the sky, and the telescope operates at a wavelength of 2 microns.

**QUESTION 2**

A two-element interferometer consists of two telescopes whose light is combined and interfered. Sketch the response of such an interferometer to a nearby red giant star, as a function of the (projected) separation between the two telescopes. The red giant subtends one-fiftieth of an arc second on the sky, and the telescope operates at a wavelength of 2 microns.

**QUESTION 3**

**Define and describe the ‘diffraction limit’ of a telescope. List at least three scientifically important telescopes which operate at the diffraction limit, and at least three which do not. For the ones which do not, explain why they do not. In both categories include at least one telescope not operating in optical/near IR wavelengths.**

### QUESTION 3

Define and describe the ‘diffraction limit’ of a telescope. List at least three scientifically important telescopes which operate at the diffraction limit, and at least three which do not. For the ones which do not, explain why they do not. In both categories include at least one telescope not operating in optical/near IR wavelengths.

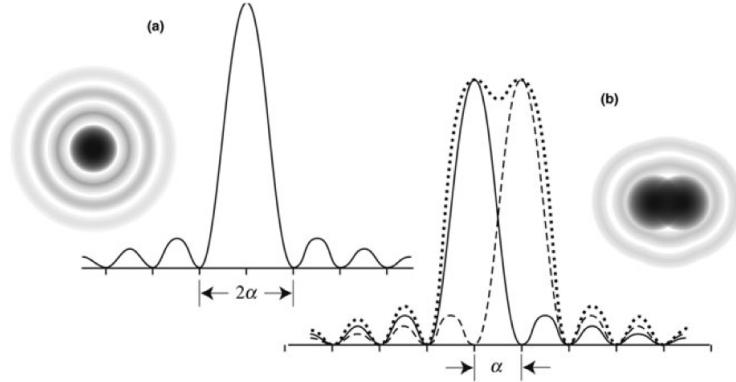


FIG. 104.— (left) (a) The Airy pattern caused by the diffraction of light through a circular aperture is shown as a negative image. The irradiance is also plotted as intensity versus angular radius. (b) The negative image of two identical monochromatic point sources separated by an angle  $\alpha$  in order to show the limiting resolution of the Rayleigh criterion. The plot shows intensity along the line joining the two images. The unblended images are shown as the solid and dashed curves while their sum is shown as the dotted curve. Image taken from (Frederick 2010). (right) The superimposed diffraction patterns from two point sources. (a) The sources are easily resolved. (b) The sources are barely resolvable. Image taken from Carroll & Ostlie (2007).

The diffraction limit of a telescope is the fundamental limitation in the resolution of the telescope based on the blurring of images due to the diffraction of light. In what follows we will focus on the diffraction of light through a circular aperture in the Fraunhofer or far-field limit. This is the case where the source of light and the observation screen are effectively far enough from the diffraction aperture so that wavefronts arriving at the aperture and observation screen may be considered plane. In contrast, the Fresnel or near-field diffraction limit takes the curvature of the wavefront into account. In the far-field approximation, as the viewing screen is moved relative to the aperture, the size of the diffraction pattern scales uniformly, but the shape of the diffraction pattern does not change. In the near-field approximation, the situation is more complex in that both the shape and size of the diffraction pattern depend on the distance between the aperture and the screen. The detailed calculations of diffraction through a circular aperture are contained in Pedrotti et al. (2007, pg. 268).

It turns out that the irradiance for a diffracted waveform through a circular aperture of diameter  $D$  is given as

$$I(\theta) = I_0 \left( \frac{2J_1(\gamma)}{\gamma} \right)^2, \quad (434)$$

where  $\gamma \equiv kD\sin(\theta)/2$ ,  $k = 2\pi/\lambda$  is the wavenumber of the diffracted light,  $\theta$  denotes the angle between the aperture opening and observing screen, and  $I_0$  is the irradiance as  $\gamma \rightarrow 0$  (i.e., where  $\theta = 0$ ).  $J_1(x)$  is the first Bessel function and the function  $J_1(x)/x$  reaches a maximum of 1/2 as  $x \rightarrow 0$ , so that the irradiance is greatest at the centre of the pattern. Due to the symmetry of the aperture, the diffraction pattern is symmetrical about the optical axis through the centre of the circular aperture and has its first zero when  $J_1(x)$  hits zero at  $x \approx 3.832$ . The first minima in the diffraction pattern therefore occurs at an angular width of

$$\gamma = \left( \frac{k}{2} \right) D \sin(\alpha) = 3.832 \Rightarrow \alpha = \sin^{-1} \left( \frac{1.22\lambda}{D} \right) \Rightarrow \alpha \approx \frac{1.22\lambda}{D}, \quad (435)$$

where the last approximation is valid in the far-field approximation (Pedrotti et al. 2007, pg. 277).

For a telescope the parameter  $D$  used above describes the diameter of the entrance aperture; the light-gathering element of the telescope, usually the mirror or lens that the incoming wave first encounters (Frederick 2010, pg. 138). Despite the fact that the source is a point, its image will have a finite size due to the diffraction process described above. The observed diffraction pattern given by equation (434) describes a central bright disk surrounded by a series of bright concentric rings, whose brightness decreases with distance from the centre of the pattern; see Figure 104. The majority of the light (84%) is focused into the central bright spot whose angular radius is given by  $\alpha$  in equation (435).

The central bright spot of the diffraction pattern is known as the Airy disk and its angular width  $\alpha$  dictates the diffraction limit of the telescope. If two point sources lie close together, their blended Airy patterns may not be distinguishable from that of a single source. If we can say for sure that the particular pattern is due to two sources, and not a single source, then we say that the sources are resolved (Frederick 2010, pg. 138). The Rayleigh criterion for just-resolvable images – a somewhat arbitrary but useful criterion – requires that the angular separation of the centres of the Airy patterns be no less than the angular radius  $\alpha$  of the Airy disk. At this limiting resolution, the maximum intensity of the one Airy pattern coincides with the first dark ring of the

second; see Figure 104. From equation (435) we have that the diffraction limit for a telescope with a lens of diameter  $D$  is

$$\Delta\theta_{\min} = \frac{1.22\lambda}{D}. \quad (436)$$

In practice, through a careful analysis of the diffraction patterns of the sources, it is possible to resolve objects that are somewhat more closely spaced than allowed by the Rayleigh criterion (Carroll & Ostlie 2007, pg. 147).

Rayleigh's criterion is a good predictor of the performance of space telescopes. On the surface of the Earth, however, turbulence, which causes dynamic density variations in the Earth's atmosphere, limits the resolving power of all but the smallest telescopes. This loss of resolution is referred to as seeing. These distortions are caused by temperature gradients in the Earth's atmosphere. This creates compressions and rarefactions in the air and correspondingly higher and lower indices of refraction. These regions in the air retard some sections of a plane wavefront while allowing other sections to progress forward. When this distorted wavefront reaches the telescope it will create a combination of small Airy patterns that add up to make a multi-spotted image called a speckle pattern, in which each spot is the diffraction-limited image of the source as seen through a coherent clump of air. Of course, the speckle pattern changes continuously and over long exposures blurs into a single 'seeing disk' (Frederick 2010, pg. 140).

Advanced techniques called adaptive optics (AO) can correct wavefront distortions that arise from atmospheric turbulence and sharpen blurred images on long exposures. In this case a small, deformable mirror is employed so that the shape of its reflecting surface is adjusted in such a way as to exactly cancel the distortions generated from the turbulence. In order to determine the changes that need to be applied, the telescope monitors a guide star that is very near the target object<sup>66</sup>. Fluctuations in the guide star determine the adjustments that must be made to the deformable mirror (Carroll & Ostlie 2007, pg. 159). This process is somewhat easier in the near-infrared because of the longer wavelengths involved. As a result, AO have been successful in providing near-diffraction-limited images in the near IR. See Frederick (2010, Figure 6.14) for a schematic of an AO system.

Three scientifically important telescopes which operate at the diffraction limit are the Hubble Space Telescope (HST), the Wilkinson Microwave Anisotropy Probe (WMAP), and the Keck II Telescope. With a primary mirror of 2.4 m, the HST was estimated to be not quite diffraction-limited in the UV region due to extremely small imperfections in the surfaces of the mirrors. Since resolution is  $\propto \lambda$  and mirror defects become less significant as  $\lambda$  increases, HST should have been nearly diffraction-limited at the red end of the visible spectrum. This was not achieved at first due to a grinding error of the primary mirror, but has since been resolved during a repair mission in 1993 (Carroll & Ostlie 2007, pg. 151). WMAP is a Gregorian telescope with 1.4 and 1.6 m primary mirrors that provides full sky imaging at five different wavelengths with diffraction-limited resolution ranging from about 0.2 to 0.9 degrees (Patanchon et al. 2004). The full sky maps are covered in the K, Ka, Q, V, and W bands (all part of the microwave region). The 10 m Keck II telescope is equipped with the second generation Near Infrared Camera (NIRC-2) which employs a laser guide AO system (Wizinowich et al. 2006). This allows for diffraction-limited resolution in the J, H, and K passbands with  $\sim 10$  times the sensitivity of the NIR camera on Keck I (which only incorporates a natural guide star AO system). Other scientifically important telescopes operating at their diffraction limit are the Spitzer Space Telescope, the Very Large Array (VLA), and the Expanded Very Large Array (EVLA).

More info from Charles' notes: HST works in the optical ( $\sim 500$  nm), has an aperture diameter of 2.4 m, and has a resolution limit of about  $0.052''$ . Spitzer has an aperture diameter of 0.85 m, works in the NIR ( $\sim 10 \mu\text{m}$ ), and has a resolution limit of about  $2''$ . Also, the Giant Meterwave Radio Telescope (GMRT) operates at its diffraction limit. This has an effective mirror diameter (since it is an array) of 25 km, operates on the GHz frequency, and has a  $\sim 2''$  resolution. In general, space telescopes and radio telescopes operate at the diffraction limit.

Three scientifically important telescopes that do not operate at the diffraction limit are the Chandra X-ray Observatory, W.M. Keck Observatory, and XMN Newton. Chandra is composed of four concentric mirror shells whose projected areas in the plane of incidence are annuli. Their average width is  $\sim 1.5$  cm implying a diffraction limit of  $0.01''$  at 0.6 nm. However, Chandra's resolution is only  $0.5''$ . This discrepancy results from the difficulty in figuring and polishing the optics of a grazing incidence telescope for imaging such small wavelengths. Imperfections on the surfaces of the mirrors degrade the reflection efficiency and scattering cross-section of the incident X-rays, making them difficult to localize<sup>67</sup>.

More info from Charles' notes: Chandra, which has an aperture diameter of 1.2 m, looks at X-rays ( $\sim 1$  nm) and has an angular resolution limit of  $0.5''$ . This is due to imperfections in Chandra's concentric mirror shells. W.M. Keck, which has a diameter of 10 m, looks in the optical ( $\sim 500$  nm), but is seeing limited to  $\sim 1''$ . Adaptive optics decreases this limit almost to the diffraction limit. XMN Newton, with an effective diameter of 0.74 m, works in the X-ray, and has a resolution limit of about  $6''$ . In general, high-energy telescopes and telescopes on the ground do not operate at the diffraction limit.

#### MIRROR IMPERFECTIONS

A number of optical imperfections can distort images as a result of lens design or imperfections in the material it is made from (from Charles and Wikipedia):

- **Spherical Aberration:** A real lens with spherical surfaces suffers from spherical aberration: it focuses rays more tightly if they enter it far from the optic axis than if they enter closer to the axis. It therefore does not produce a perfect focal point.
- **Chromatic Aberration:** In this case there is a failure of a lens to focus all colours to the same convergence point. It occurs because lenses have a different refractive index for different wavelengths of light. Generally, the refractive index decreases with increasing wavelength.

<sup>66</sup> If a guide star does not exist close enough to the target, an artificial laser guide star may be used. In this case a laser is used to excite sodium atoms in the atmosphere at an altitude of  $\sim 90$  km.

<sup>67</sup> see <http://spie.org/x8903.xml?ArticleID=x8903>

- **Coma:** Where a paraboloidal mirror/lens generates comet-like distortions. This occurs from light entering the lens out of the axis of symmetry, causing the focus to align off axis.
- **Astigmatism:** Where cross-like distortions are generated when rays that propagate in two perpendicular planes have different foci.
- **Distortion:** A variation of magnification across the image.

**QUESTION 4**

**What's the minimum mass of a black hole you could survive a fall through the event horizon without being ripped to shreds? Why would you be ripped to shreds for smaller black holes?**

#### QUESTION 4

**What's the minimum mass of a black hole you could survive a fall through the event horizon without being ripped to shreds? Why would you be ripped to shreds for smaller black holes?**

The information for this question comes from Yevgeni and Charles' notes. We can approximate a BH using the Schwarzschild metric. The event horizon is then

$$r_S = \frac{2GM}{c^2}. \quad (437)$$

The surface gravity of a BH evaluated at the event horizon can be approximated using Newton's formula,

$$g_S = \frac{GM}{r_S^2} = \frac{c^4}{4GM}. \quad (438)$$

From this result we can immediately reach some quick conclusions about BHs. Firstly, the factor  $c^4/G$  in equation (438) suggests that the surface gravity at the event horizon is a large number indeed. Moreover, since  $g_S$  is inversely proportional to  $M$ , more massive BHs have *weaker* surface gravity.

However, when it comes to ripping objects apart through *spaghettification* it is not the absolute gravitational force that we mostly care about, but rather the difference in force that the object experiences on either end; hence, the **tidal force**. The Newtonian approximation for the tidal force per unit mass felt by an object of mass  $m$  and length  $L$  is

$$f_{\text{tide}} \approx \frac{1}{m} \frac{dg}{dr} L = \frac{1}{m} \frac{d}{dr} \left( -\frac{GMm}{r^2} \right) = \frac{2GM}{r_S^3} L = \frac{c^6}{4G^2 M^2} L. \quad (439)$$

The approximation in equation (439) is valid as long as  $dg/dr \gg d^2g/dr^2L$  **why?**. From equation (439) we see that the tidal force decreases rapidly with increasing mass of the BH. This has to do with the fact that the Schwarzschild radius, in equation (437), is proportional to  $M$  (i.e., more massive BHs are larger) so that both  $g_S$  and  $f_{\text{tide}}$  decrease with increasing mass when evaluated at  $r_S$ .

Traditional forms of execution include dismembering an individual through the tidal force of horses pulling in different directions. A very ballpark estimate suggests 100g would tear a human limb from limb; horses pulling cannot be more than  $10^4$  N, which translates, for a 100 kg person, into 100 g forces. Note that roller coasters often subject humans to a few g forces, so to an order of magnitude, 10g would be too small; a better estimate is 100g. If we suppose that a moderately tall human with  $L = 2$  m falls feet-first into a BH, then inverting equation (439) yields  $M_{\min} \approx 10^{34}$  kg =  $5 \times 10^3 M_\odot$ . Hence, rather massive BHs are required in order for humans to safely pass through the event horizon.

#### SUPERMASSIVE BLACK HOLES

Part of the evidence for the existence of a SMBH at the centre of our galaxy, Sgr A\*, come from the observed orbital paths of stars within its vicinity. Many of these are observed to have rather small semi-major axes, with correspondingly large orbital velocities, suggesting the presence of a compact massive source near the centre of the MW. The size of the compact object is limited by the small periastron distances some of these stars are found to have. The fact that stars remain intact as they come so close to the massive source is consistent with the above argument if a SMBH with  $M \sim 10^6 M_\odot$  were present at the centre. In contrast, observations of solar-mass BHs usually involve accretion disks that form as companion stars are ripped apart (Yevgeni).

A natural question arises in the context of how SMBHs formed. We have already discussed some aspects of this in a previous question. Generally, this is still an open question in astrophysics. One hypothesis is that BHs with masses  $10 - 100 M_\odot$  that are left behind by the explosions of massive stars grow by accretion of matter. BHs could also have formed out of pre-galactic halos through various instabilities, culminating in the formation of a  $\sim 10 M_\odot$  BH, which then accretes at super-Eddington rates until it reaches  $\sim 10^5 M_\odot$ . BHs could also have been formed from the Big Bang itself. The primary unsolved issue today is how to feed these BHs until they reach SMBH status. The existence of high-redshift quasars requires that SMBHs be formed soon after the first stars. No method is known to gorge primordial BHs to stimulate such rapid growth (Charles).

**QUESTION 5**

**Let's say the LHC produces microscopic black holes in their high energy proton-anti-proton collisions? What will happen to them? Will they destroy the Earth?**

## QUESTION 5

**Let's say the LHC produces microscopic black holes in their high energy proton-anti-proton collisions? What will happen to them? Will they destroy the Earth?**

The information for this question comes from Charles' notes. To begin, we note that from a basic energy consideration, it is unlikely for BHs to be produced by the LHC. The LHC produces an energy density of 1 TeV in a spatial extent on the order of  $\Delta x \sim 10^{-19}$  m. A BH with that Schwarzschild radius would have a mass  $M \approx 10^8$  kg =  $10^{35}$  GeV; some 32 orders of magnitude larger than what can be achieved by the LHC.

However, for the moment, let's suppose that a BH did form during operation at the LHC. According to BH thermodynamics, all BHs decay through the emission of Hawking radiation. In particular, a Schwarzschild BH radiates photons as a blackbody with an effective temperature of

$$T = \frac{\hbar c^3}{8\pi kGM}. \quad (440)$$

We state this relation without proof, but present a useful motivation for its form below. The total radiation from the BH is thus  $L = 4\pi r_S^2 \sigma_{\text{SB}} T^4$ , and since the only energy source that can be provided by the BH comes from its rest mass, we require  $L = -c^2 dM/dt$ . We can then relate the two expressions, integrating from the time  $t = 0$  when the BH had mass  $M = M_0$  to the time  $t = t_{\text{evap}}$  when the BH had evaporated to a mass of zero:

$$-c^2 \frac{dM}{dt} = 4\pi r_S^2 \sigma_{\text{SB}} T^4 \Rightarrow \int_{M_0}^0 M^2 dM = \kappa \int_0^{t_{\text{evap}}} dt, \quad (441)$$

where  $\kappa$  is a collection of constants containing things like  $G$ ,  $c$ ,  $\hbar$ ,  $k$ , and  $\sigma_{\text{SB}}$ . Integration equation (441), solving for  $t_{\text{evap}}$ , and plugging in fundamental constants, we obtain:

$$t_{\text{evap}} \approx 10^{-16} \text{ seconds} \left( \frac{M}{1 \text{ kg}} \right)^3. \quad (442)$$

Hence, more massive BHs radiate at *lower* effective temperatures and take *longer* times to evaporate. For a BH with a mass of  $M = 1$  TeV that could potentially be generated in the LHC, equation (442) yields  $t_{\text{evap}} \approx 10^{-89}$  seconds; a very short lifetime indeed!

Perhaps we can salvage more time for our microscopic BHs by proposing that it consume surrounding mass during its evaporation. The accretion rate is then  $dM/dt = f\pi r_S^2 \rho v$ , where  $\rho$  is the matter density of surrounding material,  $v$  is the speed of the BH in passing through this material, and  $f$  is some constant of order unity; this relation is simply derived by considering a geometrical cross section of  $\sigma = \pi r_S^2$ . In order to stave off death, we require that

$$\frac{L}{c^2} = \frac{2\sigma_{\text{SB}} \hbar^4 c^{10}}{(8\pi)^3 k^4 G^2 M^2} = f\pi r_S^2 \rho v. \quad (443)$$

If we assume that  $v \approx c$  then the minimum mass density that it must swallow is  $\rho \sim 10^{155}$  kg m<sup>-3</sup>; much greater than BH densities! In order to prevent evaporation while travelling at the speed of light through pure iron, the BH would need a rest energy of  $M_0 \sim 10^{42}$  GeV.

Even if our theoretical reasoning is incorrect, so that BHs can be produced by the LHC and survive indefinitely, it is still unlikely that they will destroy the Earth. The LHC collision of two 7 TeV proton streams is equivalent to a stationary atom in the Earth's atmosphere being struck by a  $10^8$  GeV cosmic ray. Since cosmic rays have been detected with energies up to  $10^{11}$  GeV, these collisions have surely taken place in our atmosphere. The continued existence of our planet suggests that equivalent mechanisms occurring within the LMC should not be regarded as potential armageddon.

### HAWKING RADIATION

In considering BHs, quantum effects in GR become important when the mass of a BH approaches the Planck mass,  $M_P \approx 10^{-8}$  kg. In this case, we should not trust any classical inferences about the gravitational field of microscopic BHs at radii as small as its Schwarzschild radius, and so have no reason to believe these should behave gravitationally as classical BHs (Cliff Burgess' Lecture Notes from McMaster University).

One quantum property of a BH that may arise is known as **Hawking radiation**. This effect for BHs is a special case of a more general quantum phenomenon: the spontaneous production of particles by an external field. Because of the randomness of quantum mechanics, the vacuum of empty space is better imagined as a frothing soup of particles and antiparticles that are forever trying to emerge as real particles. They normally cannot emerge, however, because their appearance is forbidden by conservation laws. In particular, charge conservation requires that only particle-antiparticle pairs can emerge from vacuum together, but it is energy conservation that keeps such pairs from emerging all the time from the vacuum around us, because such an emergence would require the production of sufficient energy to account for their masses,  $E = 2mc^2$ . Although there is a sense that the uncertainty principle allows quantum fluctuations to violate energy conservation, they can only do so very briefly, and in the long term energy conservation is inviolate. The situation changes in the presence of an electric field, for example, because the energy of a pair of oppositely charged particles is a function of their distance. Such particles can lower their energy by separating because

their opposite charges make them feel forces in opposite directions to the electric field. It is the work done by these forces that lowers their energy, and if their total energy (including their mass) can be lowered to zero in this way then energy conservation can no longer forbid their spontaneous production (Burgess).

Hawking's observation in the 1970s was that a similar phenomenon can happen in the gravitational field produced by a BH. As particles and antiparticles pop in and out of the fermenting froth of the vacuum near the Schwarzschild radius, one member of a pair can fall into the BH and so be unable to recombine with its erstwhile partner. And the energy that is released by having this member fall into the hole can be sufficient to carry its surviving partner far enough away from the BH that it can escape. The resulting prediction is that a BH should emit a constant stream of elementary particles, known as Hawking radiation. When considering GR, particle production can occur provided the particle-antiparticle pair can tunnel to a separation of order  $r \simeq r_S$ , since one particle must remain outside the event horizon (in order to escape) while the other must get deep enough inside to ensure that it reaches an area for which it loses sufficient energy to allow the other to escape. Using the quantum amplitude,  $\psi \simeq \exp(-mr)$ , for the amplitude for a pair of mass  $m$  to separate a distance  $r$  leads one to expect a particle production rate that is suppressed by a power of  $\exp(-mr_S)$ . It happens that a more precise calculation does give this result, and the distribution of particles that are released in this way closely resembles what would be expected for radiation from a hot body,  $\propto \exp(-m/T_H)$ , where  $T_H$  is given by equation (440) (Burgess). Since the BH's gravitational energy was used to produce the two particles, some of its mass must be reduced (Carroll & Ostlie 2007, pg. 645).

Let's see if we can derive this temperature. Suppose we have a BH made of photons, and each photon has a wavelength  $\lambda = r_S$ . The number of photons within the BH is then

$$N = \frac{Mc^2\lambda}{hc} = \frac{2GM^2}{hc}. \quad (444)$$

The entropy of the system is  $S = k \ln \Omega \approx k \ln(N!)$ , and if  $N$  is large then  $\ln(N!) \approx N \ln(N) - N$ . For a very rough order of magnitude, say that  $\ln N - 1 \sim 1$  (we will perhaps be an order of magnitude off). Then  $S \sim E/T \sim kN$ , and  $E = mc^2$ , which gives us

$$T \sim \frac{\hbar c^3}{8\pi kGM}. \quad (445)$$

#### ESOTERIC CONSIDERATIONS

In general, for a universe equipped with  $d$  spatial dimensions, BHs radiate as blackbodies with an effective temperature  $T \propto M^{-1/(-2+d)}$ . Hence, the temperature dependence on mass becomes shallower with increasing spatial dimensions, meaning it takes longer for BHs to evaporate. Current experiments with modified gravity at very small scales have shown  $d \leq 5$ , and even with  $d = 5$ , our analysis above does not change substantially. The minimum rest energy a BH must have to stave off thermodynamic death by accreting iron while travelling at the speed of light is  $\sim 10^{24}$  GeV.

**QUESTION 6**

**How is synchrotron radiation generated? In outline, how can it be removed from CMB maps?**

## QUESTION 6

**How is synchrotron radiation generated? In outline, how can it be removed from CMB maps?**

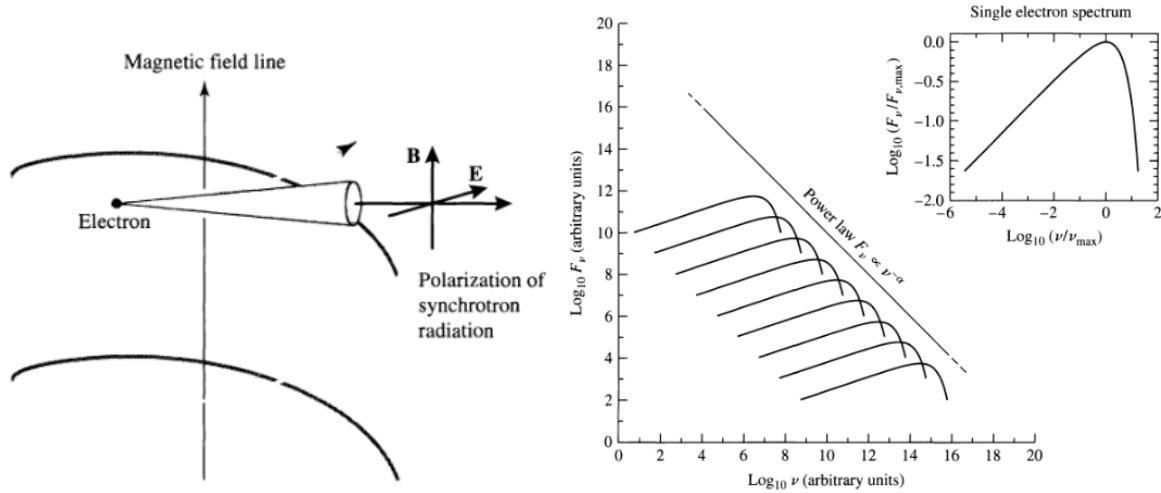


FIG. 105.— (left) Synchrotron radiation emitted by a relativistic electron as it spirals around a magnetic field line. (right) The power-law spectrum of synchrotron radiation, shown as the sum of the radiation produced by individual electrons as they spiral around magnetic field lines. The spectrum of a single electron is at the upper right. The turnover at low frequencies in the combined spectrum is not shown here. Images taken from Carroll & Ostlie (2007).

Unless otherwise cited, the information in this question was taken from Etsuko's and Charles' study notes.

We know from the form of the Lorentz force that moving electrons will spiral along magnetic field lines. In particular, the magnetic force on a moving charge  $q$  is

$$\mathbf{F} = q(\mathbf{v} \times \mathbf{B}), \quad (446)$$

so that the component of an electron's velocity  $\mathbf{v}$  perpendicular to the field lines produces a circular motion around the lines, while the component of the velocity along the lines is not affected. As they follow the curved field lines, the electrons accelerate, thereby emitting electromagnetic radiation. The emitted radiation is strongly linearly polarized in the plane of the circular motion (Carroll & Ostlie 2007, pg. 592); see Figure 105. This is because emission is always orthogonal to the direction of the field, meaning that a measure of synchrotron polarization allows us to measure the direction of the magnetic field projected onto the sky.

For nonrelativistic electrons, the process emits what is called **cyclotron radiation**, which is characterized by a single frequency equal to the frequency of gyration in the magnetic field,

$$\omega_B = \frac{qB}{m}. \quad (447)$$

For relativistic electrons, the frequency spectrum is much more complicated and can extend to many times the gyration frequency; this is known as **synchrotron radiation**. In general, if we have a power-law distribution of electron energies with spectral index  $p$ , then the total emission is also a power law,

$$I_\nu \propto \nu^{-(p-1)/2} = \nu^{-\alpha} \quad (448)$$

which can be found by Fourier transforming the square of the electric field. A pure power-law spectrum is the signature of synchrotron radiation, which is frequently encountered in astronomical situations involving relativistic electrons and magnetic fields (e.g., AGN jets, SNRs, and pulsars). As shown in Figure 105, a synchrotron spectrum is produced by the combined radiation emitted by individual electrons as they spiral around magnetic field lines (Carroll & Ostlie 2007, pg. 1089). For galactic emission,  $p$  is often equal to 8, yielding a spectral shape  $I_\nu \propto \nu^{3.5}$ .

However, the synchrotron spectrum does not continue to rise without bound as the frequency decreases. At a transition frequency, the spectrum turns over and varies as  $I_\nu \propto \nu^{5/2}$  (spectral index  $\alpha = -2.5$ ). This occurs because the plasma of spiralling electrons becomes opaque to its own synchrotron radiation, an effect known as **synchrotron self-absorption** (Carroll & Ostlie 2007, pg. 1089).

### CMB MAPS

The measured temperature distribution of the microwave radiation is a superposition of the CMB and of emission from galactic (and extragalactic) sources. In the vicinity of the galactic disk, this foreground emission dominates, whereas it seems to be considerably weaker at higher galactic latitudes. However, due to its different spectral behaviour, the foreground emission can be identified and subtracted. We note that the galactic foreground basically consists of three components: synchrotron radiation from relativistic electrons in the MW, thermal radiation by dust, and bremsstrahlung from hot gas. The synchrotron component

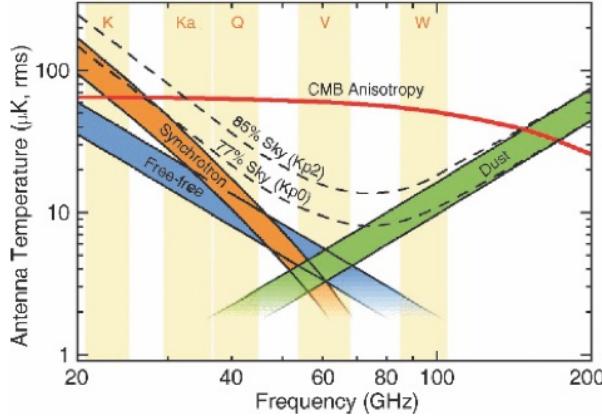


FIG. 106.— The antenna temperature ( $\propto I_\nu \nu^{-2}$ ) of the CMB and of the three foreground components discussed in the text, as a function of frequency. The five frequency bands of WMAP are marked. The dashed curves specify the average antenna temperature of the foreground radiation in the 77% and 85% of the sky, respectively, in which the CMB analysis was conducted. We see that the three high-frequency channels are not dominated by foreground emission. Image taken from Schneider (2002).

defines a spectrum of about  $I_\nu \propto \nu^{-0.8}$ , whereas the dust is much warmer than 3 K and thus shows a spectral distribution of about  $I_\nu \propto \nu^{3.5}$  in the spectral range of interest for CMB measurements. Bremsstrahlung has a flat spectrum in the relevant spectral region,  $I_\nu \approx \text{const}$ . This can be compared to the spectrum of the CMB, which has a form  $I_\nu \propto \nu^2$  in the Rayleigh-Jeans region<sup>68</sup> (Schneider 2002, pg. 342).

There are two ways to extract the foreground emission from the measured intensity distribution. First, by observing at several frequencies the spectrum of the microwave radiation can be examined at any position, and the three aforementioned foreground components can be identified by their spectral signature and subtracted. As a second option, external datasets may be taken into account. At larger wavelengths, the synchrotron radiation is significantly more intense and dominates. From a sky map at radio frequencies, the distribution of synchrotron radiation can be obtained and its intensity at the frequencies used in the CMB measurements can be extrapolated. In a similar way, the IR emission from dust, as measured, e.g., by the IRAS satellite, can be used to estimate the dust emission of the MW in the microwave domain. Finally, one expects that gas that is emitting bremsstrahlung also shows strong Balmer emission of hydrogen, so that the bremsstrahlung pattern can be predicted from an H $\alpha$  map of the sky. Both options, the determination of the foregrounds from multifrequency data in the CMB experiment and the inclusion of external data, are utilized in order to obtain a map of the CMB which is as free from foreground emission as possible (Schneider 2002, pg. 342).

The optimal frequency for measuring the CMB anisotropies is where the foreground emission has a minimum; this is the case at about 70GHz (see Figure 106). Unfortunately, this frequency lies in a spectral region that is difficult to access from the ground. Another issue associated with observing the CMB from the ground is IR emission from the atmosphere (Schneider 2002, pg. 342).

#### AGN SPECTRA

In the SED of a quasar, the turnover evident at low frequency (redward of the big blue bump and IR bump) may be due to synchrotron self-absorption. However, the thermal contributions to the continuum spectrum, evident in the IR bump suggest that in other cases, the turnover may be due to the long-wavelength Rayleigh-Jeans portion of the blackbody spectrum produced by warm dust grains. It is possible that the steeper, low-frequency spectra of radio-quiet AGNs are due to the thermal spectrum of dust grains, while the shallower low-frequency spectra of radio-loud AGNs may be due to a combination of thermal and nonthermal<sup>69</sup> emission (Carroll & Ostlie 2007, pg. 1090).

The observation of AGN jets are made possible by inefficiencies in the transport of particles and energy out to the radio lobes. The spectra of the radio lobes and jets follow a power law, with a typical spectral index of  $\alpha \sim 0.65$ . The presence of power-law spectra and a high degree of linear polarization strongly suggest that the energy emitted by the lobes and jets comes from synchrotron radiation. The loss of energy by synchrotron emission is unavoidable, and in fact the relativistic electrons in jets will radiate away their energy after just  $10^4$  yr or so. This implies that there is not nearly enough time for particles to travel out to the larger radio lobes; even at the speed of light typical travel times are on the order of several Myr. This long travel time and the long lifetime of radio lobes imply that there must be some mechanism for accelerating particles in the jets and radio lobes. As one possibility, shock waves may accelerate charged particles by magnetically squeezing them, reflecting them back and forth inside the shock. Radiation pressure may also play a role, but it alone is not enough to generate the necessary acceleration (Carroll & Ostlie 2007, pg. 1125).

<sup>68</sup> Two limiting cases of the Planck function are of particular interest. For low frequencies,  $h\nu \ll kT$ , one can apply the expansion of the exponential function for small arguments. The leading-order term in this expansion then yields  $B_\nu(T) \approx 2\nu^2 kT/c^2$ , which is called the **Rayleigh-Jeans approximation** of the Planck function. In the other limiting case of very high frequencies,  $h\nu \gg kT$ , the exponential factor in the denominator becomes very much larger than unity, so that we obtain  $B_\nu(T) \approx 2h\nu^3 \exp(-h\nu/kT)/c^2$ , known as the **Wein approximation** of the Planck spectrum (Schneider 2002, pg. 419).

<sup>69</sup> Astrophysical synchrotron sources are often called nonthermal sources because the energy distribution of the relativistic electrons is a power law and there is no single electron temperature  $T$ . For more information see <http://www.cv.nrao.edu/course/astr534/SynchrotronSrcs.html>.

**QUESTION 7**

**What are “forbidden lines” of atomic spectra? In what conditions are they observationally important?**

## QUESTION 7

**What are “forbidden lines” of atomic spectra? In what conditions are they observationally important?**

Observations of atomic spectra and the theory of quantum mechanics both demonstrate that the energy states available to any bound electron are quantized. That is, an electron can only exist in certain permitted energy and angular momentum states. The reason for this quantization is due to the wave structure of the electron. In the situation in which the electron is bound in the potential well created by the positive charge of an atomic nucleus, the electron’s wave function undergoes constructive interference at particular energies, and destructive interference at all others. Since the square of the wave function gives the probability density of the electron existing at a certain location and time, the electron cannot have energies that cause the wave function to destructively interfere with itself and go to zero (Frederick 2010, pg. 197).

This quantization leads to the familiar description of the wave function of an electron bound to an atomic nucleus. In this case we describe the wave function using a set of four quantum numbers:  $n$ ,  $l$ ,  $m$ , and  $s$ .  $n$  is the so-called principal quantum number and takes integer values  $n = 1, 2, 3, \dots, \infty$ .  $l$  is known as the angular momentum quantum number and takes on the integer values  $l = 0, 1, \dots, n-1$  (the states  $l = 0, 1, 2, 3, \dots$  are also referred to as the states  $s, p, d, f, g, \dots$ ).  $m$  is the magnetic quantum number and takes on the integer values  $m = -l, \dots, l$ . Finally,  $s$  is the spin quantum number which, for an electron, takes on  $s = \pm 1/2$  (Marleau 2010). Though the wave function itself depends on the first three quantum numbers, the energy of the electron depends only on the value of  $n$  (in particular, the energy levels are  $\propto n^2$ ).

In order to understand the physical processes behind the forbidden lines of atomic spectra we must delve into the realm of quantum mechanics. Unless specified otherwise, what follows will be a brief summary of the information presented in Griffiths (2005, Chapter 9).

The physics of atomic line transitions is one of the few exactly solvable problems involving a time-dependent Hamiltonian. In particular, this problem is solved using time-dependent perturbation theory in which the Hamiltonian is separated into a time-independent term ( $H^0$ ) and a time-dependent term ( $H'$ ). For simplicity we assume that there are only two states of the time-independent term,  $\psi_a$  and  $\psi_b$ , which have energies  $E_a$  and  $E_b$ , and are mutually orthonormal. Since the states  $\psi_a$  and  $\psi_b$  constitute a complete set, any state  $\Psi$  can be written as a linear combination of them. This is true even when we switch on the time-dependent term  $H'$ . We therefore write any general state as

$$\Psi(t) = c_a(t)\psi_a e^{-iE_a t/\hbar} + c_b(t)\psi_b e^{-iE_b t/\hbar}, \quad (449)$$

where  $c_a$  and  $c_b$  are time-dependent coefficients.

Suppose that  $\psi_b$  is some excited state and that  $\psi_a$  is the ground state. The problem of examining the transition probability of the electron from one state to the other reduces to determining the coefficients  $c_a$  and  $c_b$  as functions of time. It turns out that the time derivatives of these coefficients can be determined by evaluating

$$\dot{c}_a = -\frac{i}{\hbar} H'_{ab} e^{-i\omega_0 t} c_b, \quad \dot{c}_b = -\frac{i}{\hbar} H'_{ba} e^{i\omega_0 t} c_a \quad (450)$$

where  $H'_{ij} \equiv \langle \psi_i | H' | \psi_j \rangle$  and  $\omega_0 \equiv (E_b - E_a)/\hbar$  (Griffiths 2005, pg. 343).

Let’s now specify our problem as to that of an atom exposed to an electromagnetic wave. In the presence of passing light, an atom responds primarily to the electric component of the radiation. If the wavelength is large compared to the size of the atom then we can ignore the spatial variation in the field<sup>70</sup>. We thus have that the atom is exposed to a sinusoidally oscillating electric field

$$\vec{E} = E_0 \cos(\omega t) \hat{k}, \quad (451)$$

assuming that the light is monochromatic and polarized along the  $z$  axis. In this case we have that

$$H' = -qE_0 z \cos(\omega t) \Rightarrow H'_{ab} = -dE_0 \cos(\omega t), \quad (452)$$

where  $q$  is the charge on the electron and  $d \equiv q\langle \psi_b | z | \psi_a \rangle$  and is supposed to remind us of the electric dipole moment (technically it is the off-diagonal matrix element of the  $z$  component of the dipole momentum operator  $q\vec{r}$ ). Because of its association with electric dipole moments, radiation governed by equation (452) is called **electric dipole radiation**; it is overwhelmingly the dominant kind of radiation, at least in the visible region (Griffiths 2005, pg. 349).

Griffiths (2005, pg. 346) solves equation (450) for the time-dependent Hamiltonian given by equation (452) using time-dependent perturbation theory. From their result the transition probability – the probability that an electron which started in the ground state  $\psi_a$  will be found, at time  $t$ , in the state  $\psi_b$  – is obtained by taking  $|c_b(t)|^2$ . This turns out to be

$$P_{a \rightarrow b}(t) = \left( \frac{|d|E_0}{\hbar} \right)^2 \frac{\sin^2[(\omega_0 - \omega)t/2]}{(\omega_0 - \omega)^2}. \quad (453)$$

This is the process known as absorption in which the atom absorbs energy  $E_b - E_a = \hbar\omega_0$  from the electromagnetic field. By looking at the denominator of equation (453) we see that the probability of transition is greatest when the frequency  $\omega$  of the photon is close to the natural frequency of the transition  $\omega_0$ . The same procedure can be used to compute the transition probability  $P_{b \rightarrow a}$  of an electron falling from an excited state to the ground state. This describes the processes of stimulated emission and it turns out that the rates are precisely equal:  $P_{a \rightarrow b} = P_{b \rightarrow a}$ . In the case of stimulated emission we shine light on the atom, causing

<sup>70</sup> For visible light  $\lambda \sim 500$  nm while the diameter of an atom is  $\sim 10$  nm, so this approximation is justified; but this will not be the case for, say, X-rays.

a transition downwards so that an energy  $\hbar\omega_0$  is gained by the field. In general, we modify the form of equation (453) into an integral in order to deal with non-monochromatic light and we modify the form of  $d$  to assume a flux of radiation from all directions (Griffiths 2005, pg. 353).

There is another mechanism for emission known as spontaneous emission where the electron makes a transition downwards from an excited state without any applied electromagnetic field to initiate the process. However, spontaneous emission is not spontaneous at all, for if the electron were really free from external perturbations it would not make a transition downwards. The perturbation responsible for this transition is actually from the non-zero electromagnetic field that exists within the atom even in the ground state (just as how the harmonic oscillator has nonzero energy in the ground state). So in a sense spontaneous emission is really just another form of stimulated emission.

Suppose that we are interested in systems like hydrogen in which we specify the states with the quantum numbers  $n$ ,  $l$ , and  $m$ . From the presence of  $d$  in equation (453), the problem of calculating spontaneous emission rates ( $R = dP/dt$ ) is reduced to evaluating matrix elements of the form  $\langle \psi_a | \vec{r} | \psi_b \rangle$ . Through the use of angular momentum operators it can be shown that there are a set of selection rules that restrict certain transitions from taking place. The first of these selection rules requires that

$$\Delta m = \pm 1 \text{ or } 0. \quad (454)$$

This is really a statement of the conservation of (the  $z$  component of) angular momentum. Since the photon carries a spin of 1, its value of  $m$  can be either -1, 0, or 1. The emitted photon will therefore take on one of these values and our conservation law requires that the atom give up whatever the photon takes away (Griffiths 2005, pg. 361). The next selection rule requires that

$$\Delta l = \pm 1. \quad (455)$$

Again, this is simply another statement of conservation of angular momentum. Since the photon has spin a spin of 1, the rules for the addition of angular momentum state that the value  $l'$  of the state after emission can only take on the values  $l' = l+1$ ,  $l' = l$ , or  $l' = l-1$ , where  $l$  is the value of angular momentum before transmission. Although the middle value is allowed is permitted through conservation of angular momentum, it does not occur for dipole radiation (Griffiths 2005, pg. 362).

Similar selection rules also exist for many electron atoms that involve the total spin quantities  $L$ ,  $S$ , and  $J$ . These are:

- Parity, defined as  $\sum_i l_i$ , must change.
- The change in total angular momentum must be  $\Delta L = 0, \pm 1$ .
- The change in total angular plus spin momentum must be  $\Delta J = 0, \pm 1$ , but  $J = 0 \rightarrow 0$  is forbidden.
- Only one single electron wavefunction  $nl$  changes, with  $\Delta l = \pm 1$ .
- Total spin cannot change:  $\Delta S = 0$ .

Allowed transitions are those that follow all of these **selection rules**. Semi-forbidden transitions are those that follow the first four rules, but break the last one concerning spin. Forbidden transitions are those that break any of the first four rules (Charles).

The selection rules state that not all transitions to lower-energy states can proceed by spontaneous emission. The transitions which violate the above selection rules are called (misleadingly) **forbidden transitions**. In general, however, there is always some probability for radiative transmission between two states, though in some cases this probability can be exceedingly small. The “forbidden” transitions violating the selection rules above are only disallowed in electric dipole radiation. They arise because equation (451) is only an approximation of the electric field in the vicinity of an atom; its true form is

$$\vec{E} = E_0 \cos(\vec{k} \cdot \vec{r} - \omega t) \vec{k}, \quad (456)$$

If the atom is centred at the origin then  $\vec{k} \cdot \vec{r} \sim r/\lambda$  since  $|\vec{k}| = 2\pi/\lambda$ . The dipole approximation takes  $r/\lambda \ll 1$ , which allowed us to drop that term from equation (456). However, if we take the first-order correction then we have

$$\vec{E} = E_0 [\cos(\omega t) + (\vec{k} \cdot \vec{r}) \sin(\omega t)]. \quad (457)$$

The new term in equation (457) is called the electric quadrupole term or magnetic dipole term (Griffiths 2005, pg. 366). Their presence now allow for forbidden transitions to take place (and even higher order powers of  $\vec{k} \cdot \vec{r}$  lead to even more “forbidden” transitions, associated with higher multipole moments).

Another equivalent condition for the applicability of the dipole approximation is that  $v/c \ll 1$ . This arises because  $\vec{k} \cdot \vec{r}$  can be written as an expansion in  $v/c$  (Rybicki & P. 2004, pg. 272). The reason why both the magnetic dipole and electric quadrupole term arise as the first-order correction in equation (457) is that their strengths are of the same order of magnitude since the magnetic force is weaker than the electric force by a factor of  $v/c$ . An example of radiation arising from the magnetic dipole term is the 1909 Å emission line of C III] while an example of radiation arising from the electric quadrupole term is the 5007 Å emission line of [O III] (Marleau 2010). Another type of forbidden line is the 21 cm line that results when the electron flips its spin in a hydrogen atom. This is a “forbidden” line that only becomes allowed when we consider the magnetic dipole moment introduced by the proton in the nucleus (Griffiths 2005, pg. 283).

This analysis has led to the classification scheme of allowed transitions, semi-forbidden transitions, and forbidden transitions. The allowed transitions are those that satisfy the selection rules of electric dipole radiation. They have the largest transition probabilities and the lifetime of excited states that decay through this process is of order  $10^{-8}$  s. Forbidden transitions are

characterized as having decay time-scales of the order of 1 s, whereas semi-forbidden transitions have time-scales intermediate those of allowed and forbidden transitions (Schneider 2002, pg. 197). The distinction between semi-forbidden and forbidden transitions is accomplished by denoting the former with a single square bracket (as in C III]) and denoting the latter with double square brackets (as in [O III]).

Forbidden transitions are observationally important as their presence can provide estimates into the density of the gas we are observing. An excited atom can transit into a lower-energy state either through the emission processes described above, or by losing energy through collisions with other atoms. The rate at which the radiate transmit through collisional de-excitation depends on the gas density. If the density of the gas is high, the mean time between two collisions is much shorter than the average lifetime of forbidden or semi-forbidden radiational transitions. Therefore the corresponding line photons are not observed<sup>71</sup>. The absence of forbidden lines is then used to derive a lower limit for the gas density, and the occurrence of semi-forbidden lines yields an upper bound for the density (Schneider 2002, pg. 197). To minimize the dependence of this argument on the chemical composition of the gas, transitions of the same element are preferentially used for density estimates.

#### EINSTEIN'S COEFFICIENTS

Suppose that there are two states, 1 and 2, n a system at thermodynamic equilibrium. This requires the following:

$$n_1 B_{12} \bar{J} = n_2 (A_{21} + B_{21} \bar{J}), \quad (458)$$

where  $B_{ab}$  denotes stimulated transition between states  $a$  and  $b$ ,  $A_{ab}$  denotes spontaneous emission, and  $\bar{J}$  is the intensity of photons evaluated at the frequency required for transition between the two states. Since the system is in thermodynamic equilibrium, we have that

$$\frac{n_2}{n_1} = \frac{g_2}{g_1} e^{-h\nu_0/kT}, \quad (459)$$

where  $\nu_0$  is the wavelength difference between states 1 and 2. Moreover, a system in thermodynamic equilibrium radiates as a blackbody, so that  $\bar{J} \approx B_\nu(\nu_0)$ , where the approximate arises since there is some variation in  $\nu$  over a line profile, though this variation is small. Combining equations (458) and (459) yields

$$B_\nu(\nu_0) = \frac{A_{21}/B_{21}}{\frac{g_2 B_{21}}{g_1 B_{12}} \exp(h\nu_0/kT) - 1}, \quad (460)$$

and equating this to the Planck spectrum yields:

$$g_1 B_{12} = g_2 B_{21} \quad \text{and} \quad A_{21} = \frac{2h\nu^3}{c^2} B_{21}. \quad (461)$$

The explicit form of the stimulated coefficients,  $B_{12}$  and  $B_{21}$ , can be computed from time-dependent perturbation theory (Charles).

<sup>71</sup> The gas density must be very low in order to observe forbidden lines. The density is so low, in fact, that it cannot be achieved in the laboratory; these lines are indeed forbidden in the lab.

**QUESTION 8**

**What is a polytropic equation of state? Give examples of objects for which this is a very good approximation.**

## QUESTION 8

**What is a polytropic equation of state? Give examples of objects for which this is a very good approximation.**

Unless otherwise cited, information pertaining to this question was taken from Yevgeni's and Charles' study notes.

The three basic stellar properties are pressure  $P$ , density  $\rho$ , and temperature  $T$ . In general, these quantities are co-dependent, and an equation of state relating them takes the form  $P = f(\rho, T)$ . This is analytically difficult to work with, since there are essentially two free parameters and setting  $P$  cannot uniquely determine both  $\rho$  and  $T$ . In many situations, however,  $T$  is a function of  $\rho$ , and in these situations  $P = f(\rho)$ , known as a **barytropic** equation of state. A subset of these equations of state, where

$$P = K\rho^\gamma = K\rho^{1+(1/n)}, \quad (462)$$

are known as a **polytropic** equation of state. Here  $\gamma$  is called the polytropic exponent,  $K$  the polytropic constant, and  $n$  the polytropic index.  $K$  can either be fixed (as we will see below) or remains as a free-parameter that varies from star to star. A polytropic equation of state is useful since it simplifies the equations of stellar structure.

A particularly famous polytropic equation of state is the ideal gas law. In this case we have that:

$$P = \left( \frac{kT}{\mu m_H} \right) \rho, \quad (463)$$

so that  $\gamma = 1$ ,  $n = \infty$ , and  $K$  is collected in the brackets.

Of course, in many astrophysical situations, polytropes – and in particular the ideal gas law – are not realistic equations of state. However, in some situations they can be used as very good approximants. These include:

- **Radiation-dominated Stars:** For many stars we can consider the pressure to arise from a sum of ideal thermal pressure and radiation pressure so that

$$P = \frac{k}{\mu m_H} \rho T + \frac{a}{3} T^4 = \frac{k}{\mu m_H \beta} \rho T, \quad (464)$$

where  $\beta \equiv P_{\text{ideal}}/P$ , and is assumed to be constant throughout the star. Hence,  $1 - \beta = P_{\text{rad}}/P = aT^4/3P$ , and we can replace the temperature in equation (464) with a function of pressure:

$$P = K\rho^{4/3} = \left( \frac{3k^4}{a\mu^4 m_H^4} \right)^{1/3} \left( \frac{1-\beta}{\beta^4} \right)^{1/3} \rho^{4/3}, \quad (465)$$

which is an  $n = 3$  polytrope, provided that  $\beta$  is constant. Extremely massive stars are nearly completely radiation-dominated, completely convective (i.e., roughly isentropic), which for radiation means  $\rho \propto T^3$ ; combining this with  $P \propto T^4$  yields  $\gamma = 4/3$ , which is an  $n = 3$  polytrope and consistent with our prior derivation. The Sun can be approximated by an  $n = 3$  polytrope, though the match would not be perfect since the Sun has a convective envelope.

- **Zero-temperature Degenerate Stars:** We have considered these in detail in another question. Suffice it to say, a non-relativistic completely degenerate gas has  $P \propto \rho^{5/3}$  and a relativistic completely degenerate gas has  $P \propto \rho^{4/3}$ . An  $n = 3/2$  polytrope is inherently unstable, and that manifests itself here in a maximum mass for degenerate stars. The non-relativistic equation works well for cold WDs, BDs, and gas giants.

- **Adiabatic Stars:** Consider a gas with  $2\alpha$  degrees of freedom. Then  $U = \alpha N k T = \alpha P V$ . The first law of thermodynamics states that  $dU = -P dV$ . Substituting  $dU = \alpha P dV + \alpha V dP$ , we obtain  $dP/P = -(\alpha+1)/\alpha(dV/V)$ , so that

$$P \propto \rho^{(\alpha+1)/\alpha}. \quad (466)$$

For a monotonic gas,  $\alpha = 3/2$ , giving  $\gamma = 5/3$ , or  $n = 3/2$ .

- **Isothermal Stars:** In this case  $P \propto \rho$  with the polytropic constant  $K = kT/\mu m_H$ . True isothermal stars do not exist, but this can give the structure of the inert core of a shell-burning star.

- **Neutron Stars:** NSs are supported by neutron degeneracy pressure, and hence should have an equation of state similar to that of WDs, except for a few parameter changes. At the extreme densities of NSs, general relativistic effects cannot be ignored, which means that determining an effective equation of state for their interior becomes much more difficult.

### LANE-EMDEN EQUATION

The advantage of using a polytropic equation of state is that it immediately simplifies the equation of hydrostatic balance:

$$\frac{dP}{dr} = -\frac{d\Phi}{dr} \rho \Rightarrow \frac{d\Phi}{dr} = -\gamma K \rho^{\gamma-2} \frac{d\rho}{dr}. \quad (467)$$

This relation can be integrated to obtain

$$\rho = \left( \frac{-\Phi}{(n+1)K} \right)^{n/(n-1)}, \quad (468)$$

$n$	$z_n$	$\left(-z^2 \frac{dw}{dz}\right)_{z=z_n}$	$\rho_c/\bar{\rho}$
0	2.4494	4.8988	1.0000
1	3.14159	3.14159	3.28987
1.5	3.65375	2.71406	5.99071
2	4.35287	2.41105	11.40254
3	6.89685	2.01824	54.1825
4	14.97155	1.79723	622.408
4.5	31.8365	1.73780	6189.47
5	$\infty$	1.73205	$\infty$

FIG. 107.— Results from integrating the Lane-Emden equation for a given polytrope. Integration ends when  $z = z_n$ , at which point  $w(z) = 0$ . Image taken from Kippenhahn & Weigert (1990).

provided that  $\gamma \neq 1$ , and where we have set an appropriate boundary condition that  $\Phi = 0$  at  $\rho = 0$ . The Poisson equation assuming spherical symmetry reduces to

$$\nabla^2 \Phi = 4\pi G\rho \Rightarrow \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial \Phi}{\partial r} \right) = 4\pi G\rho \Rightarrow \frac{d^2 \Phi}{dr^2} + \frac{2}{r} \frac{d\Phi}{dr} = 4\pi G \left( \frac{-\Phi}{(n+1)K} \right)^n, \quad (469)$$

after making use of equation (468). We can now define a number of dimensionless variables to simplify this expression:

$$z = Ar \text{ where } A = \frac{4\pi G}{(n+1)K} \rho_c^{1-1/n} \text{ and } w = \frac{\Phi}{\Phi_c} = \left( \frac{\rho}{\rho_c} \right)^{1/n}, \quad (470)$$

with subscript  $c$  referring to central values. At the centre we have  $z = 0$ ,  $\Phi = \Phi_c$ , and  $w = 1$ . Equation (469) can then be rewritten as

$$\frac{d^2 w}{dz^2} + \frac{2}{z} \frac{dw}{dz} + w^n = 0 \Leftrightarrow \frac{1}{z^2} \frac{d}{dz} \left( z^2 \frac{dw}{dz} \right) + w^n = 0. \quad (471)$$

This is known as the **Lane-Emden equation**. Since we are only interested in physical solutions (i.e., ones that do not diverge), we require that  $dw/dz(z=0) = 0$ .

While there are only a few analytical solutions, this is a relatively easy integral to solve; to remove the regular singularity at zero we may expand  $w(z)$  into a power series. We are then able to determine the density  $\rho$  as a function of the central density as a function of radius. To scale this solution to a star, we use the relation

$$M = 4\pi \rho_c R^3 \left( -\frac{1}{z} \frac{dw}{dz} \right)_{z=z_n}. \quad (472)$$

If we do not fix  $K$ , then all we need to specify are  $n$ , and two out of three variables:  $M$ ,  $R$ , and  $\rho_c$ . If we do fix  $K$ , then all we need to specify are  $n$  and one out of  $M$ ,  $R$ , and  $\rho_c$ . Fixing  $K$  naturally gives the mass-radius relationship

$$R \propto M^{\frac{1-n}{3-n}}. \quad (473)$$

Figure 107 shows a table of Lane-Emden values for different polytropic models. These are obtained by choosing an  $n$  and numerically solving equation (471) with appropriate boundary conditions to derive  $w(z)$  and  $w'(z)$ ; then  $\rho = \rho_c w^n(z)$  is determined as a function of  $z$ . We can also utilize the values in Figure 107 to compute mixed polytropic models. For example, solar mass stars on the MS have convective envelopes, but radiative cores; these can be fitted using two different polytropic models. Note that the radius tends to infinity for polytropes with  $n \geq 5$ .

**QUESTION 9**

**What was the solar neutrino problem, and how was it resolved?**

## QUESTION 9

### What was the solar neutrino problem, and how was it resolved?

Information pertaining to this question was taken from Etsuko's and Charles' study notes, which were based on Carroll & Ostlie (2007, pg. 356).

The solar neutrino problem was first discovered by Davis et al. in 1970. Davis et al.'s neutrino detector at Homestake Gold Mine in South Dakota contained approximately a thousand metric tonnes of cleaning fluid,  $\text{C}_2\text{Cl}_4$ , in order to observe the weak reaction



${}_{18}^{37}\text{Ar}$  has a half-life of 35 days. The threshold energy for reaction, 0.814 MeV, is less than the energies of the neutrinos produced in every step of the pp chain except the crucial first one,  ${}^1\text{H} + {}^1\text{H} \rightarrow {}^2\text{H} + e^+ + \nu_e$ . However, the reaction that accounted for 77% of the neutrinos detected in the Davis experiment is the decay of  ${}^8\text{B}$  in the PP III chain:



Unfortunately, this reaction is very rare, producing only for one pp chain termination in 5000.

Careful calculations of solar neutrino emission and their reaction rate with  $\text{C}_2\text{Cl}_4$  in equation (??) suggested that 7.9 SNU<sup>72</sup> should have been recorded in the Homestake experiment. The actual data found  $2.56 \pm 0.16$  SNU; only one argon atom was produced every two days in the 100,000 gallon tank! To determine this value, the tank was carefully purged of accumulated argon atoms once every few months. The measured value was subsequently corroborated by other detectors, such as Japan's Super-Kamiokande (which used pure  $\text{H}_2\text{O}$  surrounded by detectors; neutrinos scattering off electrons produce Cherenkov radiation, which can then be detected) and the Soviet-American Gallium Experiment and GALLEX in Gran Sasso, Italy (which exploited the reaction of low-energy pp chain neutrinos that dominate the Sun's neutrino flux:  $\nu_e + {}_{31}^{71}\text{Ga} \rightarrow {}_{32}^{71}\text{Ge} + e^-$ ).

The resolution to this deficit of neutrinos, referred to as the **solar neutrino problem**, is known as the **Mikheyev-Smirnov-Wolfenstein (MSW) effect**, which involves the transformation of neutrinos from one flavour to another. This idea is an extension of the electroweak theory of particle physics that combines the electromagnetic theory with the theory of weak interactions governing some types of radioactive decay. All neutrinos produced by the Sun are electron neutrinos ( $\nu_e$ ), but during their journey from the centre of the Sun to Earth they oscillate between electron, muon ( $\nu_\mu$ ) and tau ( $\nu_\tau$ ) neutrinos. These oscillations occur during the passage of neutrinos through the interior of the Sun and are caused by interactions with electrons. Because the chlorine, water, and gallium detectors have different threshold energies and they are sensitive to only the electron neutrino, their results were determined to be consistent with MSW theory.

One testable consequence of the MSW effect is that if neutrinos oscillate between flavours, they must necessarily have mass. This is because the change of neutrino flavour can only occur between neutrinos having different masses. The required mass difference needed for the MSW solution to the solar neutrino problem is much less than the current experimentally established upper limit on the mass of the electron neutrino of  $m_{\nu_e} = 2.2$  eV. Even though the standard electroweak theory does not predict masses for the neutrinos, many reasonable extensions of this theory do allow for masses in the right range (e.g., theories such as GUTs).

In 1998 Super-Kamiokande was used to detect atmospheric neutrinos produced when high-energy cosmic rays collide with Earth's upper atmosphere, which is capable of creating electron and muon neutrinos, but not tau neutrinos. They found that the muon neutrino flux coming from underneath the detector was substantially less than the flux coming from overhead. This can be interpreted as muon neutrinos generated on the other side of the planet oscillating into tau neutrinos on their way to the detector (i.e., through the Earth's interior), while the muon neutrinos generated on the near side of the planet have not had the time to do so. In addition, the Sudbury Neutrino Observatory has been able to detect all three flavours of neutrino.

### NEUTRINO MASS

One testable consequence of the MSW effect was that flavour oscillation was only allowed by GUTs if neutrinos had mass (electroweak theory does not predict neutrinos to have mass). The current established upper limit for neutrino masses is around 0.3 eV, which is still greater than the mass needed for the MSW effect to operate. Experimentally, neutrino mass can be determined via particle accelerator experiments, which have set a lower limit for the total neutrino mass of all three flavours to be  $m_{\min} \gtrsim 0.05$  eV. Since neutrinos are considered hot dark matter, they suppress both the CMB power spectrum and the matter power spectrum at small scales by free-streaming. Observations of small-scale power constrain the abundance of HDM to  $\Omega_\nu h^2 < 0.0076$ , implying a *maximum* neutrino mass of  $m_{\max} \lesssim 0.23$  eV (Schneider 2002, pg. 352).

<sup>72</sup> This stands for **solar neutrino unit**, defined as  $10^{-36}$  reactions per target atom per second. With roughly  $10^{30}$  atoms of  ${}_{17}^{37}\text{Cl}$  atoms in the tank, if only one argon atom was produced each day, this rate would have corresponded to 5.35 SNU.

**QUESTION 10**

**Why is nuclear fusion stable inside a main-sequence star? Under what conditions is nuclear fusion unstable? Give examples of actual objects.**

### QUESTION 10

**Why is nuclear fusion stable inside a main-sequence star? Under what conditions is nuclear fusion unstable? Give examples of actual objects.**

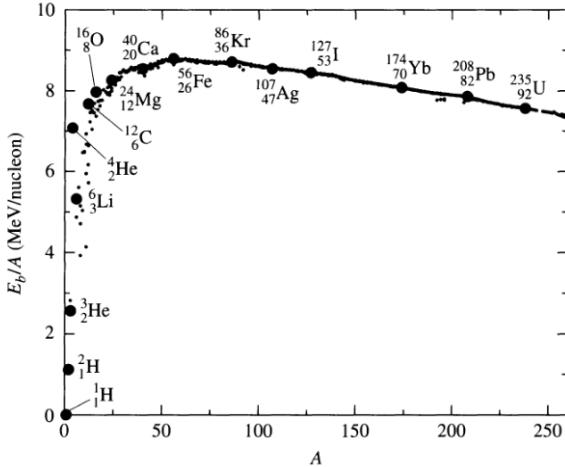


FIG. 108.— The binding energy per nucleon as a function of mass number  $A$ . Note that several nuclei have abnormally high values of  $f$  relative to others of similar mass. Among these unusually stable nuclei are  ${}^4_{\text{He}}$ ,  ${}^{16}_{\text{O}}$ , and  ${}^{12}_{\text{C}}$  which, along with  ${}^1_{\text{H}}$  are the most abundant nuclei in the universe. Image taken from Carroll & Ostlie (2007).

MS stars are powered by thermonuclear fusion reactions occurring in their cores. In nuclear fusion reactions, several light nuclei fuse together to form a heavier nuclei, whose mass is actually slightly lower than the total mass of the constituent parts that went into making it up. For example, a helium nucleus, composed of two protons and two neutrons, can be formed by a series of nuclear reactions involving four hydrogen nuclei ( $4 \text{ H} \rightarrow \text{He} + \text{low mass remnants}$ ). The total mass of the four hydrogen atoms is 4.03130013 u whereas the mass of the helium nucleus is 4.002603 u. Obviously, there is a net loss of mass in this reaction, which, through Einstein's famous equation, must have been compensated by a release of energy. For the fusion of helium, the net release of energy is 26.7 MeV, roughly 0.7% of the original masses involved. This energy is known as binding energy and it corresponds physically to the amount of energy required to break a nucleus up into its constituent protons and neutrons (Carroll & Ostlie 2007, pg. 299).

A nucleus of mass  $M_{\text{nuc}}$  and atomic mass number  $A$ , containing  $Z$  protons of mass  $m_p$  and  $(A-Z)$  neutrons of mass  $m_n$ , has a binding energy of

$$E_B = [(A-Z)m_n + Zm_p - M_{\text{nuc}}] c^2. \quad (476)$$

When comparing different nuclei it is more convenient to consider the average binding energy per nucleon,

$$f \equiv \frac{E_B}{A}, \quad (477)$$

which is also called the binding fraction. WIth the exception of hydrogen, typical values of  $f$  are around 8 MeV, with relatively small differences for nuclei of very different  $A$ . Figure 108 shows  $f$  as a function of atomic number  $A$ .

We see that  $f(A)$  increases sharply from hydrogen, then quickly flattens out and reaches a maximum of 8.5 MeV at  $A = 56$  (i.e., at  ${}^{56}\text{Fe}$ ), after which it continues to drop slowly. The increase for  $A < 56$  is a surface effect: particles at the surface of the nucleus experience less attraction by nuclear forces than those in the interior, which are completely surrounded by other particles. In a densely packed nucleus the surface area of the nucleus increases with radius slower than the volume (i.e., the number  $A$ ) such that the fraction of surface particles drops. The competing effect that causes the curve to fall back down for  $A > 56$  is the repulsive forces that the protons exert on each other. The Coulomb forces that the protons in the nucleus exert on each other is much further reaching than the nuclear forces (range is  $\sim 10^{-12}$  cm) and therefore does not experience the saturation of the nuclear forces (Kippenhahn & Weigert 1990, pg. 148).

The maximum at  ${}^{56}\text{Fe}$  indicates that it is the most tightly bound nuclei. In other words, the nucleus of  ${}^{56}\text{Fe}$  has the smallest mass per nucleon, so that any nuclear reaction brining the nucleus closer to this maximum will be exothermic. This either occurs via fusion from the left of  ${}^{56}\text{Fe}$  or via fission of heavy nuclei from the right. The nuclear fusion of  ${}^{56}\text{Fe}$  therefore provides a natural finishing point for the nuclear engines of stars (Kippenhahn & Weigert 1990, pg. 148).

In order to obtain fusion within the interior of stars, a Coulomb potential energy barrier between the positively charged nuclei must be overcome. If we assume that the energy required to overcome this barrier is provided by the thermal energy of the gas, and that all nuclei are moving non-relativistically, then the temperature  $T_{\text{classical}}$  required to overcome the barrier can be estimated:

$$\frac{3}{2} k T_{\text{classical}} = \frac{1}{4\pi\epsilon_0} \frac{Z_1 Z_2 e^2}{r}, \quad (478)$$

where  $Z_1$  and  $Z_2$  are the numbers of protons in each nucleus and  $r$  is their distance of separation (Carroll & Ostlie 2007, pg. 301). Assuming that the radius of a typical nucleus is on the order of 1 fm ( $10^{-15}$  m), we find that  $T_{\text{classical}} \sim 10^{10}$  K for a collision between two protons. We should expect this to be considerably higher than the average kinetic energy for stellar interiors since in normal stars we observe a slow energy release, rather than a nuclear explosion. However, this value is about  $10^3$  times larger than the central temperature of the Sun, which implies that, through classical effects only, we should not expect any fusion reactions to take place at all! In the high-energy tail of the Maxwell-Boltzmann distribution, the exponential factor here drops to  $e^{-1000} \sim 10^{-434}$ , which leaves absolutely no chance for the “mere”  $10^{57}$  nucleons in the entire Sun (Kippenhahn & Weigert 1990, pg. 149).

The saving grace that allows nuclear fusion to take place within stellar interiors arises from quantum tunnelling. The uncertainty principle asserts that we cannot know the position and momentum of a particle to arbitrary precision. The uncertainty in the position of one proton colliding with another may be so large that even though the kinetic energy of the collision is insufficient to overcome the classical Coulomb barrier, one proton may nevertheless find itself within the central potential well defined by the strong force of the other proton. A crude estimate of the quantum-mechanical temperature  $T_{\text{quantum}}$  required for nuclear reactions, we assume that a proton must be within one de Broglie wavelength of its target in order to tunnel through the Coulomb barrier. With this assumption we find that  $T_{\text{quantum}} \sim 10^7$  K for the case of two colliding protons. For larger nuclei this temperature increases due to the increased height of the Coulomb barrier and results in the well-separated phases of different nuclear burning during stellar evolution. This last part arises from what is known as the Gamow peak which describes a narrow energy band where the likelihood of a nuclear reaction is greatest (Carroll & Ostlie 2007, pg. 305).

The minimum stellar mass for nuclear fusion is  $0.08 M_{\odot}$  to ignite the fusion of hydrogen into helium (Schneider 2002, pg. 427). For more massive stars, ever heavier elements are generated by fusion throughout its lifetime: once the hydrogen is used up, helium will be burned, then carbon, oxygen, etc. This chain creates nuclei with masses progressively nearer the iron peak in Figure 108. As a result, the timescale for each succeeding sequence becomes shorter<sup>73</sup>. Once the iron nucleus is reached no more energy can be gained from fusion to heavier elements so that the pressure, which is normally balancing the star’s own gravity, can no longer be maintained. At this point the star collapses under its own gravity leading to WDs for low mass stars or to core-collapse supernovae (SNe II and SNe Ib,c) for more massive stars with  $M \gtrsim 8 M_{\odot}$  (Schneider 2002, pg. 48).

Classical novae are an example of an astrophysical situation in which fusion leads to unstable conditions. Novae are members of a general class of cataclysmic variables, which survive their release of energy (unlike SNe) and the outburst process can reoccur (Carroll & Ostlie 2007, pg. 673). The accepted theoretical model of a nova consists of a WD in a semidetached binary system that accretes matter at a rate of about  $10^{-8}$  to  $10^{-9} M_{\odot} \text{ yr}^{-1}$ . The hydrogen-rich gases accumulate on the surface of the white dwarf, where they are compressed and heated. At the base of this layer, turbulent mixing enriches the gases with the carbon, nitrogen, and oxygen of the white dwarf. The base of this enriched layer is supported by electron degeneracy pressure. Once this mixture reaches  $10^6$  K a shell of CNO-cycle hydrogen burning begins. For highly degenerate matter the pressure is independent of the temperature, so the shell source cannot dampen the reaction rate by expanding and cooling. The result is a runaway thermonuclear reaction, with temperatures reaching  $10^8$  K before the electrons lose their degeneracy. When the luminosity exceeds the Eddington limit, radiation pressure can lift the accreted material and expel it into space (Carroll & Ostlie 2007, pg. 683).

Another example of degenerate conditions leading to unstable fusion reactions is the explosion of a WD into a Type Ia SNe. The standard model that is assumed for these explosions is the destruction of a WD in a binary system with another star. If sufficient mass falls onto the WD, its mass can be driven near the Chandrasekhar limit, producing a catastrophic explosion. It is, however, still unclear the exact mechanisms that trigger this explosion. We will focus on the so-called single-degenerate model. In one version of this model, as the material from the secondary falls onto the primary, the helium in the gas will settle on top of the carbon-oxygen WD, becoming degenerate. When enough helium has accumulated, a helium flash will occur. This helium flash not only burns the helium into carbon and oxygen, but also sends a shock wave downward into the degenerate core, causing ignition of a runaway fusion of carbon and oxygen. Another version of this model doesn’t incorporate the helium flash, but rather supposes that the ignition of the carbon-oxygen core occurs when the star nears the Chandrasekhar limit (Carroll & Ostlie 2007, pg. 688).

<sup>73</sup> For a  $20 M_{\odot}$  star, the MS lifetime (core hydrogen burning) is  $\sim 10^7$  years, core helium burning requires  $10^6$  years, carbon burning lasts 300 years, oxygen burning takes 200 days, and silicon burning is completed in two days (Carroll & Ostlie 2007, pg. 531).

**QUESTION 11**

**Why do neutrons inside a neutron star not decay into protons and electrons?**

### QUESTION 11

**Why do neutrons inside a neutron star not decay into protons and electrons?**

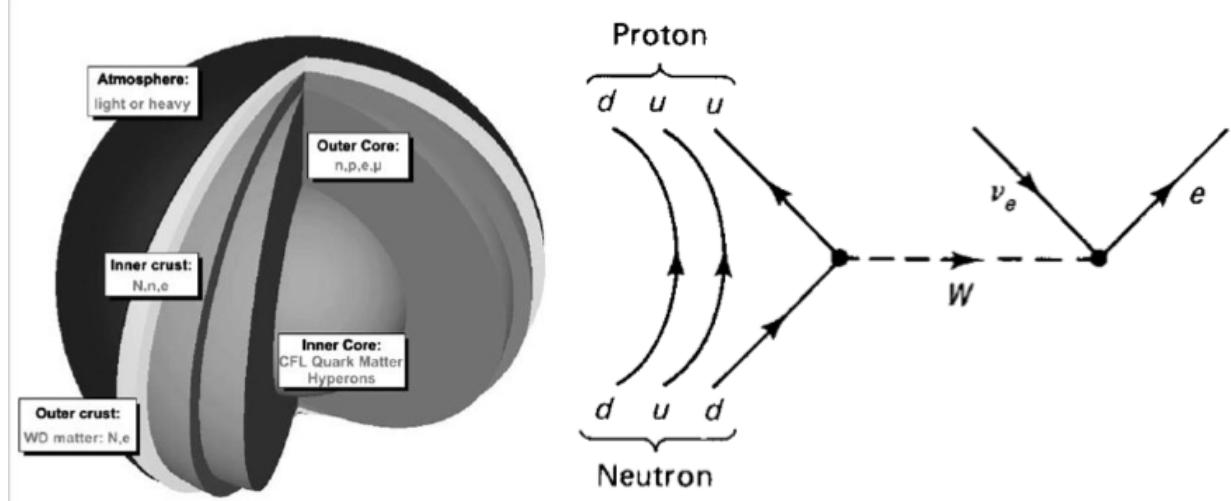


FIG. 109.— (left) Schematic representation of the interior of a NS. Image taken from Charles' study notes. (right) Feynman diagram of the  $\beta$ -decay of a neutron. Here a down quark (with charge  $-1/3$ ) is converted to an up quark (with charge  $+2/3$ ) during the transformation of a neutron [ddu] into a proton [uud] and electron. The  $W^-$  boson carries away the balance charge and subsequently decays into the electron and electron neutrino. Image taken from Yevgeni's study notes.

Degeneracy pressure is physically generated by forcing electrons into increased momentum states so as to satisfy the Pauli exclusion principle – the quantization of phase space requires that electrons sharing the “same” position must have different momenta. As such, an increased compression generates a higher distribution of momentum states (i.e., the top of the Fermi sea, given by Fermi energy  $E_F$ , is higher) (Charles).

The mass of a neutron is  $\sim 1.001$  the mass of a proton, and the mass of an electron is  $\sim 10^{-4}$  the mass of a proton. Electron capture,

$$e^- + p^+ \rightarrow n + \nu_e, \quad (479)$$

therefore requires about  $10^{-4}m_pc^2$  worth of kinetic energy in the reactants, approximately the same value as the rest energy of the electron. When the neutrino energy is accounted for, approximately 3 times the rest energy of the electron is needed. It is only at high densities when the electrons within the NS possess a significant amount of kinetic energy that reaction (479) can proceed. Prior to this, the nucleons are found in iron nuclei, assuming that previous to the NS was an iron WD at the centre of a massive supergiant star. This is the outcome of the minimum-energy compromise between the repulsive Coulomb force between the protons and the attractive nuclear force between all of the nucleons (Carroll & Ostlie 2007, pg. 579).

For ordinary (even relativistic) WDs,  $E_F \lesssim m_ec^2$ , so WDs are stable to electron capture. A collapsing WD, however, will increase its Fermi energy to infinity (as  $E_F \rightarrow \infty$  there is still insufficient electron degeneracy pressure to halt collapse, since there are still an enormous supply of low-energy electrons), making electron capture possible once the electron possess enough energy for reaction (479) to proceed. Once electron capture occurs, it cannot come into equilibrium with the reverse reaction,

$$n + \nu_e \rightarrow e^- + p^+, \quad (480)$$

called *assisted*  $\beta$ -decay, because neutrinos emitted from reaction (479) easily escape the WD. As electron capture occurs within the interior, the most stable arrangement of nucleons is one where the neutrons and protons are found in a lattice of increasingly neutron-rich nuclei so as to decrease the energy due to Coulomb repulsion between protons. This process is known as **neutronization** and produces a sequence of nuclei such as  $^{56}\text{Fe}$ ,  $^{62}\text{Ni}$ ,  $^{86}\text{Kr}$ , etc (Carroll & Ostlie 2007, pg. 580).

Ordinarily, the supernumerary neutrons produced during electron capture would revert back to protons via the standard  $\beta$ -decay process,

$$n \rightarrow e^- + p^+ + \bar{\nu}_e, \quad (481)$$

which has a half-life of about 10 minutes. This *does* occur within the collapsing NS, up until the point for which  $E_F \sim 3m_ec^2$  (Charels). Once the Fermi energy exceeds this point, there are no vacant states available for an emitted electron in reaction (481) to occupy, so the neutrons cannot decay back into protons (Carroll & Ostlie 2007, pg. 581). Hence, it is not entirely true that neutrons do not decay into protons and electrons, but after the Fermi sea has been filled, the NS will indeed become stable to  $\beta$ -decay.

When the density reaches  $\rho \sim 10^{14} \text{ kg m}^{-3}$ , the minimum-energy arrangement is one in which some of the neutrons are found *outside* the nuclei. The appearance of these free neutrons is called **neutron drip** and marks the start of a three-component mixture of a lattice of neutron-rich nuclei, nonrelativistic degenerate free neutrons, and relativistic degenerate electrons. The fluid of free

electrons has the striking property that it has no viscosity. This occurs because a spontaneous pairing of degenerate neutrons has taken place; the resulting combination of two fermions is a boson, with the collection forming a **superfluid** that flows without resistance. As the density increases further, to the point at which  $\rho \sim \rho_{\text{nuc}} \sim 10^{15} \text{ kg m}^{-3}$ , the nuclei effectively dissolve as the distinction between neutrons inside and outside of nuclei becomes meaningless. This results in a fluid mixture of free neutrons, protons, and electrons dominated by neutron degeneracy pressure, with both the neutrons and protons paired to form superfluids. As the density increases further, the ratio of neutrons:protons:electrons approaches a limiting value of 8:1:1, as determined by the balance between the competing processes of electron capture and  $\beta$ -decay inhibited by the presence of degenerate electrons (Carroll & Ostlie 2007, pg. 581).

#### INTERIOR STRUCTURE OF NEUTRON STARS

The properties of the NS material when  $\rho > \rho_{\text{nuc}}$  are still poorly understood. A complete theoretical description of the behaviour of a sea of free neutrons interacting via the strong force in the presence of protons and electrons is not yet available, and there is little experimental data on the behaviour of matter in this density range. A further complication is the appearance of sub-nuclear particles such as pions produced by the decay of a neutron into a proton and negatively charged pion, which occurs spontaneously when  $\rho > 2\rho_{\text{nuc}}$ . Figure 109 shows the interior structure of a NS obtained by numerically integrating general-relativistic version of stellar structure equations using sophisticated equations of state relating NS material to the density and pressure within the star (Carroll & Ostlie 2007, pg. 581). The different layers of a NS are:

- **Atmosphere:** About a cm in thickness.
- **Outer Crust:** Consists of heavy nuclei, in the form of either a fluid *ocean* or a solid lattice, and relativistic degenerate electrons. As the depth increases so does the density, and thus the abundance of neutron-rich nuclei. Neutron drip occurs at the bottom of the outer crust.
- **Inner Crust:** Consists of a three-part mixture of a lattice of neutron-rich nuclei, a superfluid of free neutrons, and relativistic degenerate electrons. The bottom of the inner crust occurs where  $\rho \approx \rho_{\text{nuc}}$ , and the nuclei dissolve.
- **Interior:** Consists primarily of superfluid neutrons, with a smaller number of superfluid, superconducting protons and relativistic degenerate electrons.
- **Core:** There may or may not be a solid core consisting of pions or other sub-nuclear particles.

#### NUCLEAR REACTIONS

In general, there are three types of nuclear reactions involving elementary particles, corresponding to the three types of bosonic force carriers existing in nature. Electromagnetic reactions involve photons, strong nuclear reactions involve gluons (which are able to change quarks from one type to another), and weak interactions involve one of  $W^+$ ,  $W^-$ , or  $Z_0$  bosons. If a reaction involves a photon it is immediately classified as electromagnetic, and if a neutrino is involved a reaction is immediately classified as weak; other reactions require closer investigation into their origin. For instance, the  $\beta$ -decay of a neutron in reaction (481) is a weak interaction, since it involves the emission of a neutrino, and is mediated by a  $W^-$  boson; see Figure 109. For nuclear reactions there are a number of conservation laws that must be met: (i) conservation of colour, (ii) conservation of charge, (iii) conservation of baryon number, and (iv) conservation of lepton number (Yevgeni).

**QUESTION 12**

**Give examples of degenerate matter.**

**QUESTION 12**

**Give examples of degenerate matter.**

**QUESTION 13**

**What is the typical temperature of matter accreting on a star, a white dwarf, a neutron star, a stellar mass black hole, and a supermassive black hole? In what wavelength range would one best find examples of such sources?**

**QUESTION 13**

**What is the typical temperature of matter accreting on a star, a white dwarf, a neutron star, a stellar mass black hole, and a supermassive black hole? In what wavelength range would one best find examples of such sources?**

**QUESTION 14**

The weak equivalence principle for gravity corresponds to being able to find a coordinate system for a region of spacetime. What kind of coordinate system is this, and why?

### QUESTION 14

**The weak equivalence principle for gravity corresponds to being able to find a coordinate system for a region of spacetime. What kind of coordinate system is this, and why?**

The weak equivalence principle (WEP) asserts that the trajectory of a freely falling “test” body (one not acted upon by such forces as electromagnetism and too small to be affected by tidal gravitational forces) is independent of its internal structure and composition. In the simplest case of dropping two different bodies in a gravitational field, WEP states that the bodies fall with the same acceleration (Will 2006).

A direct test of WEP is the comparison of the acceleration of two laboratory-sized bodies of different composition in an external gravitational field. If the principle were violated, then the accelerations of the two bodies would differ. In classical mechanics there are two a priori independent concepts of mass: inertial mass  $m_I$ , which accounts for the resistance against acceleration, and the gravitational mass  $m_G$ , which is the mass that gravity couples to. There is no reason to assume that  $m_I = m_G$  and so in the presence of a gravitational field  $g$ , we expect the object to pick up an acceleration given by  $m_I a = m_G g$ . We note that the inertial mass of a typical laboratory body is made up of several types of mass-energy: rest energy, electromagnetic energy, weak-interaction energy, and so on. If one of these forms contributes differently to  $m_G$  differently than it does to  $m_I$  then a violation of WEP will result. A variety of experiments have been successful in showing that  $m_I$  and  $m_G$  are equivalent to one part in  $10^{-13}$  (Will 2006).

The fact that  $m_I = m_G$  is a profound concept and led Einstein to argue that locally the effects of gravity and acceleration are indistinguishable. This can be visualized with the following thought experiment. Suppose that we place Alice in a sealed elevator somewhere in outer space. In the absence of any external forces Alice will float as well as two rocks that she drops out of her hands. Now suppose that Bob ties a rope to the top of the elevator and begins to pull on it with a constant acceleration. In this case Alice will be pushed to the bottom of the elevator with a constant force and so will the rocks. Let’s now change gears and imagine the same elevator but this time we fix it to a cable in the presence of a constant gravitational field. Once again Alice will be pushed to the bottom of the elevator in addition to the rocks she releases from her hands. Now suppose that we cut the cable holding the elevator so that it falls freely in the gravitational field. Alice and the rocks will now float, just as they did when the elevator was placed in an empty party of the universe.

This thought experiment leads us to two important concepts. Firstly, we see that, locally, the effects of gravity and acceleration are indistinguishable. Moreover, we see that the effects of gravity can be eliminated by going to a freely-falling reference frame (or coordinate system) (Blau 2010, pg. 15). This should not come as a surprise. In classical mechanics, if the free fall in a constant gravitational field is described by the equation

$$\ddot{x} = g \text{ (+ other forces)}, \quad (482)$$

then in the accelerated coordinate system

$$\xi(x, t) = x - gt^2/2, \quad (483)$$

the same physics is described by the equation

$$\ddot{\xi} = 0 \text{ (+ other forces)}, \quad (484)$$

so that the effect of gravity has been eliminated by going to the freely falling coordinate system  $\xi$ .

We should note that the above thought experiment was somewhat restrictive in that it assumed a uniform gravitational field. In a non-uniform gravitational field the effects of gravity cannot be eliminated by going to a freely falling coordinate system. This is only possible locally, on such scales on which the gravitational field is essentially constant. To see this, consider the same elevator attached to a cable, but this time suspended in a spherically symmetric gravitational field. In this case, when we cut the cable the rocks will move closer to Alice, so that she will be forced to conclude that there is some sort of force responsible for this (Blau 2010, pg. 17).

Let’s now try to reformulate all of this in the language of general relativity. Through the use of the Euler-Lagrange equations we write the geodesic equation as

$$\ddot{x}^a + \Gamma_{bc}^a \dot{x}^b \dot{x}^c = 0, \quad (485)$$

where  $\Gamma_{bc}^a$  is the Christoffel symbol of the second kind, which is related to the metric  $g_{ab}$  via

$$\Gamma_{bc}^a = \frac{1}{2} g^{am} (g_{mb|c} + g_{mc|b} - g_{bc|m}), \quad (486)$$

and the notation  $g_{bc|m}$  simply represents the partial derivative of  $g_{bc}$  along the coordinate  $x^m$ . The geodesic equation describes the world line of a particle, that is the path that minimizes proper time.

The aim here is to find a coordinate system such that, at least in a small non-empty neighbourhood of some point  $P$ , the coordinate curves are themselves geodesics. In such a coordinate system, the geodesic equation in the applicable neighbourhood reduces to the form

$$\ddot{x}^a = 0. \quad (487)$$

It turns out that this is always possible to do at exactly the position  $P$  and its applicability outside of this point depends on the curvature of spacetime there. This is because it is always possible to find coordinates such that  $\Gamma_{bc}^a = 0$  at  $P$ , but  $\Gamma_{bc}^a$  will in general not vanish if we move away from  $P$ . In this coordinate system the coordinate curves are locally geodesic curves. Since at this

point, the observer fixed in these coordinates has no forces acting on them in the sense of Newton's equations of motion, these coordinates define locally the instantaneous freely-falling reference frame (Dyer 2010).

We can learn several things from the equivalence principle. First, we note that equation (485) is true for all metrics, and therefore includes the Minkowski metric  $\eta_{ab} \equiv \text{diag}(+1, -1, -1, -1)$ . It is only when flat spacetime is expressed using coordinates of an inertial reference frame that the Christoffel symbols vanish and we can write the geodesic equation as equaiton (487). Of course, non-inertial observers in Minkowski space will have modified metrics in which  $\Gamma_{bc}^a \neq 0$  and such an observer would conclude that some sort of force is responsible for observed inertial motion described in these non-inertial coordinates (Blau 2010, pg. 24). This is how we arrive at the notion of fictitious forces such as the centrifugal and coriolis forces. It is the fact that the components of  $\Gamma_{bc}^a$  do not transform as the components of a tensor that these fictitious forces can vanish in some frames even if they do not in others.

For curved spacetime the Christoffel symbols play the role of the gravitational force term, and thus in this sense the components of the metric play the role of the gravitational potential. The equivalence principle suggests that we can only locally relate the curved metric to the Minkowski metric via a suitable coordinate transformation to locally geodesic coordinates. In this sense gravity plays the role of a fictitious force that can be eliminated by transforming to the instantaneous freely-falling (inertial) reference frame (Blau 2010, pg. 29). This can only be accomplished in a small neighbourhood about our point  $P$  and in general can only be accomplished for all spacetime if and only if we are in flat spacetime. To determine if one is in a truly curved spacetime would require the computation of the Riemann curvature tensor, which is constructed from the second derivatives of the metric, and has the property that its components vanish if and only if the metric is a coordinate transformation of the Minkowski metric.

**QUESTION 15**

**What are the typical detectors used in gamma-ray, X-ray, UV, visible, infrared, submm, and radio observations?**

**QUESTION 15**

**What are the typical detectors used in gamma-ray, X-ray, UV, visible, infrared, submm, and radio observations?**

**QUESTION 16**

You don't usually need to cool down the detectors for short wavelength (e.g., X-ray) observations, but it's critical to cool down the detectors in long wavelength (e.g., far-IR) observations. Why is this?

### QUESTION 16

**You don't usually need to cool down the detectors for short wavelength (e.g., X-ray) observations, but it's critical to cool down the detectors in long wavelength (e.g., far-IR) observations. Why is this?**

Detectors in long wavelength observations (e.g., IR) must be cooled down since their blackbody radiation may cause them to glow in the wavelength band of interest (Carroll & Ostlie 2007, pg. 168). A blackbody of temperature  $T$  emits a continuous spectrum that peaks at a wavelength  $\lambda_{\max}$ , which becomes smaller with increasing temperature; the relation between  $\lambda_{\max}$  and  $T$  is known as **Wien's displacement law**:

$$\lambda_{\max}T = 0.0029 \text{ m K} \approx (10 \mu\text{m})(300 \text{ K}) \approx (500 \text{ nm})(5800 \text{ K}). \quad (488)$$

We see from the above relation that the peak wavelength of a blackbody of temperature  $300 \text{ K}$  is roughly  $10 \mu\text{m}$ . Therefore, IR detectors operating at room temperature conditions must be cooled in order to prevent observations from being overwhelmed by the telescope's inherent blackbody glow. This is not necessary for short wavelength observations (e.g., X-ray) since the temperature required for a blackbody to peak in the Å range is  $\sim 10^7 \text{ K}$ .

When observing wavelengths greater than  $2.3 \mu\text{m}$  from the Earth, blackbody radiation from the telescope and the atmosphere itself begins to dominate other sources of background. Measurements of faint astronomical objects have to be made by alternatively observing the field containing the source and a nearby "empty" one. In this process, known as chopping, the latter signal is subtracted from the former in order to eliminate the strong background. Note that not all IR wavelengths are equally suitable for ground-based observations because the transparency of the Earth's atmosphere is strongly variable with wavelength. There is no possibility of cooling the optics and support structure of a ground-based telescope since this would simply cause them to become coated with ice from the water vapour always present in the air (Glass 1999, pg. 26).

Placing a cooled IR telescope in space eliminates the problems associated with the transmission and emission of the Earth's atmosphere. The telescope can be cooled either through a liquid cryogen such as helium or through more passive techniques that utilize the frigidity of space itself. Latter methods involve Sun shading, optimal satellite orientation, special surface finishes (i.e., coating the surface with gold instead of aluminum in order to reduce its infrared emissivity), and multiple radiation shielding (Glass 1999, pg. 133). Passive cooling techniques have the benefit that the telescope will have an indefinite lifetime, without the use of cryogenic liquids. The Spitzer Space Telescope (SST) is an example of an infrared telescope that has employed both methods for its operation. SST was equipped with liquid helium cryogen that cooled the telescope for 5.5 years to  $\sim 3 \text{ K}$ . The coolant eventually ran out in May 2009 which marked the onset of its "Warm Mission" where the telescope runs at  $\sim 30 \text{ K}$ . This warm phase marked the end of the instrument's longer wavelength multiband imaging photometer and its infrared spectrograph, but its two shortest-wavelength detectors in its infrared array camera continue to work<sup>74</sup>. The passive cooling techniques employed by SST involve a Sun shield to block radiation from the Sun as well as an Earth-trailing orbit that removes it from the large heat load of the planet (SST Wikipedia page).

Another reason that long wavelength detectors (specifically IR) need to be cooled is due to the requirements of the specific detectors that they employ. IR observations cannot be performed using CCDs since the large size of the forbidden band gap of silicon means that they are blind to all light with wavelengths greater than  $1.1 \mu\text{m}$  (Frederick 2010, pg. 265). Instead, IR observations usually make use of thermal detectors which detect radiation through the increase in temperature that its absorption causes in the sensing element. The two main types of IR detector are: the photoconductor for the NIR and MIR and the bolometer for the FIR (Kitchin 2003, pg. 48).

Photoconductive cells exhibit a change in conductivity with the intensity of their illumination. The mechanism for that change is the absorption of the radiation by the electrons in the valence band of a semiconductor and their consequent elevation to the conduction band. The conductivity therefore increases with increasing illumination, and is monitored by a small bias current. There is a cut-off point determined by the minimum energy required to excite a valence electron over the band gap. Typically a semiconducting material such as silicon or germanium will be used, but for the reason discussed above, it must be doped with another element. The doping agent is chosen such that electrons or gaps from the doping atom occupy levels close to the bottom of the conduction band so little energy is required to excite them (Kitchin 2003, pg. 49). The most commonly used material for wavelengths greater than  $50 \mu\text{m}$  is germanium doped with gallium, denoted by Ge(Ga). For example, the InfraRed Astronomy Telescope (IRAS) used small arrays of Ge(Ga) detectors to survey the sky at 12, 25, 60, and  $100 \mu\text{m}$ . Because of the smaller band gaps involved, dark currents in infrared photoconductive arrays can become a severe problem and therefore must operate at low temperature. Although some NIR arrays work well with liquid-nitrogen cooling ( $77 \text{ K}$ ), MIR arrays usually require liquid-helium cooling ( $4 \text{ K}$ ) instead (Frederick 2010, pg. 269).

A bolometer is a device that changes its electrical resistivity in response to heating by illuminating radiation. A bolometer works as a two-element device consisting of: (1) a thermometer which senses the temperature increase produced in (2) an absorber, when the latter is exposed to an incoming light beam. A heat sink of temperature  $T_0$  encloses the absorber and thermometer in an evacuated cavity. A strip of material with conductance  $G$  connects the absorber and the heat sink. The absorber is then exposed to incoming light that deposits energy in the absorber at a rate  $P_{\text{in}}$ . After a time, the temperature of the absorber increases by  $\Delta T$ , which is measured by the thermometer in order to determine  $P_{\text{in}}$ . To maximize the sensitivity of the measurement, which is just  $\Delta T/P_{\text{in}}$ , the conductance of the link, area of the absorber, and temperature  $T_0$  of the reservoir must all be kept small (Frederick 2010, pg. 270). Germanium doped with gallium is widely used for the bolometer material with a metal-coated dielectric as the absorber (Kitchin 2003, pg. 52).

<sup>74</sup> See the article at [http://www.nasa.gov/mission\\_pages/spitzer/news/spitzer-20090506.html](http://www.nasa.gov/mission_pages/spitzer/news/spitzer-20090506.html)

**QUESTION 17**

Compare the S/N ratios between the following two cases where photon noise is dominant (assume an unresolved point source): [A] 1-minute exposure with a 10-m telescope; [B] 10-minute exposure with a 1-m telescope.

**QUESTION 17**

**Compare the S/N ratios between the following two cases where photon noise is dominant (assume an unresolved point source): [A] 1-minute exposure with a 10-m telescope; [B] 10-minute exposure with a 1-m telescope.**

**QUESTION 18**

**Describe the difference between linear and circular polarizations.**

**QUESTION 18**

**Describe the difference between linear and circular polarizations.**

**QUESTION 19**

What's the field of view of a 2K x 2K CCD camera on a 5-m telescope with f/16 focal ratio. The pixel size is 20 micron. If you bring this to a 10-m telescope with the same focal ratio, what will be the field of view? Give your answer using the Etendue conservation rule.

**QUESTION 19**

**What's the field of view of a 2K x 2K CCD camera on a 5-m telescope with f/16 focal ratio. The pixel size is 20 micron. If you bring this to a 10-m telescope with the same focal ratio, what will be the field of view? Give your answer using the Etendue conservation rule.**

**QUESTION 20**

Sketch and give the equations for each of the following distributions: 1. Gaussian (Normal distribution); 2. Poisson distribution; 3. Log-normal distribution. Give two examples from astrophysics where each of these distributions apply.

## QUESTION 20

**Sketch and give the equations for each of the following distributions:** 1. Gaussian (Normal distribution); 2. Poisson distribution; 3. Log-normal distribution. Give two examples from astrophysics where each of these distributions apply.

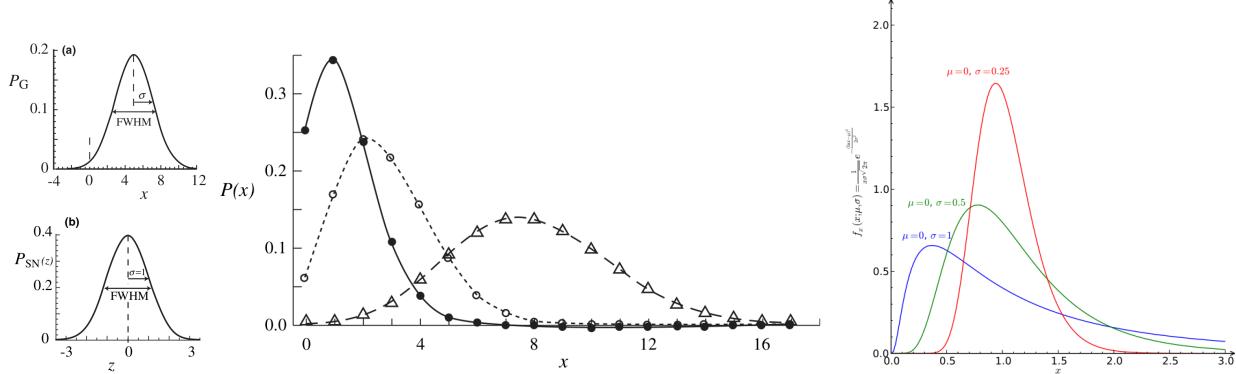


FIG. 110.— (left) The top plot shows a Gaussian distribution with  $\mu = 5$  and  $\sigma = 2.1$ . The bottom plot shows the standard normal distribution which is the Gaussian distribution with  $\mu = 0$  and  $\sigma = 1$ . Image taken from Frederick (2010). (middle) The Poisson distribution for values of  $\mu = 1.4$  (filled circles),  $\mu = 2.8$  (open circles), and  $\mu = 8.0$  (open triangles). Only the plotted symbols have meanings as probabilities, not the continuous lines used to denote the different distributions. We see that the dispersion of the distribution increases as  $\mu$  increases. In addition, as  $\mu$  increases the Poisson distribution begins to resemble a Gaussian distribution. Often we can approximate a Poisson distribution with a more tractable Gaussian as long as  $\mu \gtrsim 10$  (Abraham 2011). Image taken from Frederick (2010). (right) Lognormal distribution plotted with a linear abscissa. Image taken from Wikipedia.

Let's consider the problem of determining a parameter (e.g., the brightness of a star) by making several measurements under nearly identical circumstances. we define the population under investigation as the hypothetical set of all possible measurements that could be made with an experiment substantially identical to our own. We then imagine that we make our actual measurements by drawing a finite sample from this much larger population. There are usually two quantities related to the underlying population that we wish to compute: the population mean ( $\mu$ ) and the standard deviation ( $\sigma$ ) of the population. We estimate these quantities from our sample of  $N$  measurements by computing the sample mean ( $\bar{x}$ ) and standard deviation ( $s$ ) of the sample (Frederick 2010, pg. 42).

Suppose we have a large population,  $Q$ , from which we wish to make a measurement of some quantity  $x$ . We call  $x$  a random variable in the sense that its value will be unpredictable and depend on the features of  $Q$ . We define  $P_Q$  to be the function that describes how likely it is to obtain a particular value of  $x$  in a single trial. We call  $P_Q$  the probability distribution of  $x$  in  $Q$  and we have that  $P_Q(x)dx$  is the probability that the result of a single trial will return a value between  $x$  and  $x+dx$ . In experimental situations, we sometimes know or suspect something about the probability distribution before conducting any quantitative trials. In fact, nature seems to favour a small number of distributions. We will now focus on three of those distributions that have significant astrophysical relevance (Frederick 2010, pg. 47).

### GAUSSIAN (NORMAL) DISTRIBUTION

The Gaussian, or normal, distribution is the most important continuous distribution in the statistical analysis of data. If a population has a Gaussian distribution, then in a single trial the probability that  $x$  will have a value between  $x$  and  $x+dx$  is

$$P_G(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad (489)$$

where  $\mu$  and  $\sigma$  are the mean and standard deviation of the distribution (Frederick 2010, pg. 49). It is useful to describe the dispersion of a Gaussian by specifying its full width at half maximum (FWHM) which denotes the separation in  $x$  between the two points satisfying  $P_G(x) = 1/2$ . The FWHM is related to the variance via  $\text{FWHM} = 2.354\sigma$ . The dispersion of a distribution determines how likely it is that a single sample will be close to the population mean.

The Gaussian distribution is particularly important due to its connection with the **Central Limit Theorem**. Suppose that we have a general probability distribution  $P(x)$  which has some mean  $\mu$  and variance  $\sigma^2$ . The Central Limit Theorem states that if  $\{x_1, x_2, \dots, x_n\}$  is a sequence of  $n$  independent random variables drawn from  $P$ , then as  $n$  becomes large, the distribution of the variables

$$\bar{x}_n = \frac{1}{n} \sum x_i \quad (490)$$

will approach a Gaussian distribution with mean  $\mu$  and and variance  $\sigma^2/n$  (Frederick 2010, pg. 52).

Since so many measurements in science are averages of individual experiments, the Central Limit Theorem means that the Gaussian distribution will be central to the analysis of experimental results. In addition, since the variance of the average is proportional to  $1/n$  relates directly to the problem of estimating uncertainty. Since  $s$ , the estimated standard deviation, is the best

guess for  $\sigma$ , we should estimate  $\sigma_\mu(n)$ , the standard deviation of  $\bar{x}_n$ , the mean, as

$$\sigma_\mu(n) = \frac{s}{\sqrt{n}}. \quad (491)$$

This value is commonly quoted as the uncertainty in a measurement.

The Gaussian distribution has a wide variety of applications in astrophysics. One application is its use in approximating the shapes of many observables including the width of spectral lines (their true profile is closer to Lorentzian) as well as the Doppler broadening produced by absorbers with a thermal velocity distribution (Draine 2011, pg. 58). It is also used to approximate the shape of the Airy disk that results from diffraction of the image of a point source. The distribution of noise in radio receivers and other astronomical instruments is also well governed by a Gaussian distribution. Moreover – and as described above – the Gaussian distribution is used in the analysis of astronomical data to gauge the significance of a measurement. This typically involves citing how many standard deviations the experimental result is accurate to<sup>75</sup>.

#### POISSON DISTRIBUTION

The Poisson distribution describes a population encountered in certain counting experiments. These are cases in which the random variable,  $x$ , is the number of events counted in a unit time (e.g. the number of photons hitting a detector every minute). For counting experiments where non-correlated events occur at an average rate,  $\mu$ , the probability of counting  $x$  events in a single trial is

$$P_p(x) = \frac{\mu^x}{x!} e^{-\mu}. \quad (492)$$

Here  $P_p(x)$  is the Poisson distribution in which  $x$  is restricted to positive integer numbers.

Figure 110 plots the Poisson distribution for three different values of  $\mu$ . We see that as  $\mu$  increases, so does the dispersion of the distribution. The reason for this is that the variance of a Poisson distribution is exactly equal to its mean:

$$\sigma^2 = \mu. \quad (493)$$

This property has important consequences for planning and analyzing experiments. For example, suppose we count the number of photons  $N$  that arrive at our detector in  $t$  seconds. If we count  $N$  things in a single trial we can estimate that the average result of a single trial of length  $t$  seconds will be a count of  $\mu \approx \bar{x} = N$  photons. The uncertainty in this approximation can be judged by the standard deviation of the population from which the measurement was drawn, namely the Poisson distribution. The uncertainty of the measurement should then be  $\sigma = \sqrt{\mu} \approx \sqrt{N}$  and the fractional uncertainty in counting  $N$  events is therefore  $\sigma/\mu \approx 1/\sqrt{N}$ . Thus, to decrease the uncertainty in our estimate of the photon arrival rate we should increase the number of photons we count, either by increasing the exposure time or size of the telescope (Frederick 2010, pg. 48).

The most applicable example of the use of the Poisson distribution in astronomy involves counting photons. We are justified in using the Poisson distribution since the arrival of successive photons are independent from each other. Hence, if the integration over time,  $t$ , of photons arriving at a rate,  $\lambda$ , has a mean of  $\mu = \lambda t$  photons, then the fluctuation on this number will be  $\sigma = \sqrt{\mu}$ . In practice, we estimate  $\mu$  since we only know the number of photons in a single exposure. For photon-limited observations, such as CCD images or spectra  $\mu = \lambda t$  while  $\sigma = \sqrt{\lambda t}$ . Hence, if we integrate over more time, then since  $\sigma \propto \sqrt{t}$  and the signal strength is  $\propto t$ , we have that the signal-to-noise ratio is  $\propto \sqrt{t}$ , at least in the sky-limited (exposure in which the limiting factor is the background sky flux) case (Wall & Jenkins 2003, pg. 29).

The Poisson distribution can be applied to other astronomical counting calculations, such as those that count SN explosions. If supernovae are assumed to occur independently of each other and are observed to occur at some rate  $R$ , then the number of supernovae one expects to count in some time interval can be judged by the Poisson distribution. Cappellaro et al. (1997) estimated the rate of different types of supernovae (namely Ia, Ib/c, and II) as a function of galaxy morphology by combining data from five supernova searches. By combining the data from different searches, the statistical error in reporting the rate of supernovae explosions was decreased by increasing the number of counts.

#### LOGNORMAL DISTRIBUTION

The Lognormal distribution is one in which the logarithm of the random variable is normally distributed. The mathematical form of the Lognormal distribution is similar to equation (489), but with the following modifications:

$$P_L(x)dx = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right]. \quad (494)$$

Note the addition of  $x$  in the denominator of the prefactor to the exponential. A plot of the Lognormal distribution for various values of  $\mu$  and  $\sigma$  is displayed in Fig. 110.

There are a variety of uses of the Lognormal distribution in astrophysics. For example, Chabrier (2003) has used it to model the initial mass function for stars with masses below  $1 M_\odot$ . In general, if the fluctuations of a quantity are proportional to the quantity then the distribution of values will tend to be lognormal (Gaskell 2004).

<sup>75</sup> For related information see <http://astronomy.swin.edu.au/cms/astro/cosmos/g/Gaussian+Function>

**QUESTION 21**

You are trying to determine a flux from a CCD image using aperture photometry, measuring source(+sky) within a 5-pixel radius, and sky within a 20–25 pixel annulus. Assume you find 10000 electrons inside the aperture and 8100 electrons in the sky region, and that the flux calibration is good to 1%. What is the fractional precision of your measurement? (Ignore read noise.) More generally, describe how you propagate uncertainties, what assumptions you implicitly make, and how you might estimate errors if these assumptions do not hold.

**QUESTION 21**

You are trying to determine a flux from a CCD image using aperture photometry, measuring source(+sky) within a 5-pixel radius, and sky within a 20–25 pixel annulus. Assume you find 10000 electrons inside the aperture and 8100 electrons in the sky region, and that the flux calibration is good to 1%. What is the fractional precision of your measurement? (Ignore read noise.) More generally, describe how you propagate uncertainties, what assumptions you implicitly make, and how you might estimate errors if these assumptions do not hold.

**QUESTION 22**

Suppose you measure the brightness of a star ten times (in a regime where source-noise dominates; you will be given 10 values for the total number of counts): (1) How do you calculate the mean, median, and mode and standard deviation?

(2) How can you tell if any points are outliers? Say some points are outliers, what do you do now (ie. how does this impact the calculation of the quantities in part 1)?

**QUESTION 22**

Suppose you measure the brightness of a star ten times (in a regime where source-noise dominates; you will be given 10 values for the total number of counts): (1) How do you calculate the mean, median, and mode and standard deviation?

(2) How can you tell if any points are outliers? Say some points are outliers, what do you do now (ie. how does this impact the calculation of the quantities in part 1)?

**QUESTION 23**

Suppose you do an imaging search for binaries for a sample of 50 stars, and that you find companions in 20 cases. What binary fraction do you infer? Suppose a binary-star fraction of 50% had been found previously for another sample (which was much larger, so you can ignore its uncertainty). Determine the likelihood that your result is consistent with that fraction.

**QUESTION 23**

Suppose you do an imaging search for binaries for a sample of 50 stars, and that you find companions in 20 cases. What binary fraction do you infer? Suppose a binary-star fraction of 50% had been found previously for another sample (which was much larger, so you can ignore its uncertainty). Determine the likelihood that your result is consistent with that fraction.

## APPENDIX

Here we include a list of acronyms referenced in this document and commonly used in astronomy.

TABLE 11

Acronym	Meaning
AGB	Asymptotic Giant Branch
AGN	Active Galactic Nuclei
DM	Dark Matter
dSph	Dwarf Spheroidal Galaxy
EW	Equivalent Width
EUV	Extreme Ultraviolet
FIR	Far Infrared
GMC	Giant Molecular Cloud
HB	Horizontal Branch
HR	Hertzsprung-Russell
IMF	Initial Mass Function
IR	Infrared
ISM	Interstellar Medium
ISRF	Interstellar Radiation Field
LBV	Luminous Blue Variable
LINER	Low-Ionization Nuclear Emission-Line Region
LMC	Large Magellanic Cloud
MHD	Magnetohydrodynamical
MIR	Mid Infrared
MLR	Mass-to-Light Ratio
MS	Main Sequence
MW	Milky Way
NIR	Near Infrared
PAH	Polycyclic Aromatic Hydrocarbon
PDMF	Present Day Mass Function
PDR	Photodissociation Region
PN	Planetary Nebula
RGB	Red Giant Branch
SED	Spectral Energy Distribution
SFR	Star Formation Rate
SGB	Subgiant Branch
SMBH	Supermassive Black Hole
SMC	Small Magellanic Cloud
SN	Supernova
SNR	Supernova Remnant
UIR	Unidentified IR
ULIRG	Ultra-luminous Infrared Galaxy
UV	Ultraviolet
WD	White Dwarf
WR	Wolf-Rayet Star
ZAMS	Zero Age Main Sequence

## REFERENCES

- Abraham, R. 2011, AST 2040 Lecture Notes
- Ajello, M. 2009, ArXiv e-prints
- Albrecht, A. in , NATO ASIC Proc. 565: Structure Formation in the Universe, ed. R. G. CrittendenN. G. Turok, 17
- Appenzeller, I. & Mundt, R. 1989, A&A Rev., 1, 291
- Bahcall, J. N. & Tremaine, S. 1981, ApJ, 244
- Bekki, K. & Couch, W. J. 2011, MNRAS, 415, 1783
- Bennett, J., Donahue, M., Schneider, N., & Voit, M. 2012, The Essential Cosmic Perspective (San Francisco: Addison-Wesley)
- Binney, J. & Tremaine, S. 1994, Galactic Dynamics (Princeton: Princeton University Press)
- Birkinshaw, M. 1999, Phys. Rep., 310, 97
- Blau, M. 2010, Lecture Notes on General Relativity (Online)
- Bromm, V., Coppi, P. S., & Larson, R. B. 2002, ApJ, 564, 23
- Bruzual A., G. & Charlot, S. 1993, ApJ, 405, 538
- Cappellaro, E., Turatto, M., Tsvetkov, D. Y., Bartunov, O. S., Pollas, C., Evans, R., & Hamuy, M. 1997, A&A, 322, 431
- Carlberg, R. G., Dawson, P. C., Hsu, T., & Vandenberg, D. A. 1985, ApJ, 294, 674
- Carroll, B. W. & Ostlie, D. A. 2007, An Introduction to Modern Astrophysics: Second Edition (San Francisco: Addison Wesley)
- Chaboyer, B., Demarque, P., Kernan, P. J., & Krauss, L. M. 1998, ApJ, 494, 96
- Chabrier, G. 2003, PASP, 115, 763
- Chen, C.-H. R., Chu, Y.-H., Gruendl, R. A., Gordon, K. D., & Heitsch, F. 2009, ApJ, 695, 511
- Chiang, E. I. & Goldreich, P. 1997, ApJ, 490, 368
- . 1999, ApJ, 519, 279

- Cole, G. H. & Woolfson, M. W. 2002, Planetary Science: The Science of Planets Around Stars (Bristol:IPP)
- Coles, P. & Lucchin, F. 2002, Cosmology: The Origin and Evolution of Cosmic Structure (New York:Wiley)
- Davis, T. M. & Lineweaver, C. H. 2004, *pasa*, 21, 97
- Dodelson, S. 2003, Modern Cosmology (San Diego: Academic Press)
- Draine, B. T. 2011, Physics of the Interstellar and Intergalactic Medium (Princeton: Princeton University Press)
- Dressler, A. 1980, *ApJ*, 236, 351
- Dyer, C. C. 2010, AST 2060 Lecture Notes
- Dyson, J. E. & Williams, D. A. 1997, The Physics of the Interstellar Medium (London: Institute of Physics Publishing)
- Eisenstein, D. J., Zehavi, I., Hogg, D. W., Scoccimarro, R., Blanton, M. R., Nichol, R. C., Scranton, R., Seo, H.-J., Tegmark, M., Zheng, Z., Anderson, S. F., Annis, J., Bahcall, N., Brinkmann, J., Burles, S., Castander, F. J., Connolly, A., Csabai, I., Doi, M., Fukugita, M., Frieman, J. A., Glazebrook, K., Gunn, J. E., Hendry, J. S., Hennessy, G., Ivezić, Z., Kent, S., Knapp, G. R., Lin, H., Loh, Y.-S., Lupton, R. H., Margon, B., McKay, T. A., Meiksin, A., Munn, J. A., Pope, A., Richmond, M. W., Schlegel, D., Schneider, D. P., Shimasaku, K., Stoughton, C., Strauss, M. A., SubbaRao, M., Szalay, A. S., Szapudi, I., Tucker, D. L., Yanny, B., & York, D. G. 2005, *ApJ*, 633, 560
- Farinella, P. & Zappalà, V. 1997, Advances in Space Research, 19, 181
- Filippenko, A. V. 1997, *ARA&A*, 35, 309
- Fitzpatrick, E. L. & Massa, D. 2007, *ApJ*, 663, 320
- Frederick, R. C. 2010, To Measure the Sky: An Introduction to Observational Astronomy (Cambridge: Cambridge University Press)
- Gaskell, C. M. 2004, *ApJ*, 612, L21
- Glass, I. S. 1999, Handbook of Infrared Astronomy (Cambridge: Cambridge University Press)
- Gómez, P. L., Nichol, R. C., Miller, C. J., Balogh, M. L., Goto, T., Zabludoff, A. I., Romer, A. K., Bernardi, M., Sheth, R., Hopkins, A. M., Castander, F. J., Connolly, A. J., Schneider, D. P., Brinkmann, J., Lamb, D. Q., SubbaRao, M., & York, D. G. 2003, *ApJ*, 584, 210
- Gratton, R. G., Bragaglia, A., Carretta, E., Clementini, G., Desidera, S., Grundahl, F., & Lucatello, S. 2003, *A&A*, 408, 529
- Gray, R. O. & Corbally, C. J. 2009, Stellar Spectral Classification (Princeton: Princeton University Press)
- Griffiths, D. J. 2005, Introduciton to Quantum Mechanics (New Jersey: Pearson Prentice Hall)
- Hamada, T. & Salpeter, E. E. 1961, *ApJ*, 134, 683
- Hansen, B. M. S., Anderson, J., Brewer, J., Dotter, A., Fahlman, G. G., Hurley, J., Kalirai, J., King, I., Reitzel, D., Richer, H. B., Rich, R. M., Shara, M. M., & Stetson, P. B. 2007, *ApJ*, 671, 380
- Hansen, C. J., Kawaler, S. D., & Trimble, V. 2004, Stellar Interiors: Physical Principles, Structure, and Evolution (New York: Springer)
- Harrison, E. 1993, *ApJ*, 403, 28
- Hartmann, W. K. 2005, Moons and Planets (Belmont:Cole)
- Hawking, S. W. 1974, in IAU Symposium, Vol. 63, Confrontation of Cosmological Theories with Observational Data, ed. M. S. Longair, 283–286
- Henry, R. C. 1999, *ApJ*, 516, L49
- Hollenbach, D. J. & Tielens, A. G. G. M. 1999, *Reviews of Modern Physics*, 71, 173
- Holmberg, J., Nordström, B., & Andersen, J. 2007, *A&A*, 475, 519
- Hu, W. & Dodelson, S. 2002, *ARA&A*, 40, 171
- Hu, W., Sugiyama, N., & Silk, J. 1997, *Nature*, 386, 37
- Hubbard, W. B., Burrows, A., & Lunine, J. I. 2002, *ARA&A*, 40, 103
- Hubbard, W. B., Guillot, T., Lunine, J. I., Burrows, A., Saumon, D., Marley, M. S., & Freedman, R. S. 1997, *Physics of Plasmas*, 4, 2011
- Janes, K. A. & Phelps, R. L. 1994, *AJ*, 108, 1773
- Jarosik, N., Bennett, C. L., Dunkley, J., Gold, B., Greason, M. R., Halpern, M., Hill, R. S., Hinshaw, G., Kogut, A., Komatsu, E., Larson, D., Limon, M., Meyer, S. S., Nolta, M. R., Odegard, N., Page, L., Smith, K. M., Spergel, D. N., Tucker, G. S., Weiland, J. L., Wollack, E., & Wright, E. L. 2011, *ApJS*, 192, 14
- Jungman, G., Kamionkowski, M., Kosowsky, A., & Spergel, D. N. 1996, *PRD*, 54, 1332
- Karttunen, H., Kroger, P., Oja, H., Poutanen, M., & Donner, K. J. 2006, Fundamental Astronomy, 5th Edition (Berlin: Springer)
- Khedekar, S. & Majumdar, S. 2010, *Phys. Rev. D*, 82, 081301
- Kippenhahn, R. & Weigert, A. 1990, Stellar Structure and Evolution (New York: Springer)
- Kitayama, T., Susa, H., Umemura, M., & Ikeuchi, S. 2001, *MNRAS*, 326, 1353
- Kitchin, C. P. 2003, Astrophysical Techniques (London: Institute of Physics Publishing)
- Koval', V. V., Marsakov, V. A., & Borkova, T. V. 2009, *Astronomy Reports*, 53, 785
- Kroupa, P. 2001, *MNRAS*, 322, 231
- Larson, D., Dunkley, J., Hinshaw, G., Komatsu, E., Nolta, M. R., Bennett, C. L., Gold, B., Halpern, M., Hill, R. S., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Odegard, N., Page, L., Smith, K. M., Spergel, D. N., Tucker, G. S., Weiland, J. L., Wollack, E., & Wright, E. L. 2011, *ApJS*, 192
- Le Feuvre, M. & Wieczorek, M. A. 2011, *Icarus*, 214, 1
- Leinert, C., Bowyer, S., Haikala, L. K., Hanner, M. S., Hauser, M. G., Levasseur-Regourd, A.-C., Mann, I., Mattila, K., Reach, W. T., Schlosser, W., Staude, H. J., Toller, G. N., Weiland, J. L., Weinberg, J. L., & Witt, A. N. 1998, *A&AS*, 127, 1
- Longair, M. S. 2008, Galaxy Formation (Berlin: Springer)
- Lotz, J. M., Miller, B. W., & Ferguson, H. C. 2004, *ApJ*, 613, 262
- Marleau, F. 2010, Radiative Processes Lecture Notes - Atomic Physics
- Marsakov, V. A., Koval', V. V., Borkova, T. V., & Shapovalov, M. V. 2011, *Astronomy Reports*, 55, 667
- Matsuoka, Y., Ienaka, N., Kawara, K., & Oyabu, S. 2011, *ApJ*, 736, 119
- McQuinn, M., Oh, S. P., & Faucher-Giguère, C.-A. 2011, *ApJ*, 743, 82
- Miller, G. E. & Scalo, J. M. 1979, *ApJS*, 41, 513
- Montmerle, T. 1994, The cold universe : proceedings of the XXVIIth Rencontre de Moriond, XIIIth Moriond Astrophysics Meeting (France: Editions Frontieres)
- Murthy, J. 2009, *Ap&SS*, 320, 21
- Olive, K. A. 1991, *Science*, 251, 1194
- Osterbrock, D. E. & Ferland, G. J. 2006, Astrophysics of Gaseous Nebulae and Active Galactic Nuclei (University Science Books)
- Partridge, R. B. 2006, 3 K: The Cosmic Microwave Background Radiation (Cambridge: Cambridge University Press)
- Pasquini, L., Bonifacio, P., Randich, S., Galli, D., & Gratton, R. G. 2004, *A&A*, 426, 651
- Patanchon, G., Delabrouille, J., & Cardoso, J. . 2004, ArXiv Astrophysics e-prints
- Pedrotti, F. L., S., P. L., & M., P. L. 2007, Introduciton to Optics (New Jersey: Pearson Prentice Hall)
- Peebles, P. J. E. 1993, Principles of Physical Cosmology (Princeton: Princeton University Press)
- Percival, W. J., Nichol, R. C., Eisenstein, D. J., Frieman, J. A., Fukugita, M., Loveday, J., Pope, A. C., Schneider, D. P., Szalay, A. S., Tegmark, M., Vogeley, M. S., Weinberg, D. H., Zehavi, I., Bahcall, N. A., Brinkmann, J., Connolly, A. J., & Meiksin, A. 2007, *ApJ*, 657, 645
- Pravec, P. & Harris, A. W. 2000, *Icarus*, 148, 12
- Prochaska, J. X., Worseck, G., & O'Meara, J. M. 2009, *ApJ*, 705, L113
- Ribaudo, J., Lehner, N., & Howk, J. C. 2011, *ApJ*, 736, 42
- Rich, J. 2010, Fundamentals of Cosmology (Berlin:Springer)
- Riotto, A. & Trodden, M. 1999, Annual Review of Nuclear and Particle Science, 49, 35
- Rybicki, G. B. & P. L. A. 2004, Radiative Processes in Astrophysics (Germany: Wiley-VCH)
- Ryden, B. 2002, Introduction to Cosmology (San Franciso: Addison Wesley)
- Ryter, C. E. 1996, *Ap&SS*, 236, 285
- Salpeter, E. E. 1955, *ApJ*, 121, 161
- Sandage, A. 2010, *AJ*, 139, 728
- Sandage, A. & Perelmutter, J.-M. 1991, *ApJ*, 370, 455
- Santos, N. C. 2008, *NAR*, 52, 154
- Schmitt, H. R., Kinney, A. L., Calzetti, D., & Storchi Bergmann, T. 1997, *AJ*, 114, 592
- Schneider, P. 2002, Extragalactic Astronomy and Cosmology (Berlin:Springer)
- Schommer, R. A., Suntzeff, N. B., Olszewski, E. W., & Harris, H. C. 1992, *AJ*, 103, 447
- Seabroke, G. M. & Gilmore, G. 2007, *MNRAS*, 380, 1348
- Seager, S. 2011, Exoplanets
- Shore, S. N. 2003, The Tapestry of Modern Astrophysics (Hoboken:Wiley)
- Singal, J., Stawarz, Ł., Lawrence, A., & Petrosian, V. 2010, *MNRAS*, 409, 1172
- Smolin, L. 2004, ArXiv High Energy Physics - Theory e-prints
- Sparke, L. & Gallagher, J. 2007, Galaxies in the Universe: An Introduction (Cambridge: Cambridge University Press)
- Springel, V., White, S. D. M., Jenkins, A., Frenk, C. S., Yoshida, N., Gao, L., Navarro, J., Thacker, R., Croton, D., Helly, J., Peacock, J. A., Cole, S., Thomas, P., Couchman, H., Evrard, A., Colberg, J., & Pearce, F. 2005, *Nature*, 435, 629
- Tegmark, M. 2002, *Science*, 296, 1427
- Townsend, P. K. 1997, ArXiv General Relativity and Quantum Cosmology e-prints

- van den Bergh, S. 2008a, MNRAS, 385, L20  
 —. 2008b, MNRAS, 390, L51  
 van der Kruit, P. C. & Freeman, K. C. 2011, ARA&A, 49, 301  
 Walcher, J., Groves, B., Budavári, T., & Dale, D. 2011, Ap&SS, 331, 1  
 Wall, J. V. & Jenkins, C. R. 2003, Practical Statistics for Astronomers  
 (Cambridge: Cambridge University Press)  
 Walsh, K. J., Richardson, D. C., & Michel, P. 2008, Nature, 454, 188  
 Watkins, L. L., Evans, N. W., & An, J. H. 2010, MNRAS, 406, 264  
 Weinberg, M. D. 2000, ApJ, 532, 922  
 Will, C. M. 2006, Living Reviews in Relativity, 9  
 Wilms, J., Allen, A., & McCray, R. 2000, ApJ, 542, 914  
 Wizinowich, P. L., Le Mignant, D., Bouchez, A. H., Campbell, R. D., Chin, J. C. Y., Contos, A. R., van Dam, M. A., Hartman, S. K., Johansson, E. M., Lafon, R. E., Lewis, H., Stomski, P. J., Summers, D. M., Brown, C. G., Danforth, P. M., Max, C. E., & Pennington, D. M. 2006, PASP, 118, 297  
 Wyse, R. F. G. & Gilmore, G. 2005, ArXiv Astrophysics e-prints  
 Xiang, F. Y., Li, A., & Zhong, J. X. 2011, ApJ, 733, 91  
 Xue, X. X., Rix, H. W., Zhao, G., Re Fiorentin, P., Naab, T., Steinmetz, M., van den Bosch, F. C., Beers, T. C., Lee, Y. S., Bell, E. F., Rockosi, C., Yanny, B., Newberg, H., Wilhelm, R., Kang, X., Smith, M. C., & Schneider, D. P. 2008, ApJ, 684, 1143  
 Younger, J. D. & Hopkins, P. F. 2010, eprint arXiv:1003.4733