Astrophysics

Y. Ding
September 23, 2025

Contents

Name	Symbol	Value	Units
Speed of light in vacuum	c	2.997×10^8	$\mathrm{m\,s^{-1}}$
Gravitation constant	G	6.67×10^{-11}	$\mathrm{m^3\,kg^{-1}\,s^{-2}}$
Planck constant	$\hbar$	1.054×10^{-34}	$\mathrm{J\,s}$
		6.582×10^{-16}	$\mathrm{eV\,s}$
Boltzmann constant	k	1.38×10^{-23}	$\mathrm{J\,K^{-1}}$
		8.617×10^{-5}	$\mathrm{eV\,K^{-1}}$
Fine structure constant	α	$1/137$	
Stefan-Boltzmann constant	σ	5.67×10^{-8}	$\mathrm{W\,m^{-2}\,K^{-4}}$
Radiation constant	a	7.57×10^{-16}	$\mathrm{J\,K^{-4}\,m^{-3}}$
Mass of proton	m_p	1.67×10^{-27}	kg
		938.27	$\mathrm{MeV/c^2}$
Mass of electron	m_e	9.11×10^{-31}	kg
		0.511	$\mathrm{MeV/c^2}$

Table 0.1: Fundamental constants.

Not so fundamental constants:

Name	Symbol	Value	Units
Radius of the Sun	$R_\odot$	6.95×10^8	m
Mass of the Sun	$M_\odot$	1.998×10^{30}	kg
Surface luminosity of the Sun	$L_\odot$	3.85×10^{24}	W
Astronomical unit	AU	1.50×10^{11}	m
Light year	ly	9.46×10^{15}	m
Parsec	pc	3.09×10^{16}	m
Year	yr	3.16×10^7	s
Hubble constant	H_0	68 ± 2	$\mathrm{km\,s^{-1}\,Mpc^{-1}}$

Note: the units in this notes are sometimes written in the CGS system: length in cm, mass in g, and energy in erg $= 10^{-7}$ J.

Part I

Stellar Structure and Evolution

*Notation:

E/U: total energy/internal energy.
E_g: gravitational potential energy.
E_k: kinetic energy.
Q: heat.
P: pressure.
V: volume (or potential energy in quantum mechanics parts).
T: temperature.
F: energy flux.
L: luminosity.
N: particle number.
n: number density (or polytropic index in polytrope parts).
$\mathcal{P}$: probability
ϵ/u: energy density (or energy generation rate in nuclear reaction parts).
ρ: mass density.
$\bar{m}$: average mass.
ν: frequency.
a: radiation constant.
c_s: sound speed.
κ: opacity.
b: impact parameter.
τ: optical depth.
α: fine-structure constant or alpha particle, depending on the context.
μ: mean molecular weight or chemical potential, depending on the context.
σ: cross-section or Stefan-Boltzmann constant, depending on the context.
γ: adiabatic constant, photon, or Lorentz factor, depending on the context.

1 Introduction

Stars are self-gravitating objects that undergoes nuclear fusion. When a massive cloud start to collapse under its own gravity, its temperature, pressure, and density start to increase. Eventually, high enough temperature trigger the nuclear fusion process. A star has several basic properties:

- Pressure, temperature, and density are higher at the center.
- It is in hydrostatic equilibrium: pressure balances the gravity.
- It radiates energy, which comes from gravitational contraction and nuclear fusion.

Most stars undergo hydrogen fusion ("burning") as **main sequence stars**. Once they exhaust hydrogen in core, they evolve to become **red giants**. If they reach high enough temperatures, they will begin helium fusion, then carbon and so on, up to iron. Elements with higher atomic number are created by exploding massive stars, white dwarfs, dying low-mass stars, and merging neutron stars. Stars less then $8\,M_\odot$ will become a **white dwarf**. More massive stars will leave behind a **neutron star** or a **black hole** after exploding as a **supernova**.

1.1 Observation of Stars

Photons are continuously absorbed and emitted in the stellar interior, until they reach the **photosphere**, where we can start to see the lights. We cannot see into the stars except neutrinos from the Sun. Observed from earth, stars appear in different brightness/energy flux, color, and chemical composition. The color of a star tells about its temperature because it is roughly a **blackbody**. Carefully analyzing the spectrum of a star gives its chemical composition.

1.1.1 Color and Temeperature

We define the **effective temperature** T as the temperature of a perfect blackbody with the same luminosity L and radius R. The spectrum of a blackbody is described by the Planck function. The energy radiated per unit time per unit area per unit wavelength interval is

$$F_\lambda = \frac{2\pi hc^2}{\lambda^5}\frac{1}{e^{hc/\lambda kT}-1}. \tag{1.1}$$

The total power per unit area (or flux) is

$$P/A = \int_0^\infty F_\lambda\, d\lambda = \sigma T^4.$$

Stefan-Boltzmann law relates the effective temperature to the luminosity (the power radiated by a star),

$$L = 4\pi R^2 \sigma T^4, \tag{1.2}$$

where $\sigma \approx 5.67\times 10^{-5}\,\mathrm{erg\,cm^{-2}\,s^{-1}\,K^{-4}}$ is the Stefan-Boltzmann constant. The wavelength of maximum photons emitted is given by **Wien's displacement law** by taking the derivative of the Planck function:

$$\lambda_{\max} = \frac{0.3\,\mathrm{cm\,K}}{T} \tag{1.3}$$

This means a hotter star is bluer. For example, $\lambda_{\max}$ for the sun with $T \approx 5800\,\mathrm{K}$ is the green light wavelength. The star Spica with $T \approx 22400\,\mathrm{K}$ has $\lambda_{\max}$ in the ultraviolet range.

1.1.2 Luminosity, Brightness and Magnitude

The energy **flux**/**brightness** observed is related to luminosity by

$$F = \frac{L}{4\pi d^2} \tag{1.4}$$

where d is the distance from the observer to the star.

The **apparent magnitude** m is a simplified scale to measure the brightness of a star. A smaller magnitude gives a brighter star. Two stars that are 5 mangitudes apart differ in brightness by a factor of 100:

$$m_1 - m_2 = -2.5 \log_{10} \left(\frac{F_1}{F_2} \right). \tag{1.5}$$

Usually, the reference apparent magnitude is the star Vega with $m = 0$ (for historical reasons). The **absolute magnitude** M_{abs} of a star is its apparent magnitude if it was placed at a distance of 10 parsec. Since $F \propto 1/d^2$, the flux received at 10 pc away would be $(d/10\,\text{pc})^2$ smaller. The difference of a star's absolute magnitude from apparent magnitude is

$$M_{\text{abs}} - m = -2.5 \log_{10} \left(\frac{d}{10\,\text{pc}} \right)^2 = 5 - 5 \log_{10} \left(\frac{d}{\text{pc}} \right).$$

In terms of luminosity, the absolute magnitude is expressed as

$$M_{\text{abs}} = -2.5 \log_{10} \left(\frac{L}{L_0} \right) \quad \text{where} \quad L_0 = 85.4\, L_\odot.$$

The distance from the star can be determined by the parallax angle p. By small angle approximation,

$$\frac{p}{1''} = \frac{\text{pc}}{d}.$$

2 Equations of Stellar Structure

Assumptions:

1. Stars have spherical symmetry: rotations are neglected. Each physical quantity is a function of radius r: pressure $P(r)$, temperature $T(r)$, density $\rho(r)$, etc. Also, magnetic fields are neglected.
2. Stars are isolated such that their evolution depend on intrinsic properties like mass and composition.
3. Stars are made of homogeneous material.

2.1 Hydrostatic Equilibrium

2.1.1 Pressure

With spherical symmetry, a star can be sliced into shells of radius r and thickness dr with uniform properties like pressure, density, etc. The mass enclosed by a shell of radius r obeys

$$m(r) = \int_0^r 4\pi r'^2 \rho(r')\, dr' \quad \Longleftrightarrow \quad \frac{dm}{dr} = 4\pi r^2 \rho(r).$$

If the upper bound of the integral is the radius of the star R, then the integral turns out to be the total mass M,

$$M = \int_0^R 4\pi r^2 \rho\, dr.$$

Stars do not seem to be changing on human timescales. We already know that stars tends to collapse under gravity. Some estimation of this timescale can be done using Newton's law of universal gravitation and Newton's second law:

$$\frac{d^2 r}{dt^2} = -\frac{Gm(r)}{r^2}.$$

A rough estimation could be

$$\frac{d^2 r}{dt^2} \sim -\frac{R}{t_{\rm dyn}^2} \quad \Longrightarrow \quad t_{\rm dyn} \sim \sqrt{\frac{R^3}{GM}} \sim \frac{1}{\sqrt{G\bar{\rho}}},$$

where ρ is the average density of the star and $t_{\rm dyn}$ is called the **dynamic time** or the **free-fall time**. For the Sun, $t_{\rm dyn} \sim 1\,\text{hr}$. This time scale is short and humans could have just observed it. But the Sun is not collapsing, so there must be something preventing it from collapsing. This outward force is the pressure P. The force on an infinitesimal mass $dm = \rho\, dV$ by pressure is given by the **pressure gradient** ∇P. Meanwhile, the mass dm is pulled in by gravity $-\rho(r)g(r)\, dV$. Since the sun is not collapsing, we must have

$$\nabla P = -\rho g \quad \text{or} \quad \boxed{\frac{dP}{dr} = -\frac{Gm(r)}{r^2}\rho(r)} \tag{2.1}$$

in spherical coordinates (and spherical symmetry). This equation, relating pressure and density, is called the **equation of hydrostatic equilibrium**.

2.1.2 Some Rough Estimations

Astrophysics estimations is about getting the order-of-magnitudes—there will be a lot of $\sim$'s.

Central Pressure Invoke the hydrostatic equilibrium (2.1),

$$\frac{dP}{dr} = -\rho(r)\frac{Gm(r)}{r^2}.$$

The LHS is approximately

$$\frac{dP}{dr} \sim \frac{P(0) - P(R)}{0 - R} = -\frac{P}{R},$$

where we used the continuous boundary condition $P(R) = 0$. For the RHS of the equation, using $m(r) \sim M$, $r \sim R$, and $\rho(r) \sim M/R^3$, we have

$$-\frac{P}{R} \sim -\frac{M}{R^3}\frac{GM^2}{R^2} \quad \Longrightarrow \quad P \sim \frac{GM^2}{R^4}.$$

For the Sun, our estimation would give $P \sim 10^{15}\,\mathrm{dyne\,cm^{-2}}$, while $P_{\mathrm{true}} \approx 2 \times 10^{17}\,\mathrm{dyne\,cm^{-2}}$, not a bad estimate! The temperature of a star can be obtained from the ideal gas law

$$PV = NkT \tag{2.2}$$

In terms of the number density $n = N/V$, the pressure is

$$P = nkT = \frac{\rho}{\bar{m}}kT,$$

where $\bar{m}$ is the average mass of the particles inside a volume. Again, taking $\rho \sim M/R^3$,

$$P \sim \frac{M}{R^3}\frac{kT}{\bar{m}} \sim \frac{GM^2}{R^4} \quad \Longrightarrow \quad kT \sim \frac{GM\bar{m}}{R}.$$

The LHS is of the order of thermal energy per particle, while the RHS is of the order of gravitational potential energy per particle. In hydrostatic equilibrium, they are roughly equal. In the Sun, an ionized hydrogen has one electron and one proton, but most of the mass is concentrated in the proton. Thus, the mean particle mass is half the proton mass, $\bar{m} \approx 0.5\,m_p \sim 10^{-24}\,\mathrm{g}$. Plug this into the equation above gives $T \sim 10^7\,\mathrm{K}$, a very good estimate to $T_c \approx 1.5 \times 10^7\,\mathrm{K}$, the central temperature of the Sun. Note that the surface temperature of the sun is only about 6000 K.

2.1.3 Timescales of Stars

Dynamical/Free-Fall Timescale For an object already close to hydrostatic equilibrium, the dynamic timescale is the time for the object to readjust to slight perturbations in hydrostatic equilibrium. The dynamical timescale is also about equal to the time for sound waves to traverse the stellar surface. The adiabatic sound speed is give by

$$c_s = \sqrt{\gamma\frac{P}{\rho}},$$

where $\gamma = C_P/C_V$ is the adiabatic index ($\gamma = 5/3$ for a monatomic ideal gas). To show this, again assume $P \sim GM^2/R^4$ and $\rho \sim M/R^3$. The sound-crossing time is

$$t_{\mathrm{sound}} = \int_0^R \frac{dr}{c_s} \sim \frac{R}{c_s} \sim \frac{R}{\sqrt{P/\rho}} \sim R\sqrt{\frac{M/R^3}{GM^2/R^4}} = \sqrt{\frac{R^3}{GM}} \sim t_{\mathrm{dyn}}. \tag{2.3}$$

Indeed, oscillations within the Sun are to the order of 5 minutes, ($t_{\mathrm{dyn}} \approx 30$ minutes).

Kelvin-Helmholtz Timescale The next timescale is the Kelvin-Helmholtz timescale. Without nuclear fusion, the hydrostatic equilibrium relies only on gravitational potential energy. The **Kelvin-Helmholtz timescale** is roughly the time for a star to sustain current luminosity only by gravity,

$$t_{\mathrm{KH}} \sim \frac{|E_g|}{L} \sim \frac{GM^2}{RL}. \tag{2.4}$$

For the Sun, $t_{\mathrm{KH}} \approx 3 \times 10^7$ yr. One example of Kelvin-Helmholtz contraction is during pre-main sequence phase, where the star has obtained hydrostatic equilibrium, but not with enough temperature to start nuclear fusion. It contracts in a Kelvin-Helmholtz timescale and gain some thermal energy from gravitational ones. Before nuclear fusion was found, people thought that gravitational energy was powering the sun. However, the Kelvin-Helmholtz timescale was much shorter than the age of the Sun (deduced from rock formation time on Earth), so there must be other mechanism that powers the Sun.

Nuclear Timescale The **nuclear timescale** is the timescale on which the core of a star fuses and on which the structure of a star will change. A typical hydrogen fusion process is (there are some intermediate steps but we will ignore them here):

$$4\,^1\mathrm{H} + 4e^- \longrightarrow \,^4\mathrm{He} + 2e^-.$$

It is known that $m_p + m_e = 1.0078\,\mathrm{u}$, four copies of them is $4.0312\,\mathrm{u}$, and $m_{\mathrm{He}} + 2m_e = 4.0026\,\mathrm{u}$. The overall mass after the reaction is $\eta \approx 0.7\%$ less. The total rest mass energy of a star is Mc^2, but only $f \sim 10\%$ of it releases as hydrogen fusion before the core runs out of fuel, i.e. during the main sequence. From this we can calculate the nuclear timescale for the Sun:

$$t_{\mathrm{nuc}} \sim \frac{\eta f M c^2}{L} \approx 10^{10}\,\mathrm{yr}.$$

The Sun's age now is about $\frac{1}{2}t_{\mathrm{nuc}}$. Most main sequence stars obeys the mass-luminosity relation $L \propto M^{3.5}$. Hence the nuclear timescale has relation to mass as

$$t_{\mathrm{nuc}} \propto \frac{M}{M^{3.5}} = M^{-2.5}. \tag{2.5}$$

Massive stars have shorter lifetimes. In a star cluster, the most massive stars will leave the main sequence before the least massive stars begin main sequence evolution.

2.2 Virial Theorem

Consider a shell of mass dm and radius r from the center of the star. The gravitational potential energy of this shell is

$$dE_g = -\frac{Gm(r)dm}{r} = -\frac{Gm(r)}{r}(4\pi r^2 \rho\, dr) = -\frac{Gm(r)}{r^2}(4\pi r^3) = 4\pi r^3 \frac{dP}{dr}dr,$$

where in the last equality hydrostatic equilibrium (2.1) is used. The total gravitational potential energy is

$$E_g = \int_0^R dE_g = \int_0^R 4\pi r^3 \frac{dP}{dr}dr.$$

Integrating by parts,

$$E_g = \cancel{4\pi r^2 P(r)\Big|_0^R} - 3\int_0^R 4\pi r^2 P(r)dr = -3\int P\,dV = -3\langle P\rangle V.$$

The first term vanishes because $P(R) = 0$ at the stellar surface. Rearranging some terms, we get

$$\langle P\rangle = -\frac{1}{3}\frac{E_g}{V}. \tag{2.6}$$

This is called the **virial theorem** for a self-gravitating system. Recall that the typical pressure inside a star is $P \sim GM^2/R^4$. Let the total gravitational potential energy be $E_g = -\alpha GM^2/R$ where α is some factor of order unity ($\alpha = 3/5$ for a uniform density sphere). Plug this into the virial theorem and the result is

$$\langle P\rangle = \frac{\alpha}{4\pi}\frac{GM^2}{R^4},$$

the same order as the typical pressure.

2.2.1 Pressure of an Ideal Gas

In thermodynamics, it can be proved (using Newton's second law and particles in a box classically) that the pressure for an ideal gas is

$$P = \frac{1}{3}n\langle \mathbf{p}\cdot\mathbf{v}\rangle = \frac{1}{3}n\langle pv\rangle,$$

where $\mathbf{p}$ is the momentum of the particle and $\mathbf{v}$ is its velocity. This relation is valid for both relativistic and non-relativistic particles. we are trying to relate pressure to the translational kinetic energy of both type of particles inside a star. We will use the relativistic energy of a particle

$$E = \sqrt{p^2c^2 + m^2c^4}.$$

Note that p here is the relativistic momentum γmv (γ is the Lorentz factor).

If the particle is non-relativistic ($v \ll c$), then the energy of a particle is just its rest mass energy plus the ordinary kinetic energy,

$$E_{\rm nr} = mc^2 + \frac{p^2}{2m} + \mathcal{O}(v^2/c^2).$$

If the particle is relativistic ($v \approx c$), then

$$pc \gg mc^2 \quad \text{and} \quad E_{\rm r} \approx pc \approx pv.$$

In the two cases above, a particles kinetic energy differs by a factor of 2:

$$E_{k,\rm nr} = \frac{p^2}{2m} = \frac{1}{2}pv, \quad \text{while} \quad E_{k,\rm r} = pv.$$

Multiplying the number density n by $\langle pv \rangle$ gives the *(kinetic) energy density* ϵ_k:

$$P = \begin{cases} \dfrac{2}{3}\epsilon_k, & \text{[non-relativistic]}, \\ \dfrac{1}{3}\epsilon_k, & \text{[relativistic]}. \end{cases} \tag{2.7}$$

2.2.2 Virial Theorem for Non-Relativistic Ideal Gas

Like volume average of potential energy, once the pressure and kinetic energy density relation is known, we have

$$\langle P \rangle = \frac{2}{3}\frac{E_k}{V},$$

where E_k is the total kinetic energy of the star. Then we can easily obtain

$$2E_k = -E_g \quad \Longrightarrow \quad E_{\rm tot} = \frac{1}{2}E_g.$$

For non-relativistic stars, they are almost composed of ionized hydrogen and helium, so the particle number N contains all of the electrons and atomic nuclei. By the ideal gas law (2.2),

$$P = nkT = \frac{\rho}{\bar{m}}kT.$$

In statistical mechanics, the equipartition theorem says that the average translational kinetic energy of a particle in an ideal gas is $\frac{3}{2}kT$, so

$$P = \frac{2}{3}\left(\frac{3}{2}nkT\right) = \frac{2}{3}\epsilon_k,$$

which agrees with the result for non-relativistic gas.

Stars are obviously hotter than surroundings, so they will spontaneously give up energy (mainly through radiation). The rate of losing energy by radiation is given by the luminosity L, while the nuclear fusion will supply the energy by $L_{\rm nuc}$. Hence the total rate of change of energy is $\dot{E}_{\rm tot} = L_{\rm nuc} - L$. If there is no nuclear fusion, then $\dot{E}_{\rm tot} < 0$, which means E_g is decreasing and E_k is increasing. That is, the star contracts and heats up. One example of such stars is pre-main sequence stars: gravitational potential energy converts into kinetic energy and increases the temperature until high enough for nuclear fusion. From the above analysis, we say that a star has a negative heat capacity.

For main sequence stars, $L_{\rm nuc} \approx L$. there will be oscillations in temperature if there is some perturbation. For example, let's say the perturbation raises the nuclear fusion rate ($\dot{E}_{\rm tot} > 0$). The star will expand because $\dot{E}_g > 0$ and the temperature lowers. Since nuclear fusion rate is strongly correlated with temperature (discussed in later sections), it will come back down to a normal rate.

Note that this type of analysis does not apply to relativistic stars with either strong gravity or strong radiation pressure. White dwarfs (with electron degeneracy pressure) and neutron stars (neutron degeneracy pressure) belong to the first category. High temperature stars where photon pressure is important (photons are definitely relativistic) belong to the second category.

2.2.3 Virial Theorem for Relativistic Gas

If we combine the virial theorem with pressure of relativistic gas, we find that

$$E_k = -E_g \quad \Longrightarrow \quad E_{\text{tot}} = 0.$$

This means yo can change the star's size without adding or removing any energy. This happens for very massive stars which is extremely unstable. It is hard for such stars to reach hydrostatic equilibrium. More likely, they will either expand to infinity, or undergo a collapse.

2.2.4 Virial Theorem for a Generalized Equation of State

In general, the pressure is related to internal energy density or kinetic energy density as

$$P = (\gamma - 1)\epsilon_k.$$

Again γ is the adiabatic index, the same in adiabatic process in thermodynamics ($PV^\gamma = \text{const.}$). For a non-relativistic gas, $\gamma = 5/3$, and for a relativistic one, $\gamma = 4/3$.

2.3 Equation of States

We already know that pressure is needed to balance gravity in order to achieve stars' hydrostatic equilibrium. Now we are to study where the pressure comes from: to derive the equation of state. The **equation of state** of a star relates pressure to density, temperature, chemical composition, etc, for a ideal gas

$$P = P(\rho, T, \{X_i\}).$$

Here the "ideal gas" has a different meaning from the classical ideal gas. In stars, they refer to non-interacting particles since they can almost never be composed of molecules. Instead, most can be composed of electrons, protons, or photons, but they behave differently.

2.3.1 Deriving Pressure

Recall that the pressure is related to momentum and pressure of an ideal gas by

$$P = \frac{1}{3}n\langle \mathbf{p} \cdot \mathbf{v} \rangle = \frac{1}{3}n\langle pv \rangle.$$

First let us define a few terms. We will consider the momentum space of particles. The total number of states occupied is

$$n = \int_0^\infty n(p)\,dp,$$

where $n(p)$ is the number density of particles between momentum $(p, p+dp)$. The average of momentum and velocity $\langle pv \rangle$ is then

$$\langle pv \rangle = \frac{\int_0^\infty pvn(p)\,dp}{\int_0^\infty n(p)\,dp} = \frac{1}{n}\int_0^\infty pvn(p)\,dp.$$

Thus the pressure is an integral over momentum space (magnitude):

$$P = \frac{1}{3}\int_0^\infty pv(p)n(p)\,dp.$$

Here the speed v is also a function of p, $v \approx p/m$ for non-relativistic particles and $v \approx c$ for relativistic ones. According to statistical mechanics, in order to know the number density of a gas, we need to know:

- The **density of states** between momentum $(p, p+dp)$: $g(p)\,dp$. This indicates how many possible states are there for a particle with momentum p to occupy.

- The **occupation number**: $f(p)$. Not all of the states will be occupied, so the occupation number gives the fraction of occupation.
- The **spin degeneracy**: g_s, $g_s = 2$ for electrons, protons, neutrons, photons, etc. For neutrinos, $g_s = 1$. If a particle has spin, then g_s indicates how many identical particles can occupy the same spatial state. For example, electrons can have spin $1/2$ and $-1/2$, so $g_s = 2$. g_s can also indicates number of polarization states for photons.

Given some momentum distribution, the pressure in terms of all quantities defined above is

$$P = \frac{1}{3}\int_0^\infty p v g_s g(p) f(p)\, dp.$$

Specifically, the density of state function is (see derivation in Appendix A.2)

$$g(p)\, dp = \frac{4\pi p^2 dp}{h^3},$$

The occupation number is given by (see derivation in Appendix A.3)

$$f(p) = \frac{1}{e^{[\epsilon(p)-\mu]/kT} \pm 1},$$

Notice that there is a ± 1 in the occupation number. The $+1$ is for fermions (e^-, p^+, n, ν_e, etc., particles obeying Fermi-Dirac distribution) and -1 is for bosons (particles obeying Bose-Einstein distribution). In most cases, the only boson we will encounter is the photon. This ± 1 term is unimportant in a classical limit of ideal gas, so for particles obeying Boltzmann distribution,

$$f(p) = e^{-[\epsilon(p)-\mu]/kT}.$$

Classical Ideal Gas Now let's calculate the number density in the classical limit. Plug in the density of states $g(p)\, dp$ and the occupation number $f(p)$ into n,

$$n = \int g_s g(p) f_{\text{class}}(p)\, dp = \int_0^\infty g_s \frac{4\pi p^2}{h^3} e^{-[\epsilon(p)-\mu]/kT} dp = \frac{4\pi g_s}{h^3} e^{\mu/kT} \int_0^\infty p^2 e^{-\epsilon(p)/kT} dp.$$

In the non-relativistic limit, $\epsilon(p) = mc^2 + p^2/2m$,

$$n = g_s \frac{4\pi}{h^3} e^{(\mu - mc^2)/kT} \int_0^\infty p^2 e^{-p^2/2mkT} dp.$$

The integral is just a simple Gaussian integral, the answer is

$$n = g_s \left(\frac{2\pi mkT}{h^2}\right)^{3/2} e^{(\mu - mc^2)/kT}.$$

The chemical potential μ can be written as

$$\mu = mc^2 + kT \ln\left(\frac{n}{g_s n_Q}\right).$$

The term in the bracket (with the power of $3/2$) is often called the **quantum concentration**, n_Q. It has the unit of number density. Once the number density is close to the quantum concentration, the ± 1 term in the occupation number can no longer be ignored. Hence the classical limit means $n \ll n_Q$. The mean particle separation $\sim n^{-1/3}$ should be much greater than the thermal de Broglie wavelength $\lambda_{\text{dB}} = h/\sqrt{2mkT} \sim n_Q^{-1/3}$ of a particle with energy kT. Note that the quantum concentration is related to mass by $n_Q \propto m^{3/2}$. Since $m_p \approx 10^3 m_e$, $n_{Q,p} \approx 10^5 n_{Q,e}$, protons would require 10^5 times denser than electrons for quantum effects to happen. In stars, neutrons and protons are always classical, unless they are in a neutron star. However, if the number density of electrons reaches $n_e \approx 8 \times 10^{31}\ \text{m}^{-3}$, or equivalent mass density $\rho = 1.2 \times 10^5\ \text{kg/m}^3$, quantum effects matter for electrons.

In the relativistic limit $v \approx c$, one can show that

$$n_Q = 8\pi \left(\frac{kT}{hc}\right)^3, \quad \mu = kT \ln\left(\frac{n}{g_s n_Q}\right), \quad \text{and} \quad \lambda_{\text{dB}} = \frac{hc}{kT}.$$

The pressure integral with Boltzmann distribution can also be calculated easily like the number density integral. Writing μ in terms of the number density n, we achieve the ideal gas law (see Appendix A.5)

$$P = nkT = \frac{\rho}{\bar{m}}kT.$$

In conclusion, the classical limit applies when the particle number density is much less than the quantum concentration, $n \ll n_Q$. The mean particle separation is much greater than the thermal de Broglie wavelength, $n^{-1/3} \gg \lambda_{\mathrm{dB}}$. The ideal gas law is the equation of states for both relativistic and non-relativistic gas.

2.3.2 Ions and Electrons in the Ideal Gas Law

Most stars are composed of ions and electrons. Assuming that the star is fully ionized, we can divide the total pressure into two parts, one from ions and the other from electrons:

$$P = P_{\mathrm{I}} + P_e = n_{\mathrm{I}}kT + n_e kT,$$

where n_{I} and n_e are the ions and electron number densities, respectively.

In a star, the ith ion is characterized by the mass fraction X_i, the atomic mass A_i, and the proton number (or the charge) Z_i. The number density of ions is given by the sum

$$n_{\mathrm{I}} = \sum_i n_i = \sum \frac{X_i}{A_i m_p}\rho.$$

We define the **total mean molecular weight** of ions, μ_{I}:

$$\frac{1}{\mu_{\mathrm{I}}} = \sum_i \frac{X_i}{A_i} \approx X + \frac{Y}{4},$$

where X is the hydrogen ion mass fraction and Y is the helium ion mass fraction, and we ignore heavier elements. The Sun has $\mu_{\mathrm{I}} \approx 1.3$. Similarly, the electron number density is given by the sum

$$n_e = \sum_i n_i = \sum_i \frac{Z_i X_i}{A_i m_p}\rho \equiv \frac{\rho}{\mu_e m_p} \quad \text{where} \quad \frac{1}{\mu_e} = \sum_i \frac{X_i Z_i}{A_i} \approx X + \frac{Y}{2}.$$

The total mean molecular weight is defined as

$$\frac{1}{\mu} = \frac{1}{\mu_{\mathrm{I}}} + \frac{1}{\mu_e},$$

which gives the ideal gas law

$$P = \frac{\rho}{\mu m_p}kT.$$

Physically, the mean molecular weight is how masses are distributed by particles. For pure hydrogen atom, there is one molecule with atomic mass 1, so its mean molecular weight is just 1. If the hydrogen is ionized, there are two particles (one nucleus and one electron) distributing the mass, so $\mu = 1/2$. Likewise, for helium atom, its mean molecular weight is 4. If it is ionized, there are three particles (one nucleus and two electrons) distributing the mass, $\mu = 4/3$.

2.3.3 Degeneracy Pressure

We shall look at electron gas for now. The following results also work for other fermions. We first assume that the temperature is at $T = 0$. At absolute zero, the Fermi-Dirac distribution becomes a step function:

$$\bar{n}_{\mathrm{FD}} = \frac{1}{e^{(\epsilon-\mu)/kT}+1} \quad \longrightarrow \quad \bar{n} = \begin{cases} 1, & \epsilon < \mu \\ 0, & \epsilon > \mu. \end{cases}$$

This means the low-energy states are always filled while the high-energy states are always empty. The maximum momentum a fermion can have is called the **Fermi momentum**, p_{F}. (There is also a derivation using **Fermi energy** in Appendix A.6. The ideas are essentially the same.)

We can compute the number density as

$$n = \frac{4\pi g_s}{h^3}\int_0^\infty p^2 f(p)\,dp = \frac{4\pi g_s}{h^3}\int_0^{p_\mathrm{F}} p^2\,dp = \frac{4\pi g_s}{3h^3}p_\mathrm{F}^3.$$

Electrons have $g_s = 2$, so the Fermi momentum is

$$p_\mathrm{F} = \left(\frac{3n_e}{8\pi}\right)^{1/3} h,$$

where n_e is the number density of the electron gas.

To visualize this result, suppose now there are N electrons in a box of volume V. By the uncertainty principle, each h^3 box of a six-dimensional (three position and three momentum) *phase space* can have two electrons (with opposite spins), so N electrons occupy a volume of $Nh^3/2$ in the phase space. Meanwhile, these electrons will fill up the momentum space volume with $p < p_\mathrm{F}$, so

$$V_\mathrm{phase} = V_p V = \frac{4\pi}{3}p_\mathrm{F}^3 V.$$

Setting this to $Nh^3/2$,

$$\frac{4\pi}{3}p_\mathrm{F}^3 V = \frac{Nh^3}{2} \quad\Longrightarrow\quad p_\mathrm{F}^3 = \frac{N}{V}\frac{3h^3}{8\pi} = \frac{3h^3}{8\pi}n_e.$$

Even though those electrons are at zero temperature, they still have some momenta (and hence some energy). Moreover, if the electrons are denser, they get more momenta. In other words, they are forced to occupy higher momentum (energy) states. Such force is called the **degeneracy pressure**.

We know the pressure is given by

$$P = \frac{4\pi g_s}{3h^3}\int_0^\infty p^3 v(p) f(p)\,dp.$$

In the zero temperature and non-relativistic limit $v = p/m$, and for electrons with $g_s = 2$, this integral follows

$$P_{nr} = \frac{8\pi}{3m_e h^3}\int_0^{p_\mathrm{F}} p^4\,dp = \frac{8\pi}{15 m_e h^3}p_\mathrm{F}^5 = \frac{h^2}{5m_e}\left(\frac{3}{8\pi}\right)^{2/3} n_e^{5/3}.$$

In terms of mean molecular weight per electron,

$$n_e = \frac{\rho}{\mu_e m_p}, \quad P_{nr} = \frac{h^2}{5m_e}\left(\frac{3}{8\pi}\right)^{2/3}\left(\frac{\rho}{\mu_e m_p}\right)^{5/3}. \tag{2.8}$$

In the relativistic limit $v \approx c$,

$$P_r = \frac{8\pi c}{3h^3}\int_0^{p_\mathrm{F}} p^3\,dp = \frac{2\pi c}{3h^3}p_\mathrm{F}^4 = \frac{hc}{4}\left(\frac{3}{8\pi}\right)^{1/3} n_e^{4/3} = \frac{hc}{4}\left(\frac{3}{8\pi}\right)^{1/3}\left(\frac{\rho}{\mu_e m_p}\right)^{4/3}. \tag{2.9}$$

These expressions are in fully degenerate limit (zero-temperature), but they give good approximation to finite-temperature conditions. (For more information, see Appendix A.6.1, elaborated through the energy derivation.) Also note the powers of n_e in the relativistic and non-relativistc degeneracy pressure, $4/3$ and $5/3$ respectively. They are the adiabatic indices in the two cases.

The following figure shows the total pressure as a function of mass density and temperature.

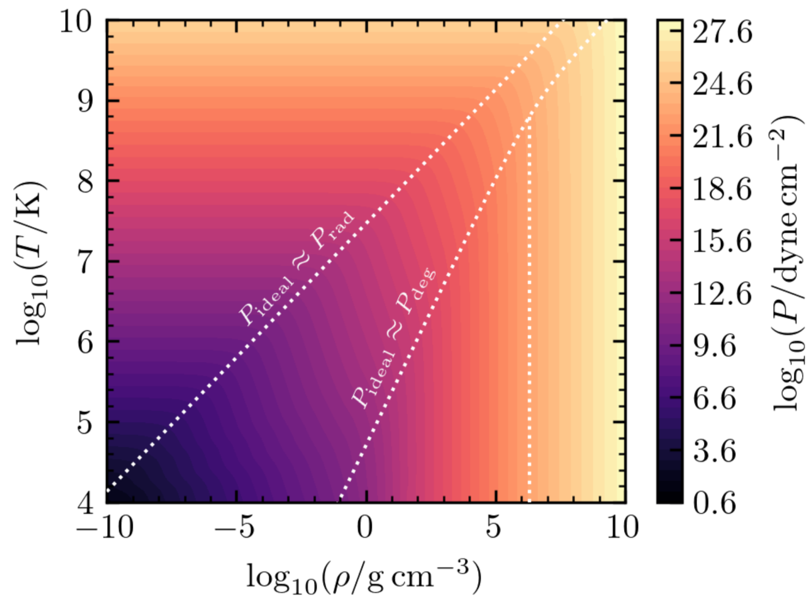


Figure 2.1: The total pressure of a star as a function of mass density ρ and temperature T. The composition is pure helium. There are several areas in this graph. Below the $P_{\rm ideal} \approx P_{\rm deg}$ line, electron degeneracy pressure dominates. The pressure in this region only depend on ρ but not T. This is because the degeneracy pressure is a strong function of ρ until T is high enough to recover the classical limit (the middle area). The vertical line in this area shows the relativistic threshold. To the right, the electron gas becomes more relativistic. The top left corner above the $P_{\rm ideal} \approx P_{\rm rad}$ line is where radiation pressure dominates.

When do degeneracy and relativity matter? Electrons are almost degenerate when the kinetic energy corresponding to the Fermi momentum is large compared to the average thermal energy $\frac{3}{2}kT$. In the non-relativistic limit, it means

$$\frac{p_{\rm F}^2}{2m_e} \gg \frac{3}{2}kT \quad \Longrightarrow \quad n_e^{2/3} \gg \frac{m_e kT}{h^2},$$

or simply $n_e \gg n_Q$ where $n_Q = (2\pi m_e kT/h^2)^{3/2}$. In the relativistic limit similarly,

$$p_{\rm F}c \gg \frac{3}{2}kT \quad \Longrightarrow \quad n_e^{1/3} \gg \frac{kT}{hc},$$

or equivalently $n_e \gg n_Q$ where $n_Q = 8\pi(kT/hc)^3$.

Electrons become relativistic when their energy are comparable to their rest-mass energy. Without degeneracy, it means $kT \approx m_e c^2$, which gives a temperature of $T \sim 10^9$ K. When there is degeneracy, it means $p_{\rm F} = m_e c$, which gives a number density of $n \sim 10^{35}\,\mathrm{m^{-3}}$. For $\mu_e = 2$ like ionized helium, carbon or oxygen, this number density is equivalent to a mass density $\rho \sim 10^9\,\mathrm{kg/m^3}$.

Example 2.1. White dwarf mass-radius relation

White dwarfs are stellar remnants supported by electron degeneracy pressure. By the virial theorem, the average pressure is about

$$P = -\frac{1}{3}\frac{E_g}{V} \sim \frac{GM^2/R}{R^3} = \frac{GM^2}{R^4}.$$

Assuming non-relativistic degeneracy, its equation of state says

$$P \propto \rho^{5/3} \propto \frac{M^{5/3}}{R^5}.$$

Setting these two pressure relations equal gives

$$R \propto M^{-1/3}.$$

A more massive white dwarf is smaller. This results in a density increase, and relativistic effects become more and more important. In the relativistic limit,

$$P \propto \rho^{4/3} \propto \frac{M^{4/3}}{R^4}.$$

The virial theorem and relativistic degeneracy pressure both have $1/R^4$, so they give $M = \text{const}$. It turns out that there is a upper limit of mass a white dwarf can have. This limit is called the **Chandrasekhar mass**, about $M_{\text{Ch}} \approx 1.4\, M_\odot$. As we already know that the virial theorem says a relativistic star has zero total energy. Any perturbation would destroy the star.

Indeed, just before a white dwarf reaches the Chandrasekhar limit, nuclear reactions are triggered at its center because of the high density. A carbon-oxygen white dwarf would have carbon ignition at the center, resulting in a type Ia supernova. An oxygen-neon white dwarf would either collapse and form a neutron star, or also undergo a supernova.

2.3.4 Radiation Pressure

Photons are bosons with zero chemical potential because they can be easily created or destroyed (e.g. change in energy level of an electron absorbs or create a photon). They obey the Bose-Einstein distribution, which reduces to the Planck distribution with $\mu = 0$,

$$f(p) = \frac{1}{e^{\epsilon/kT} - 1}.$$

The number density is

$$n = \frac{4\pi g_s}{h^3} \int_0^\infty p^2 f(p)\, dp = \frac{8\pi}{h^3} \int_0^\infty \frac{p^2\, dp}{e^{pc/kT} - 1}.$$

For photons, there are two polarizations ($g_s = 2$) and they are relativistic ($\epsilon = pc$). To evaluate the integral, let $x = pc/kT$, and then

$$n = \frac{8\pi}{h^3} \left(\frac{kT}{c}\right)^3 \int_0^\infty \frac{x^2}{e^x - 1}\, dx = 8\pi \left(\frac{kT}{hc}\right)^3 (2\zeta(3))$$

where $\zeta(z)$ is the Riemann-zeta function,

$$\zeta(z) = \frac{1}{1^z} + \frac{1}{2^z} + \frac{1}{3^z} + \cdots = \frac{1}{\Gamma(z)} \int_0^\infty \frac{x^{z-1}}{e^z - 1}\, dx, \quad \Gamma(z) = (z-1)!.$$

We find that the number density is proportional to T^3,

$$n = 16\pi \left(\frac{k}{hc}\right)^3 \zeta(3) T^3 \equiv bT^3, \tag{2.10}$$

where $b = 2.07 \times 10^7\, \text{K}^{-3}\text{m}^{-3}$. The pressure integral is similar,

$$P = \frac{1}{3} \frac{4\pi g_s}{h^3} \int_0^\infty p^3 v(p) f(p)\, dp = \frac{1}{3} \frac{8\pi c}{h^3} \int_0^\infty \frac{p^3\, dp}{e^{pc/kT} - 1}.$$

Do the same substitution $x = pc/kT$, we have

$$P = \frac{1}{3} \frac{8\pi c}{h^3} \left(\frac{kT}{c}\right)^4 \int_0^\infty \frac{x^3}{e^x - 1}\, dx = \frac{1}{3} \frac{8\pi k^4}{h^3 c^3} T^4 (6\zeta(4)) \equiv \frac{1}{3} a T^4. \tag{2.11}$$

For Riemann-zeta function with even arguments, there are exact values, $6\zeta(4) = \pi^4/15$. The constant $a \equiv 8\pi^5 k^4 / 15 h^3 c^3 \approx 7.57 \times 10^{-16} \text{J} \cdot \text{K}^{-4}\text{m}^{-3}$ is called the **radiation constant**. Since $P = \frac{1}{3}u$, the energy density u is (according to equation 2.7):

$$u = aT^4. \tag{2.12}$$

The mean photon energy is then

$$\langle \epsilon_\gamma \rangle = \frac{u}{n} = \frac{aT^4}{bT^3} = 2.7kT.$$

A hotter blackbody emits photons with shorter wavelengths. We can also derive the rate of energy emitted by a blackbody, namely the **Stefan-Boltzmann law**. It should be proportional to both T^4 and c. It turns out that (see Appendix A.7.3 for proof) the is an additional 1/4 correction factor,

$$F = \frac{1}{4}uc = \frac{ac}{4}T^4 = \sigma T^4,$$

where $\sigma \approx 5.67 \times 10^{-8}\,\mathrm{J\,s^{-1}m^{-2}K^{-4}}$.

Massive Stars Recall that the virial theorem for a classical ideal gas says

$$E_g = -2E_k.$$

We estimate that the total number of particles is $N \sim M/\mu m_p$ where μ is the mean molecular mass. Assume that kT is the internal energy of each particle, so $E_k \sim MkT/\mu m_p$, and

$$-\frac{GM^2}{R} \sim \frac{MkT}{\mu m_p} \quad \Longrightarrow \quad kT \sim \frac{GM\mu m_p}{R}.$$

Compare the radiation pressure and the ideal gas pressure from the electrons and ions, we find that

$$\frac{P_{\mathrm{rad}}}{P_{\mathrm{gas}}} = \frac{aT^4/3}{\rho kT/\mu m_p} \propto \frac{T^3}{\rho} \propto \frac{(M/R)^3}{M/R^3} \propto M^2.$$

As shown in Figure 2.1, the radiation pressure dominates in low density and high temperature massive stars. Also, the total pressure starts to be independent of density in this region.

2.3.5 Ionization: Saha equation

To discuss ionization inside a star, we make a few assumptions:

1. The star has low density/high temperature so that electrons and ions are classical. In other words, we rule out the degeneracy. Also, if electrons are relativistic, *everything* are ionized anyway so there is nothing interesting about ionization. Thus, ionization does not apply to relativistic stars either.

2. The star is assumed to be in local thermal equilibrium (constant temperature across a thermal distribution) and in ionization/chemical equilibrium. For example, consider the ionization equation

$$p + e^- \;\longleftrightarrow\; \mathrm{H} + \gamma.$$

   The rates of forward and backward reaction are the same. In chemical equilibrium, the chemical potential on each side are equal, $\mu_p + \mu_e = \mu_{\mathrm{H}} + \mu_\gamma$. Since $\mu_\gamma = 0$, the equilibrium condition is $\mu_p + \mu_e = \mu_{\mathrm{H}}$.

In classical non-relativistic limit, we know that

$$\mu = mc^2 + kT\ln\left(\frac{n}{gn_Q}\right), \quad n_Q = \left(\frac{2\pi mkT}{h^2}\right)^{3/2}.$$

Here g_s is changed to g because if an atom is not ionized, we have to take into account the degeneracy of electron orbital states. A little algebra will give

$$m_{\mathrm{H}}c^2 - m_pc^2 - m_ec^2 = kT\ln\left(\frac{n_en_p}{n_{\mathrm{H}}}\frac{g_{\mathrm{H}}}{g_eg_p}\frac{n_{Q,\mathrm{H}}}{n_{Q,e}n_{Q,p}}\right).$$

If the electron is in the ground state (when the temperature is low), $g_{\mathrm{H}} = 4$, $g_e = g_p = 2$. Moreover, since $m_{\mathrm{H}} \approx m_p$, we shall assume that $n_{Q,p} \approx n_{Q,\mathrm{H}}$. The binding energy of hydrogen is $-13.6\,\mathrm{eV}$, so we define

$$m_{\mathrm{H}}c^2 - m_pc^2 - m_ec^2 = -13.6\,\mathrm{eV} \equiv -\chi.$$

Exponentiating this term and plug in $g_p = g_e = 2$ and $g_{\mathrm{H}} = 4$ in the ground state,

$$e^{-\chi/kT} = \frac{n_en_p}{n_{\mathrm{H}}}\frac{g_{\mathrm{H}}}{g_eg_p}\frac{1}{n_{Q,e}} \quad \Longrightarrow \quad \frac{n_en_p}{n_{\mathrm{H}}} = \left(\frac{2\pi m_ekT}{h^2}\right)^{3/2} e^{-\chi/kT}. \tag{2.13}$$

This is the **Saha equation** for hydrogen in the ground state. If the proton and the hydrogen atom are not in the ground state, then its Saha equation is

$$\frac{n_e n_p}{n_{\mathrm{H}}} = \frac{2g_p}{g_{\mathrm{H}}} \left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} e^{-\chi/kT}. \tag{2.14}$$

In general, there are more than one ionization states for atoms with higher atomic number. For example, helium has first ionization state (lose one e^-) and second ionization state (lose two/all e^-s). For transition between ionization states i and $i+1$, the Saha equation is

$$\frac{n_e n_{i+1}}{n_i} \approx \frac{2g_{i+1}}{g_i} \left(\frac{2\pi mkT}{h^2}\right)^{3/2} e^{-\chi_{i\to i+1}/kT}.$$

Here the indices indicate the ionization states of an atom, and $\chi_{i\to i+1}$ is the transition energy difference. Saha equation also applies to **nuclear statistical equilibriums**, e.g.

$$\gamma + {}^{20}\mathrm{Ne} \longleftrightarrow {}^{16}\mathrm{O} + {}^{4}\mathrm{He}.$$

These reactions occur at very high temperatures, and the energy difference is much higher ($\sim$ MeV). They play important roles at the end state in the core of massive stars.

In the outer layers of stars, there can be a **partial ionization zone**. We shall stick with pure hydrogen ionization first since it is the simplest one. If the ionization fraction is x,

$$x = \frac{n_p}{n_p + n_{\mathrm{H}}} = \frac{n_e}{n_e + n_{\mathrm{H}}},$$

we will have the number density relation

$$n_{\mathrm{H}} + n_p = \frac{\rho}{m_{\mathrm{H}}} \quad \text{or} \quad n_e = n_p = x\frac{\rho}{m_{\mathrm{H}}}.$$

Then the Saha equation says that

$$\frac{x^2}{1-x}\frac{\rho}{m_{\mathrm{H}}} = \left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} e^{-\chi/kT}. \tag{2.15}$$

This gives the ionization fraction x as a function of temperature and density, which can be solved just by the quadratic formula. One would expect that an atom would ionize at a temperature where $kT \sim \chi$, or $T \sim 10^5\,\mathrm{K}$ for hydrogen. However, for x around 50%, the temperature indicated by the Saha equation is only about $T \sim 10^4\,\mathrm{K}$. Similar for other atoms, the ionization temperature is around $kT \sim 10^{-1}\chi$.

Spectral Lines Different elements have different ionization energy:

- $\chi \approx 5\,\mathrm{eV}$: metallic elements like Li, Na, Mg, Al, K, Ca, etc.
- $10 \lesssim \chi \lesssim 20\,\mathrm{eV}$: H, C, N, O, Ar,
- $\chi \gtrsim 20$: noble gases like He, Ne, etc.

The strength of spectral lines depends on both the ionization state and the excitation states of the elements. The Balmer lines for hydrogen are strong when there are large number of hydrogen atoms in the first excited state ($n = 2$). According to Boltzmann statistics, the number density (or probability) of neutral hydrogen being in an excited state n is

$$\mathcal{P}(n) \propto g_n e^{-\epsilon_n/kT}, \quad \epsilon_n = -\frac{13.6\,\mathrm{eV}}{n^2}.$$

where $g_n = 2n^2$ for hydrogen. This degeneracy can be obtained from quantum mechanical analysis of the hydrogen atom. For example, the ratio of number densities of first excited and ground states hydrogen is

$$\frac{n_2}{n_1} = \frac{g_2}{g_1} e^{+13.6\,\mathrm{eV}(1/4-1)/kT} = 4e^{-10.2\,\mathrm{eV}/kT}.$$

To see Balmer lines, you will need enough hydrogen atoms in the first excited state, and hence a high enough temperature. Starting from the coolest stars to hottest stars (type M, K, G, F, A, B, O), More and more hydrogen atom are in the first excited state due to higher temperature. However, if the temperature is too high, the hydrogen will be in higher excited states or even be ionized. Thus, the Balmer lines is the strongest in type A stars. In general, as temperature goes up, the order of disappearance of lines is: molecular lines, metals with low first ionization potential, hydrogn, helium.

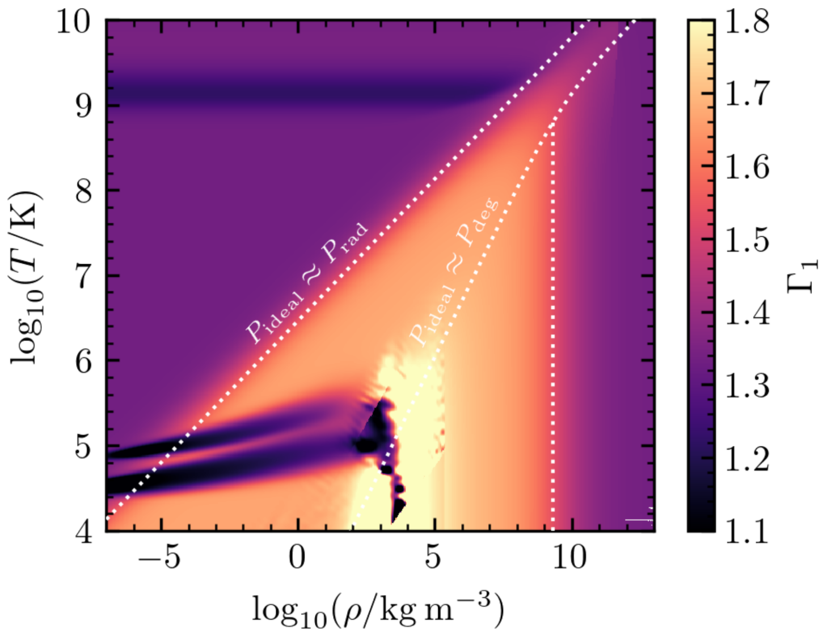


Figure 2.2: The first adiabatic exponent Γ_1 as a function of temperature and density for pure helium.

Ionization Effects on the Equation of State When there is ionization, the total pressure is

$$P = (n_p + n_e + n_{\rm H})kT.$$

Plug in the number density and ionization fraction relation,

$$P = (1+x)\frac{\rho}{m_{\rm H}}kT.$$

Internal energy density is not just $\frac{3}{2}nkT$, as it needs to account for ionization energy,

$$u = \frac{3}{2}(n_p + n_e + n_{\rm H})kT + n_p\chi \quad \Longrightarrow \quad u = \frac{\rho}{m_{\rm H}}\left[\frac{3}{2}(1+x)kT + x\chi\right].$$

Recall that in an adiabatic process, PV^γ is constant, where γ is the adiabatic index. When ionization happens, γ drops because some energy goes into ionization, so the temperature and pressure don't increase as much. Figure 2.2 is a plot of the first adiabatic exponent Γ_1 (or simply the adiabatic index) as a function of temperature and density.

We can divide the plot into two main colors, one for $\gamma = 5/3$ (non-relativistic) and the other for $\gamma = 4/3$ (relativistic). There are three regions where γ drops significantly. The first two are at $\rho \lesssim 10^{-2}\,\mathrm{g\,cm^{-3}}$ and $T \approx 10^{4-5}\,\mathrm{K}$. These are the two partial ionization zones for helium because helium has two electrons. Around $T \sim 10^9$, electron-positron pair production ($2\gamma \longleftrightarrow e^+ + e^-$, this γ refers to photons) happens. The adiabatic constant drops at this band. There is also a region where $\gamma > 5/3$. This region is poorly understood for now due to many-body effects.

Pressure Ionization Near the center of the star with higher density, Saha equation predicts that are substantial neutral atoms. This is not true. At higher densities, electrostatic potentials start to superimpose. When electron from one atom gets close to another atom, the **pressure ionization** happens. Ionization energy decreases in this case. For hydrogen atoms in the ground state, the electron distance from the ion is given by the Bohr radius $a_0 \approx 0.529 \times 10^{-10}\,\mathrm{m}$. If we assume an atom occupies a volume of $\frac{4\pi d^3}{3}$, then

$$\frac{4\pi d^3}{3}n = 1 \quad \Longrightarrow \quad d = \left(\frac{3}{4\pi n}\right)^{1/3},$$

where n is the number density. Pressure ionization happens when

$$\left(\frac{3}{4\pi n}\right)^{1/3} \lesssim a_0.$$

This corresponds to a density of

$$\rho = n m_{\mathrm{H}} \gtrsim \frac{m_{\mathrm{H}}}{4\pi a_0^3/3} \approx 3 \times 10^3\,\mathrm{kg/m^3},$$

about three times the density of water. Other elements pressure-ionize at similar densities, so at higher densities, we assume the gas is fully ionized.

Crystallization Ions are usually classical (non-degenerate), even when electrons are degenerate. However, at high densities and low temperatures, electrostatic (Coulomb) interactions between ions can be significant. These interactions cause ions to crystalize into a non-degenerate and non-relativistic lattice. The condition for crystallization is when

$$U_e = \frac{Z^2 e^2}{4\pi\epsilon_0 d} \gg kT,$$

where U_e is the electrostatic potential energy between two ions with charge Ze separated by a distance d. Again we let

$$\frac{4\pi d^3}{3} n_{\mathrm{I}} = 1 \quad \text{and} \quad n_{\mathrm{I}} = \frac{\rho}{A m_p} = \frac{Z}{A}\frac{\rho}{Z m_p} \approx \frac{2\rho}{Z m_p}.$$

In the last approximation, we take $Z/A \approx 2$, which works for most ions like helium, carbon, oxygen, etc. These are typical compositions for white dwarfs where crystallization is likely to happen. Solving for d and plug in U_e, we have

$$\Gamma_c = \frac{U_e}{kT} \propto \frac{\rho^{1/3} Z^{5/3}}{T}.$$

Γ_c is called the Coulomb parameter. When $\Gamma_c \gtrsim 1$, the ions transform from gas phase to liquid phase. At $\Gamma_c \approx 175$, crystallization happens. Thus, white dwarfs crystallize at

$$\frac{Z^2 e^2}{4\pi\epsilon_0 d} \gtrsim 175 kT.$$

2.3.6 Conclusions on Equation of States

The total pressure consists of the radiation pressure, electron pressure, and ion pressure,

$$P_{\mathrm{rad}} = \frac{1}{3} a T^4, \quad P_{\mathrm{I}} = n_{\mathrm{I}} kT, \quad P_e = n_e kT.$$

Radiation pressure dominates at high temperature and low density. When the density is high, electrons becomes degenerate, and their pressure is given by the degeneracy pressure,

$$P_{nr} = \frac{h^2}{5 m_e}\left(\frac{3}{8\pi}\right)^{2/3} n_e^{5/3}, \quad P_r = \frac{hc}{4}\left(\frac{3}{8\pi}\right)^{1/3} n_e^{4/3}.$$

In the classical regime at low temperatures $T \lesssim 10^5$ K, the gas may be partially ionized. This is where Saha equation comes into play,

$$\frac{x^2}{1-x}\frac{\rho}{m_i} = \left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} e^{-\chi/kT}.$$

2.4 Energy Transport

2.4.1 Photon Diffusion

Imagine a volume with cross-sectional area A and length l. There are electrons lining up perfectly in the volume with number density n_e, so the total number of electrons is $N = n_e A l$. We assume that the photons are scattered by Thomson scattering (cross-sections

$\sigma_\mathrm{T} \approx 6.7 \times 10^{-29}\,\mathrm{m}^2$). The total scattering area is then

$$\sigma_\mathrm{T} N = \sigma_\mathrm{T} n_e A l.$$

We expect one scattering when $A_\mathrm{sc} \approx \sigma_\mathrm{T} n_e A l$ because the Sun is marginally opaque. This gives the mean distance, or the **mean free path**, of one photon needs to travel to have one scattering,

$$l \approx \frac{1}{\sigma_\mathrm{T} n_e}. \tag{2.16}$$

For the Sun, the number density of electrons is $\sim 10^{30}\,\mathrm{m}^{-3}$, so the mean free path of a photon is about $l \sim 0.01\,\mathrm{m} \ll R_\odot$. This is just saying a photon needs to scatter many times before it leaves the Sun. In general, the mean free path can also be expressed as

$$l = \frac{1}{\kappa\rho},$$

where κ is called the **opacity**, the cross-sectional area per unit mass. This gives the opacity and Thomson scattering cross-sections relation

$$\kappa_{es} = \frac{n_e \sigma_\mathrm{T}}{\rho},$$

where κ_{es} stands for opacity of electron scattering. Suppose a star is only composed of hydrogen and helium, with hydrogen mass fraction X and helium mass fraction $Y = 1 - X$. Then if we assume full ionization, the number density of electron is

$$n_e = \frac{X\rho}{m_p} + \frac{2Y\rho}{4m_p} = \frac{X\rho}{m_p} + \frac{(1-X)\rho}{2m_p} = \frac{(1+X)\rho}{2m_p}.$$

Plugging this into the opacity, we find that the electron scattering opacity is roughly a constant,

$$\kappa_{es} = \frac{n_e\sigma_\mathrm{T}}{\rho} = \frac{(1+X)\sigma_\mathrm{T}}{2m_p} \approx 0.2(1+X)\,\mathrm{cm}^2/\mathrm{g}.$$

Suppose a photon travel N_steps in a star. On average, the distance of this photon is given by the "random walk" standard deviation times the mean free path,

$$d \approx l\sqrt{N_\mathrm{steps}}.$$

If the photon wants to get out of the star, we just set $R \sim l\sqrt{N_\mathrm{step}}$, which gives the number of steps about $N_\mathrm{step} \sim 10^{22}$. Each step will take time l/c, so the **diffusion time** of photons is

$$t_\mathrm{diff} = \frac{l}{c} N_\mathrm{steps} \approx \frac{R^2}{lc} \sim 10^4\,\mathrm{yr}.$$

Heat Transfer In stars, a pressure gradient force is opposing gravity to maintain hydrostatic equilibrium. In general, whenever there is a pressure gradient, there is also a temperature gradient—the center is hotter than the surface. By the second law of thermodynamics, heat flows outward to the surface. This is the energy transport within stars.

Now consider a setup of plane parallel temperature and energy density, where $T(x+l) < T(x) < T(x-l)$. The particles are moving with speed v and mean free path l. We are to calculate the flux arriving at x. We expect the flux is proportional to the speed of the particle, so it can transfer heat faster, and to the energy density above or below, so there is more energy to transfer. The flux from $x + l$ is

$$F(x+l) = \frac{1}{6} v u(x+l) \approx \frac{1}{6} v \left[u(x) + l\frac{du}{dx}\right].$$

Here $u(x)$ is a function, not a product. The factor of $1/6$ comes from: the $1/3$ factor from one of the three Cartesian coordinates (we only want x in this case), and $1/2$ from counting only particles in the right direction (which is $-x$). The last equality is the Taylor expansion around x. Similarly, the flux from $x - l$ is

$$F(x-l) = -\frac{1}{6} v \left[u(x) - l\frac{du}{dx}\right].$$

The minus sign comes from the fact that $F(x-l)$ is flowing opposite to $F(x+l)$. The net flux is

$$F_\mathrm{net} = F(x+1) + F(x-1) \approx \frac{1}{6} v \left[u(x) - l\frac{du}{dx}\right] - \frac{1}{6} v \left[u(x) + l\frac{du}{dx}\right] = -\frac{1}{3} v l \frac{du}{dx}.$$

The energy flux depends on the speed of the particle v, how far it could travel (the mean free path l), and the energy density gradient. Now we put the energy flux back to spherical symmetry. The energy flux is

$$F(r) = -\frac{1}{3}vl\frac{du}{dr}.$$

For photons, $v = c$, $l = (\kappa\rho)^{-1}$, and $u = aT^4$. This gives the **equation of radiative diffusion**:

$$\boxed{F(r) = -\frac{4acT^3}{3\kappa\rho}\frac{dT}{dr}.} \tag{2.17}$$

Since the pressure has a close relation to energy density, $P_{\text{rad}} = u/3$, we can also write

$$F(r) = -\frac{c}{\kappa\rho}\frac{dP_{\text{rad}}}{dr}.$$

Luminosity The **luminosity** is defined as the energy per unit time flowing outwards,

$$L(r) = 4\pi r^2 F(r).$$

We want to make an order-of-estimate for the surface luminosity $L \equiv L(R)$. The temperature gradient and the flux are estimated as

$$\frac{dT}{dr} \sim \frac{T(0) - T(R)}{0 - R} \sim -\frac{T}{R} \quad\Longrightarrow\quad F \propto \frac{1}{\kappa\rho}\frac{T^4}{R} \quad\Longrightarrow\quad L \propto R\frac{1}{\kappa\rho}T^4.$$

Here $T(R)$ is neglected because it is much smaller than the central temperature. The Virial theorem says that

$$\langle P\rangle = -\frac{1}{3}\frac{E_g}{V} \sim \frac{GM^2}{R^4}.$$

For a classical ideal gas, the typical temperature can be obtained by

$$P = \frac{\rho kT}{\mu m_p} \quad\Longrightarrow\quad kT \sim \frac{GM\mu m_p}{R} \quad\Longrightarrow\quad T \propto \frac{M\mu}{R}.$$

We now have a relation between the temperature and the mass of a star. Putting is back to the flux, using $\rho \sim M/R^3$ and assuming Thomson scattering (so κ is constant),

$$F \propto \frac{1}{M/R^3}\frac{(M\mu/R)^4}{R} \propto \frac{M^3\mu^4}{R^2} \quad \text{and} \quad L \propto R^2F \propto \mu^4 M^3.$$

There are several conclusions and caveats drawn from this

1. The luminosity of a star depends on the mass cubed. This is a good approximation for solar-like stars with $L \propto M^{3.5}$.
2. The stellar lifetime scales like M/L, or M^{-2}. That is, more massive stars have shorter lifetime.
3. As a star fuses hydrogen into helium, its mean molecular mass increases. If it is initially composed of pure hydrogen, then μ changes from $1/2$ to $4/3$ (pure He). Since $L \propto \mu^4$, the star's luminosity is constantly increasing. Indeed, the Sun is now about 30% brighter than when it was born.
4. When the Sun turns into a red giant, its mass will be about the same, so the luminosity should not change. This is inconsistent with observational data for red giant stars. The assuption of radiation diffusion is false in this case because at red giant stage, the outer layers have energy transported by convection instead.
5. If the star is dominated by radiation pressure instead of ideal gas pressure,

   $$P_{\text{rad}} = \frac{1}{3}aT^4.$$

   A different mass-luminosity relation will be obtained,

   $$L \sim \left(\frac{cGm_p}{\sigma_{\text{T}}}\right)M \propto M.$$

Stars supported by degeneracy pressure (white dwarfs and neutron stars) also has the same problem.

6. Thomson scattering is a good approximation at high temperatures and low densities, while there are other processes that scatter photons at lower temperatures and higher densities. We also neglect the dependence of electron number density on composition. These will be accounted for in later sections.

Eddington Standard Model Suppose a star transports energy via radiative diffuion. The **Eddington standard model** assumes that the classical ideal ga pressure P_{gas} is some fraction β of the total pressure P, $P_{\text{gas}} = \beta P$. The radiation pressure is then $P_{\text{rad}} = (1-\beta)P$. Edddington also assumed that β is constant throughout the star. We will derive the temperature and total pressure as a function of density.

First, the ideal gas pressure and radiation pressure are given by

$$P_{\text{gas}} = \frac{\rho}{\mu m_p} kT = \beta P \quad \text{and} \quad P_{\text{rad}} = \frac{1}{3} a T^4 = (1-\beta) P.$$

Rearranging some terms,

$$P = \frac{\rho}{\beta \mu m_p} kT = \frac{1}{3(1-\beta)} a T^4.$$

Solving for T gives

$$T = \left(\frac{1-\beta}{\beta} \frac{3}{a} \frac{k\rho}{\mu m_p} \right)^{1/3}. \tag{2.18}$$

The total pressure is the sum of the ideal gas pressure and radiation pressure:

$$\begin{aligned}
P = P_{\text{rad}} + P_{\text{gas}} &= \frac{\rho}{\mu m_p} kT + \frac{1}{3} a T^4 \\
&= \frac{\kappa\rho}{\mu m_p} \left(\frac{1-\beta}{\beta} \frac{3}{a} \frac{\kappa\rho}{\mu m_p} \right)^{1/3} + \frac{1}{3} a \left(\frac{1-\beta}{\beta} \frac{3}{a} \frac{k\rho}{\mu m_p} \right)^{4/3} \\
&= \left(\frac{1-\beta}{\beta} \right)^{1/3} \left(\frac{3}{a} \right)^{1/3} \left(\frac{k\rho}{\mu m_p} \right)^{4/3} + \left(\frac{1-\beta}{\beta} \right) \left(\frac{1-\beta}{\beta} \right)^{1/3} \left(\frac{3}{a} \right)^{1/3} \left(\frac{k\rho}{\mu m_p} \right)^{4/3} \\
&= \left(1 + \frac{1-\beta}{\beta} \right) \left(\frac{1-\beta}{\beta} \right)^{1/3} \left(\frac{3}{a} \right)^{1/3} \left(\frac{k\rho}{\mu m_p} \right)^{4/3} \\
&= \frac{1}{\beta} \left(\frac{1-\beta}{\beta} \right)^{1/3} \left(\frac{3}{a} \right)^{1/3} \left(\frac{k\rho}{\mu m_p} \right)^{4/3}.
\end{aligned}$$

$$P = \left(\frac{1-\beta}{\beta^4} \right)^{1/3} \left(\frac{3}{a} \right)^{1/3} \left(\frac{k}{\mu m_p} \right)^{4/3} \rho^{4/3}. \tag{2.19}$$

Combining the equation of hydrostatic equilibrium and radiative diffusion,

$$\frac{dP}{dr} = -\rho \frac{Gm(r)}{r^2} \quad \text{and} \quad L = 4\pi r^2 F(r) = -\frac{16acr^2T^3}{3\kappa\rho} \frac{dT}{dr},$$

one can derive dT/dP by approximating in the envelope:

$$\frac{dT}{dP} = \frac{dT}{dr} \frac{dr}{dP} = \left(-\frac{3\kappa\rho L}{16\pi acr^2T^3} \right) \left(-\frac{r^2}{Gm\rho} \right) = \frac{3}{16\pi acG} \frac{\kappa}{T^3} \frac{L}{m} \propto \frac{\kappa}{T^3} \frac{L}{m}.$$

Now we can find a way to relate β with κ, L, and m,

$$1 - \beta = \frac{dP_{\text{rad}}}{dP} = \frac{dP_{\text{rad}}}{dT} \frac{dT}{dP} = \left(\frac{4}{3} a T^3 \right) \frac{3}{16\pi acG} \frac{\kappa}{T^3} \frac{L}{m} = \frac{1}{4\pi cG} \frac{\kappa L}{m}.$$

In order for β to be constant as Eddington has assumed, the quantity $\kappa(r)L(r)/m(r)$ must be constant. In radiation pressure dominated regions, κ is mostly Thomson scattering opacity, and $L \propto m$ as shown in the **Luminosity** part, β is approximately constant.

2.4.2 Conduction

Heat transfer is also possible via conduction by ions and electrons. At some temperature T, both the ions and electrons will have the same average kinetic energy, $\frac{1}{2}mv^2 \approx kT$. However, since ions are $\sim 10^3$ times heavier than electrons, they travel at a much lower speed. Therefore, we assume that only electrons carry conduction energy and neglect ions.

The interactions between electrons with itself and ions are mainly Coulomb interactions. Consider a particle with charge Z_1 is coming to another particle with charge Z_2 with an impact parameter (perpendicular distance) b. The deviation of trajectory of Z_1 is significant when the electrostatic potential energy is comparable to its kinetic energy,

$$\frac{Z_1 Z_2}{4\pi\epsilon b} \sim kT \quad \Longrightarrow \quad b \sim \frac{Z_1 Z_2}{4\pi\epsilon_0 kT}.$$

The cross-sections is about

$$\sigma \sim \pi b^2 \sim \left(\frac{Z_1 Z_2}{4\pi\epsilon_0 kT}\right)^2.$$

For electrons-ion (abbr. ie, ion with charge Ze) and electron-electron (ee) interactions, we have

$$(n\sigma)_{ie} \sim n_{\mathrm{I}} \left(\frac{Ze^2}{4\pi\epsilon_0 kT}\right)^2 \quad \text{and} \quad (n\sigma)_{ee} \sim n_e \left(\frac{e^2}{4\pi\epsilon_0 kT}\right)^2 = Z n_{\mathrm{I}} \left(\frac{e^2}{4\pi\epsilon_0 kT}\right)^2.$$

If those collisions simply add up, then the mean free path of electrons satisfies

$$l^{-1} = (n\sigma)_{ie} + (n\sigma)_{ee} \sim Z^2 n_{\mathrm{I}} \left(\frac{e^2}{4\pi\epsilon_0 kT}\right)^2 + Z n_{\mathrm{I}} \left(\frac{e^2}{4\pi\epsilon_0 kT}\right) = (Z+1) Z n_{\mathrm{I}} \left(\frac{e^2}{4\pi\epsilon_0 kT}\right)^2.$$

Conduction vs Radiation We will make order-of-magnitude estimate to compare the efficiency of energy transport via conduction and radiation. The energy density gradient is approximated as

$$\frac{du}{dr} \sim -\frac{u}{R} \sim -\frac{P}{R}.$$

Assume that the cross-section relevant to photons is the Thomson scattering. All we need are the cross-sections and the energy flux formula,

$$n_e \sigma_{\mathrm{T}} = n_e \frac{8\pi}{3} \left(\frac{e^2}{4\pi\epsilon_0 m_e c^2}\right)^2, \quad (n\sigma)_e \sim n_{\mathrm{I}} \left(\frac{Ze^2}{4\pi\epsilon_0 kT}\right)^2, \quad \text{and} \quad F \sim -vl\frac{du}{dr}.$$

The ratio between the energy fluxes carried by electrons and by photons is

$$\frac{F_e}{F_\gamma} \sim \frac{v_e}{c} \frac{n_e \sigma_{\mathrm{T}}}{(n\sigma)_e} \frac{P_e}{P_\gamma} \sim \frac{(kT/m_e)^{1/2}}{c} \frac{[e^2/(4\pi\epsilon_0 m_e c^2)]^2}{[e^2/(4\pi\epsilon_0 kT)]^2} \frac{P_e}{P_\gamma} \approx \left(\frac{kT}{m_e c^2}\right)^{5/2} \frac{P_e}{P_\gamma}.$$

In the first step, we used the definition $l = 1/n\sigma$. The typical central temperature of a star is $\sim 10^7\,\mathrm{K}$. The first factor goes to 10^{-7}. Using the central density of the Sun, $\rho \approx 1.5 \times 10^5\,\mathrm{kg/m^3}$,

$$\frac{P_e}{P_\gamma} = \frac{\rho kT/\bar{m}}{\frac{1}{3}aT^4} \approx 10^3.$$

This gives the flux ratio

$$\frac{F_e}{F_\gamma} \sim 10^{-4}.$$

The energy transport by electrons is much less efficient than that by photons. This means we can assume that all the energy transport is by photon diffusion. However, in the above derivation the electrons are assumed to be a classical ideal gas. When electrons become degenerate, they travel much faster and have a much smaller cross-section because they can only occupy states near the Fermi momentum. In other words, electron conduction dominates at high density stars.

Total Flux In general, the total flux can be written as

$$F_{\mathrm{tot}} = F_\gamma + F_e = -\frac{4}{3} \frac{ac}{\kappa_{\mathrm{rad}}\rho} T^3 \frac{dT}{dr} - \frac{4}{3} \frac{ac}{\kappa_{ec}\rho} T^3 \frac{dT}{dr} = -\frac{4}{3} \frac{ac}{\kappa_{\mathrm{tot}}\rho} T^3 \frac{dT}{dr},$$

where κ_{ec} is the electron conduction opacity. The total opacity κ_{tot} is

$$\frac{1}{\kappa_{\text{tot}}} = \frac{1}{\kappa_{\text{rad}}} + \frac{1}{\kappa_{ec}}.$$

2.4.3 Optical Depth

It is convenient to introduce a dimensionless quantity that measures the number of photon scattering would occur if it were to continue in a straight line. The expected number of scattering over a distance D is

$$\tau = \frac{D}{l} = \kappa\rho D. \tag{2.20}$$

The quantity τ is called the **optical depth**. Like opacity, it also describes how opaque a material is. At the surface of a star, τ is defined to be zero because there will be no scattering in free space. The optical depth is increasing towards the center. This gives the differential form of optical depth:

$$d\tau = -\kappa\rho\, dr. \tag{2.21}$$

Recall that the flux of the star is given by the radiative diffusion equation (2.17). The derivative can be replaced to one respect to the optical depth,

$$F = \frac{ac}{3}\frac{1}{\kappa\rho}\frac{d(T^4)}{dr} = \frac{ac}{3}\frac{d(T^4)}{d\tau}.$$

If we assume the atmosphere of a star is plane-parallel, i.e. the thickness of it is much smaller than the radius, we have

$$F = \frac{L}{4\pi r^2} = \text{const.}$$

where $r \sim R$ is radial coordinate in the atmosphere. In terms of optical depth,

$$\frac{d(T^4)}{d\tau} = \frac{3F}{ac} = \text{const.}$$

The temperature to the forth power is proportional to τ in the plane-parallel atmosphere,

$$T^4 \approx T^4_{\tau=0} + \frac{3F}{ac}\tau. \tag{2.22}$$

Now forget about the atmosphere and do an order of magnitude estimate. For the deep interior with temperature $T_{\text{I}} \sim 10^7\,\text{K}$,

$$T_{\text{I}}^4 \approx \frac{3F}{ac}\tau \quad \Longrightarrow \quad F \approx \frac{ac}{3}T_{\text{I}}^4\frac{1}{\tau}.$$

Here $T^4_{\tau=0}$ is neglected because it is much smaller than the central temperature to the forth power. The flux looks a lot like the Stefan-Boltzmann law, but with a factor of $1/\tau$,

$$F \approx \sigma T_{\text{I}}^4\frac{1}{\tau}.$$

Therefore, even though the central temperature is much higher than the surface, their flux is reduced by $\tau \sim R/l$ due to photon diffusion. The spectrum we observe is mostly about the surface of the star.

Effective Temperature and Photosphere Define the **effective temperature** T_{eff} as

$$L = 4\pi R^2\sigma T^4_{\text{eff}}.$$

Assume that photons are isotropic everywhere except at $\tau = 0$ (this is called the **Eddington approximation**). At $\tau = 0$, half of the photons are free-streaming away and never coming back, so there are only half the photons one would expect if one were bathed in radiation of temperature T_{eff}. The energy density at the surface is

$$u_{\text{rad,surf}} \approx \frac{1}{2}aT^4_{\text{eff}}.$$

The true surface temperature follows $u_{\text{rad,surf}} \approx aT^4_{\text{surf}}$. We can solve for the surface temperature in terms of the effective temperature, $T_{\text{surf}} \approx \left(\frac{1}{2}\right)^{1/4} T_{\text{eff}} \approx 0.841\, T_{\text{eff}}$. This means the surface temperature is a bit smaller than the effective temperature. Rewriting the flux in terms of the effective temperature in the interior and substitute into (2.22),

$$F = \sigma T_{\text{eff}}^4 = \frac{ac}{4} T_{\text{eff}}^4 \quad \implies \quad T^4 = \frac{1}{2} T_{\text{eff}}^4 + \frac{3}{4} \tau T_{\text{eff}}^4 = \frac{3}{4} T_{\text{eff}}^4 \left(\tau + \frac{2}{3} \right).$$

Note that if $\tau = 2/3$, $T = T_{\text{eff}}$. Thus we define the **photosphere** to be the region where $\tau = 2/3$. This is approximately where the last photon scattering occurs. By hydrostatic equilibrium and the optical depth differential (2.21), we may determine the pressure in the photosphere,

$$\frac{dP}{dr} = -\rho g \quad \implies \quad \frac{dP}{d\tau} = \frac{g}{\kappa}.$$

If the atmosphere is thin, or equivalently, the pressure scale height is much smaller than the radius,

$$h \equiv \left| \frac{P}{dP/dr} \right| \ll R,$$

g is about constant, and we also assume that κ is constant. Then integrating on both sides,

$$\int_0^{2/3} \frac{dP}{d\tau} \, d\tau = \frac{g}{\kappa} \int_0^{2/3} d\tau \quad \implies \quad P(\tau = 2/3) - P(\tau = 0) = \frac{2g}{3\kappa}.$$

The pressure at the surface is zero where $\tau = 0$, so the pressure of the photosphere is

$$P_{\text{phot}} = \frac{2g}{3\kappa}. \tag{2.23}$$

2.4.4 Rosseland-Mean Opacity

We have talked about photon diffusion and eventually reached equation of radiative diffusion (2.17),

$$F = -\frac{4ac}{3} \frac{1}{\kappa \rho} T^3 \frac{dT}{dr}.$$

This equation is of course true, but many quantities in this equation are functions of temperature, density, and frequency of light. Let $F_\nu \, d\nu$ be the flux contribution from photons with frequency between ν and $\nu + d\nu$, similarly for κ_ν and u_ν. The equation of radiative diffusion at a certain frequency is

$$F_\nu = -\frac{1}{3} \frac{c}{\kappa_\nu \rho} \frac{du_\nu}{dr} = -\frac{1}{3} \frac{c}{\kappa_\nu \rho} \frac{du_\nu}{dT} \frac{dT}{dr},$$

where we have absorbed $4aT^3 \, dT$ into du_ν. The total flux is then

$$F = \int_0^\infty F_\nu \, d\nu = -\frac{1}{3} \frac{c}{\rho} \frac{dT}{dr} \int_0^\infty \frac{1}{\kappa_\nu} \frac{du_\nu}{dT} \, d\nu.$$

We are looking for an "average" opacity $\langle \kappa \rangle$ so that the flux can be written as

$$F = -\frac{1}{3} \frac{c}{\langle \kappa \rangle \rho} \frac{du}{dT} \frac{dT}{dr}.$$

This defines the **Rosseland mean opacity**,

$$\frac{1}{\langle \kappa \rangle} = \int_0^\infty \frac{1}{\kappa_\nu} \frac{du_\nu}{dT} \, d\nu \bigg/ \frac{du}{dT} = \int_0^\infty \frac{1}{\kappa_\nu} \frac{du_\nu}{dT} \, d\nu \bigg/ \int_0^\infty \frac{du_\nu}{dT} \, d\nu.$$

The Rosseland mean opacity still depends on temperature and density, but no longer on frequency. We can see from its definition that $1/\kappa_\nu$ is weighted by du_ν/dT. Since the flux F_ν is proportional to du_ν/dT, $1/\kappa_\nu$ is weighted more for more significant energy transport by frequency ν.

Then what exactly is u_ν? We can derive it using Planck distribution (just Bose-Einstein distribution with $\mu = 0$ for photons) and the density of states in momentum space,

$$u_p \, dp = \epsilon(p) g_s g(p) f(p) \, dp = pc \frac{8\pi p^2}{h^3} \frac{1}{e^{pc/kT} - 1} \, dp = \frac{8\pi c}{h^3} \frac{p^3}{e^{pc/kT} - 1} \, dp.$$

For photons, $pc = h\nu$, the energy density u_ν in frequency space is

$$u_\nu \, d\nu = \frac{8\pi c}{h^3} \frac{(h\nu/c)^3}{e^{h\nu/kT} - 1} \frac{h \, d\nu}{c}.$$

This gives

$$u_\nu = \frac{8\pi h}{c^3} \frac{\nu^3}{e^{h\nu/kT} - 1}. \tag{2.24}$$

Since the opacity is a function of frequency, so is the depth of the photosphere. The differential form of optical depth says $d\tau = \kappa\rho \, dr$, so higher κ_ν means smaller photosphere depth for $\tau = 2/3$, and also lower flux F_ν and lower temperature of that photosphere. Hence these frequencies will be the dark regions in the spectrum. You can see deeply into the stars only for frequency range in which photons do not scatter much by the stellar material. With this knowledge, we can deduce the star's composition using the dark regions in the spectrum.

Opacity Sources There are several absorption/scattering processes that contributes to the Rosseland mean opacity.

1. Electron scattering (Thomson scattering): a photon scatters off of a free electron,

$$\kappa_{es} = \frac{\sigma_{\mathrm{T}} n_e}{\rho} = \frac{\sigma_{\mathrm{T}}}{\mu_e m_p} \approx 0.02(1 + X) \, \mathrm{m}^2/\mathrm{kg}.$$

2. Free-free absorption (inverse bremsstrahlung): in the bremsstrahlung process, a free electron decelerates near an ion and emits a photon. The inverse is the process in which an electron absorbs a photon and accelerates. This process has cross-section $\sigma \propto \nu^{-3} T^{1/2}$. Integrating over all frequencies gives $\langle \kappa \rangle \propto \rho T^{-3.5}$. Any opacity obeying $\propto \rho T^{-3.5}$ is called **Kramers' opacity**.
3. Bound-free absorption (photoionization): at low temperature regions where partial ionization appears, a bound electron absorbs a photon and get ionized. The opacity has the form of Kramers' opacity.
4. Bound-bound absorption: a bound electron gets excited by absorbing a photon. It also has Kramers' opacity and the temperature is roughly the same as that of bound-free absorption.
5. H^- opacity, $\kappa \propto \rho^{0.5} T^9$: H^- is an ion with one proton and two electrons with ionization energy only $0.75 \, \mathrm{eV}$. It is formed by capturing a free electron ionized from alkali metals. H^- exist at $3000 \, \mathrm{K} < T < 8000 \, \mathrm{K}$. It is a significant opacity source at the Sun's photosphere and other low-mass stars or pre-main sequence stars.
6. Molecular opacity: at even lower temperatures ($T < 3000 \, \mathrm{K}$), molecules can form and can contribute to the opacity.

2.4.5 Convection

The equation of radiative diffusion (2.17) says that a temperature gradient gives rise to an energy transport via radiation,

$$F = -\frac{4acT^3}{3\kappa\rho} \frac{dT}{dr}.$$

However, the efficiency of radiative diffusion also depends on the opacity κ. If the opacity is large, the energy cannot get out from a certain region. In this section, we will discuss another way of transporting energy, namely **convection**. For a classical ideal gas in general, a higher temperature indicates a lower density. A blob with lower density tends to rise above the dense materials and cools down, giving of energy. It then gets less buoyant and sinks to the high temperature region, gains more energy, rises up, releases energy, and so on. This repetitive process is called convection.

As usual, we will divide a star into small layers at various radius. A layer (we will call it the bottom layer) is characterized by a temperature T, pressure P, and density ρ. Another layer (top layer) just above it might have a different temperature, pressure and density: $T + dT$, $P + dP$, and $\rho + d\rho$. Suppose a blob is in thermal and mechanical equilibrium (same pressure) with its surrounding at the bottom layer,

$$T_b = T, \quad P_b = P, \quad \rho_b = \rho,$$

where the subscript b denotes the quantities of the blob. If the blob is displaced by a distance δr and reaches the top layer, it will have a new set of quantities. Assume that the blob is always in mechanical equilibrium with the environment, $P_b + dP_b = P + dP$, or $dP_b = dP$. This is the quasistatic assumption, in which the blob travels at a speed less than the local sound speed. We will also assume that this process is adiabatic for now: it is too fast to have energy flowing between the blob and the environment. For an adiabatic process of a classical ideal gas,

$$PV^\gamma = \text{const.} \quad \text{or} \quad P \propto \rho^\gamma \quad \Longrightarrow \quad \ln P = \gamma \ln \rho + \text{const.},$$

where γ is the adiabatic index. Differentiating on both sides,

$$\frac{dP}{P} = \gamma \frac{d\rho}{\rho}. \tag{2.25}$$

Additionally, the ideal gas law tells that $P \propto \rho T$. Then

$$\ln P = \ln \rho + \ln T \quad \Longrightarrow \quad \frac{dP}{P} = \frac{d\rho}{\rho} + \frac{dT}{T}.$$

By two dP/P relations above, we have

$$\frac{dT}{T} = \left(1 - \frac{1}{\gamma}\right) \frac{dP}{P}.$$

This equation is often written in terms of natural logarithms,

$$d \ln T = \left(1 - \frac{1}{\gamma}\right) d \ln P \quad \Longrightarrow \quad \left(\frac{d \ln T}{d \ln P}\right)_{\text{ad.}} = 1 - \frac{1}{\gamma}, \tag{2.26}$$

where ad. refers to adiabatic processes. Apply the adiabatic process to the blob but not the environment. When it travels a displacement of δr, its new pressure is $P + \delta r(dP/dr)$. By (2.25),

$$d\rho_b = \frac{1}{\gamma} \frac{\rho_b}{P_b} dP_b = \frac{1}{\gamma} \frac{\rho}{P} dP.$$

The density of the surroundings changes by $d\rho$ after the blob is displaced by δr. If now the blob is overdense compared to its surroundings, $d\rho_b > d\rho$, it will sink. If it is underdense, $d\rho_b < d\rho$, it will continue to float up. The latter case is called **convectively unstable**, with the criterion

$$\frac{1}{\gamma} \frac{\rho}{P} dP \quad \Longrightarrow \quad \frac{1}{\gamma} \frac{dP}{P} < \frac{d\rho}{\rho}.$$

To make convection happen, we don't want the blob to oscillate around some point with small displacement δr. Instead, we want it to be convectively unstable so that it can travel a longer distance to transport energy. In general, once the blob travels long enough, it will dissolve among the environment and release energy. There are many other ways to write this criterion,

$$\frac{dT}{T} < \left(1 - \frac{1}{\gamma}\right) \frac{dP}{P}, \quad \text{or} \quad \frac{d \ln T}{d \ln P} > 1 - \frac{1}{\gamma} = \left(\frac{d \ln T}{d \ln P}\right)_{\text{ad.}},$$

where the sign flips because $dP < 0$. In convection zones in stars, we want the criterion to be related to the temperature gradient,

$$\frac{dT}{dr} < \left(\frac{\gamma - 1}{\gamma}\right) \frac{T}{P} \frac{dP}{dr} = \left(\frac{dT}{dr}\right)_{\text{ad.}}, \quad \text{or} \quad \frac{dT}{dr} < -\left(\frac{\gamma - 1}{\gamma}\right) \frac{T}{P} (\rho g). \tag{2.27}$$

In the second inequality, hydrostatic equilibrium is used and $g(r)$ is the gravitational acceleration. Because dT/dr is almost always negative, a steeper temperature gradient comparing to adiabatic temperature gradient leads to convection.

Convection in Stars It is known that radiative diffusion needs a temperature gradient

$$\left(\frac{dT}{dr}\right)_{\text{rad}} = -\frac{3\kappa\rho F}{4acT^3} = -\frac{3\kappa\rho}{4acT^3} \frac{L}{4\pi r^2}.$$

Once this temperature gradient (magnitude) exceeds a certain value, convection is triggered and dominates over radiative diffusion. We already know the criterion for convection. It can be used to determine the critical energy flux that triggers convection:

$$-\frac{3\kappa\rho}{4acT^3} \frac{L}{4\pi r^2} < \left(1 - \frac{1}{\gamma}\right) \frac{T}{P} \left(-\rho \frac{Gm}{r^2}\right).$$

Rearranging some terms,

$$\frac{L}{m} > \frac{\gamma - 1}{\gamma} \left(\frac{aT^4}{3P}\right) \frac{16\pi Gc}{\kappa}. \tag{2.28}$$

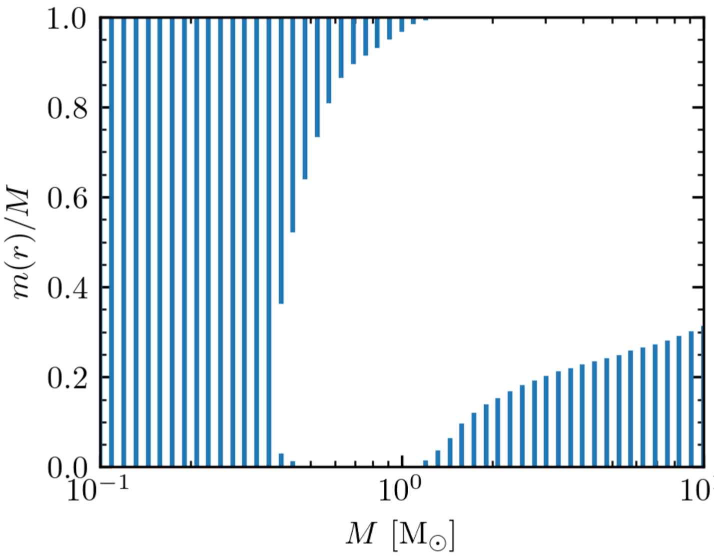


Figure 2.3: Location of convection zones (blue lines) in main sequence stars in terms of mass fraction $m(r)/M$.

Everything depends on the local condition except fundamental constants, G, c, and the radiation constant a. This inequality gives the information about where convection can happen:

1. $L(r)/m(r)$ is very large: cores of massive stars where nuclear fusion releases a large amount of energy in a small concentrated region.
2. Small $\gamma(r)$: partial ionization of H/He where the adiabatic index drops significantly.
3. Large $\kappa(r)$: partial ionization of H/He/Fe in outer envelops of stars.

Figure 2.3 shows where convection happens (in terms of mass fraction) for main sequence stars of different masses. On the main sequence, a $M \lesssim 0.5\, M_\odot$ is fully convective. A star with $0.5 \lesssim M \lesssim 1.2\, M_\odot$ has convection in the convective envelope. If a star has a mass above $1.2\, M_\odot$, then it has a convective core. Besides main sequence stars, red giants also have thick convective envelopes. A fully convective star will have an actual temperature gradient very close to the adiabatic gradient, so the adiabatic condition holds, $T \propto P^{2/5}$.

Convective Velocity To determine the motion of the blob, we invoke Archimedes' principle: the buoyant force is equal to the weight of the displaced fluid. The net force on the blob is the buoyant force minus the weight of it, which gives the equation of motion

$$\rho \frac{d^2 r}{dt^2} = g\, d\rho - g\, d\rho_b.$$

Define $\delta\rho = d\rho_b - d\rho$. The equation of motion becomes

$$\frac{d^2 r}{dt^2} = -\frac{\delta\rho}{\rho} g.$$

If the blob is displaced by δr, then

$$\delta\rho = d\rho_b - d\rho = \left(\frac{d\rho_b}{dr} - \frac{d\rho}{dr}\right)\delta r = \left(\frac{\rho}{\gamma P}\frac{dP}{dr} - \frac{d\rho}{dr}\right)\delta r = \left[\left(\frac{d\rho}{dr}\right)_{\text{ad.}} - \frac{d\rho}{dr}\right]\delta r.$$

The equation of motion reduces to

$$\frac{d^2(\delta r)}{dt^2} = -\frac{\delta\rho}{\rho} g = -\frac{g}{\rho}\left[\left(\frac{d\rho}{dr}\right)_{\text{ad.}} - \frac{d\rho}{dr}\right]\delta r \equiv -N^2 r,$$

where N is the **Brunt-Väisälä frequency**. If $N^2 > 0$, the blob is undergoing a simple harmonic motion. This type of motion is called **gravity waves**. If $N^2 < 0$, the solution is exponential, so the blob will keep floating up. This is convection. By the ideal gas law, $P \propto \rho T$ and defining $\delta P = dP_b - dP$ and $\delta T = dT_b - dT$, we have

$$\frac{\delta P}{P} = \frac{\delta \rho}{\rho} + \frac{\delta T}{T}.$$

Since $\delta P = dP_b - dP = 0$ because of the mechanical equilibrium assumption,

$$\frac{\delta \rho}{\rho} = -\frac{\delta T}{T} \quad \Longrightarrow \quad \frac{d^2 r}{dt^2} = \frac{\delta T}{T} g.$$

We will make a rough estimation of a typical convective velocity. Suppose the blob has zero initial velocity and rise a distance of $l_{\text{conv.}}$ (called the **mixing length**). Assume that $l_{\text{conv.}}$ is small so that gravity is constant. The final velocity will be

$$v_{\text{conv.}} \sim \left(\frac{\delta T}{T} g l_{\text{conv.}}\right)^{1/2}.$$

The mixing length is of the order the pressure scale height,

$$\alpha l_{\text{conv.}} \sim h \equiv \left|\frac{P}{dP/dr}\right| = \frac{P}{\rho g} \approx \frac{c_s^2}{g},$$

where c_s is the adiabatic sound speed (discussed in dynamical time scale section) and α is some constant of order unity. Thus,

$$v_{\text{conv.}} \sim \left(\frac{\delta T}{T} g \alpha h\right)^{1/2} \sim \left(\frac{\delta T}{T} \alpha\right)^{1/2} c_s.$$

By observations of stars, the parameter α is betweem 1.5 and 2. Recall the quasistatic assumption says that $v_{\text{conv.}} \ll c_s$. This means $|\delta T/T| = |\delta \rho/\rho| \ll 1$, or $dT_b \approx dT$.

Convective Luminosity After the blob rising $l_{\text{conv.}}$, it dissolves and release heat because its temperature is higher than the environment. Let the temperature difference between the blob and the surroundings be δT. The heat density it releases is

$$\delta u \approx \rho c_P \delta T, \quad \text{where} \quad c_P = \frac{\gamma}{\gamma - 1}\frac{k}{\bar{m}}.$$

Here c_P is the specific heat capacity at constant pressure. The flux carried by convection from the blob is approximated as

$$F_{\text{conv.}} \sim v_{\text{conv.}} \delta u \sim \left(\frac{\delta T}{T}\alpha\right)^{1/2} \rho c_s c_P \, \delta T = \rho c_P T \left(\frac{\delta T}{T}\right)^{3/2} \alpha^{1/2} c_s.$$

How large $\delta T/T$ would be if all the energy in the Sun is carried by convection from the center to the surface? Using the surface flux $F_\odot \approx 7 \times 10^7\,\text{J/m}^2\,\text{s}$, typical temperature $T \sim 10^7\,\text{K}$, average density $\bar{\rho} \approx 1400\,\text{kg/m}^3$, and typical sound speed $c_s \sim \sqrt{P/\rho} \sim 9 \times 10^5\,\text{m/s}$, the answer is only

$$\frac{\delta T}{T} \sim 10^{-9}.$$

A tiny temperature difference will result in a enormous amount of energy transferred by convection. The temperature gradient inside a star only needs to be slightly steeper than the adiabatic gradient to allow convection to carry energy,

$$\left(\frac{dT}{dr}\right)_{\text{conv.}} \lesssim \left(\frac{dT}{dr}\right)_{\text{ad.}}$$

It is $\lesssim$ because dT/dr is negative, and they cannot be exactly equal.

3 Nuclear Fusion

Nuclear fusion is the main energy generation process in stars. The rest mass energy in stars are released by nuclear fusion, fusing hydrogen to heavier elements.

3.1 Quantum Tunneling for Fusion

The existence of nuclear fusion in stars relies on quantum mechanics. We will treat the problem classically first, and then see why there must be quantum mechanical effects (namely quantum tunneling) for nuclear reactions to happen.

3.1.1 The Classical Regime

Nuclear fusion are mostly about atomic nuclei, as electrons do not play a role in nuclear fusion. In fact, at places where temperature is high enough for nuclear fusion to happem, the atoms are completely ionized anyway. Atomic nuclei are all positively charged ions. For those nuclei to fuse together, they must come close enough to each other and overcome the *Coulomb barrier*.

Suppose two ions of species A and B have some kinetic energy. At the Coulomb barrier r_C, the electrostatic potential energy is given by

$$E = \frac{Z_A Z_B e^2}{4\pi\epsilon_0 r_C},$$

where Z_A and Z_B denotes the atomic number of the two ions. To make nuclear fusion happen, the kinetic energy must be comparable to this potential energy. The Coulomb barrier is about $r_C \sim 10^{-15}\,\mathrm{m}$, which is comparable to the size of a nucleus. An atom with shells of electrons have typical radius of $\sim 10^{-10}\,\mathrm{m}$. If we let both ions be hydrogen nucleus, $Z_A = Z_B = 1$, and plug in $r_C = 10^{-15}\,\mathrm{m}$, the typical kinetic energy of an ion is about $1\,\mathrm{MeV}$. Thermodynamically, this corresponds to a temperature of $10^{10}\,\mathrm{K}$, which is a thousand times higher than $T \sim 10^7\,\mathrm{K}$ in the Sun's core. Boltzmann distribution says that the probability of finding a particle in energy E is proportional to

$$f = e^{-(E-\mu)/kT}.$$

How common would a particle to have an $E \approx 10^3\,kT$? Compared to typical thermal energy $E = kT$, it is e^{-1000} times less unlikely to find $E = 10^3\,kT$. From all protns in the observable universe, none of them would fuse at the Sun's core temperature if the probability is like e^{-1000}.

This tells us that the "classical formulation" of fusion does not work. However, we know that nuclear fusion is certainly happening in the Sun's core by many evidences: change in composition, detection of neutrinos, etc. Indeed, classical mechanics and electromagnetism can never explain nuclear fusion. It is a result of both quantum mechanics (how to make fusion possible) and relativity (how to account for the release of energy). So now we will turn to quantum mechanics.

3.1.2 Coulomb Penetration

To understand quantum tunneling, consider the one-dimensional finite square well for a bound particle ($0 < E < V_0$),

$$V(x) = \begin{cases} 0, & 0 < x < L, \\ V_0 & \text{otherwise.} \end{cases}$$

If the system is classical, the particle can only move in regions where $E < V(x)$, i.e. inside the well. There is no way to cross the boundaries. However, in quantum mechanics, as along as the potential is not infinite, there will always be some probability for the particle to be in the classically forbidden regions. Such "leaking" of the particle is called the **quantum tunneling effect**.

We will solve the time-independent Schrödinger equation for this potential,

$$\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x)\right]\psi(x) = E\psi(x) \quad\Longrightarrow\quad \frac{\partial^2\psi}{\partial x^2} = \frac{2m[V(x)-E]}{\hbar^2}\psi.$$

Here $\psi(x)$ is the wavefunction for the particle. The probability density of finding a particle at position x is given by $|\psi(x)|^2$, and $\int |\psi|^2\, dx = 1$ over all space. Inside the well, $V(x) = 0$, so the Schrödinger equation reduces to

$$\frac{\partial^2 \psi}{\partial x^2} = -\frac{2mE}{\hbar^2}\psi.$$

Since $E > 0$, this is a simple harmonic oscillator equation. The general solution is

$$\psi(x) = A\sin(kx) + B\cos(kx), \quad \text{where} \quad k = \frac{\sqrt{2mE}}{\hbar}.$$

Outside the well, $V(x) = V_0$, the Schrödinger equation becomes

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{2m(V_0 - E)}{\hbar^2}\psi.$$

Since at the beginning we assume a bound state, $E < V_0$, the coefficient of ψ on the RHS is positive. The general solutions are exponentials:

$$\psi(x) = Ce^{x/x_0} + De^{-x/x_0}, \quad \text{where} \quad x_0 = \frac{\hbar}{\sqrt{2m(V_0 - E)}}.$$

The wavefunction cannot blow up at positive or negative infinities, or otherwise it is not normalizable. For $x < 0$, it means $D = 0$, and for $x > 0$, it means $C = 0$. There are also continuity conditions for ψ and its derivatives that matches the integration constants A, B, C, and D. We will not do it here, since it is available in most quantum mechanics textbooks. All we need to know is that the wavefunction decays like $\psi \sim e^{-x/x_0}$ outside the well, assuming $x > 0$. For example, the probability of finding the particle $5x_0$ away outside te well is $\sim e^{-10}$.

Applying this to the Coulomb potential inside stars, the time-independent Schrödinger equation is

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{2m(Z_A Z_B e^2/4\pi\epsilon_0 x - E)}{\hbar^2}\psi.$$

Note that the potential is no longer constant but scales like $1/x$. The finite square well analysis still applies: for $x > x_C$, the particle has some sort of oscillatory wavefunction, and for $x < x_C$, the wavefunction decays. To get the approximate probability of finding the particle at $x < x_C$, it is convenient to make a WKB approximation. That is, suppose that the potential is slowly changing. Then over a small δx, the potential behaves like a finite square well. The wavefunction changes like

$$\frac{\psi(x)}{\psi(x + \delta x)} \approx e^{-\delta x\sqrt{2m[V(x)-E]}/\hbar}$$

when $E < V(x)$, which is the classically forbidden region. Assume that nuclear fusion happens when the two nuclei are on top of each other ($x \to 0$). The ratio of the wavefunction at $x = 0$ to $x = x_C$ is

$$\frac{\psi(0)}{\psi(x_C)} = \frac{\psi(0)}{\psi(\delta x)} \cdot \frac{\psi(\delta x)}{\psi(2\delta x)} \cdots \frac{\psi(x_C - \delta x)}{\psi(x_C)} \approx e^{-\sum_i \delta x\sqrt{2m[V(i\delta x)-E]/\hbar}}.$$

The sum in the exponents is a Riemann sum, which can be approximated as an integral,

$$\begin{aligned}
\sum_i \delta_x \frac{\sqrt{2m[V(i\delta x) - E]}}{\hbar} &\approx \int_0^{x_C} \frac{\sqrt{2m[V(x) - E]}}{\hbar}\, dx \\
&= \int_0^{x_C} \frac{\sqrt{2m(Z_A Z_B e^2/4\pi\epsilon_0 x - E)}}{\hbar}\, dx \\
&= \int_0^{x_C} \frac{\sqrt{2mE(x_C/x - 1)}}{\hbar}\, dx \quad \Longleftarrow \quad \text{using } E = \frac{Z_A Z_B e^2}{4\pi\epsilon_0 x_C} \\
&= \frac{\sqrt{2mE}}{\hbar}\int_0^{x_C} \sqrt{\frac{x_C}{x} - 1}\, dx \\
&= \frac{\sqrt{2mE}}{\hbar}\left(\frac{\pi}{2}x_C\right).
\end{aligned}$$

Note that the final integral is exact. Plugging this integral into the exponential,

$$\frac{\psi(0)}{\psi(x_C)} \approx e^{-\pi x_C\sqrt{2mE}/2\hbar}.$$

Squaring it to convert to the ratio of probabilities:

$$\frac{\mathcal{P}(0)}{\mathcal{P}(x_C)} = \frac{|\psi(0)|^2}{|\psi(x_C)|^2} \approx e^{-\pi x_C \sqrt{2mE}/\hbar}.$$

Now express x_C back in terms of E using $x_C = Z_A Z_B e^2/4\pi\epsilon_0 E$,

$$\frac{\mathcal{P}(0)}{\mathcal{P}(x_C)} \approx e^{-\pi Z_A Z_B e^2 \sqrt{2mE}/4\pi\epsilon E\hbar} = e^{-\sqrt{2(\pi\alpha Z_A Z_B)^2 mc^2/E}},$$

where $\alpha = e^2/4\pi\epsilon_0\hbar c$ is the fine structure constant. It is convenient to write in an even more compact form: the probability of tunneling through is

$$\frac{\mathcal{P}(0)}{\mathcal{P}(x_C)} \approx e^{-\sqrt{E_G/E}},$$

where

$$E_G = 2(\pi\alpha Z_A Z_B)^2 m_r c^2 \tag{3.1}$$

is the **Gamow energy** and $m_r = m_A m_B/(m_A + m_B)$ is the reduced mass. The Gamow energy is proportional to the product of charges squared, so it characterizes the strength of the Coulomb force between two particles. For a large Gamov energy (hence a large Coulomb repulsion), the probability of tunneling through is small as expected. For two protons, the Gamow energy is $E_G \approx 493\,\mathrm{keV}$. Recall that the typical temperature in the Sun's core is $\sim 10^7\,\mathrm{K}$, or $kT \sim 1\,\mathrm{keV}$. Thanks to quantum mechanics, the probability reduces from e^{-1000} (in the classical case) to $e^{-\sqrt{E_G/E}} \sim e^{-\sqrt{500}} \approx 10^{-10}$. This is a huge difference. The probability is still small, but nuclear fusion is no longer impossible. After all, this is the probability of tunneling through for just two particles. Think about the particle number of the order of Avogadro's number $N \sim 10^{23}$.

3.2 Rate of Nuclear Fusion

3.2.1 Reaction Cross-Section and Astrophysical S-factor

The probability of fusion is a product of the following three factors:

1. The probability of a close, head-on encounter: this is proportional to the square of the de Broglie wavelength, $\mathcal{P} \propto \lambda_{\mathrm{dB}}^2 \propto 1/E$.
2. The probability of Coulomb penetration: this is what last section is all about, $\mathcal{P} \propto e^{-\sqrt{E_G/E}}$.
3. The probability of nuclear fusion while two nuclei are next to each other: that is, even though you bring two nuclei together, they would not necessarily fuse together. This probability depends on the type of nuclear reaction, $\mathcal{P} \propto \mathcal{P}_{\mathrm{nuc}}$.

Multiplying them together, the total probability is

$$\mathcal{P} \propto n\lambda_{\mathrm{dB}}^2 \times \mathcal{P}_{\mathrm{nuc}} e^{-\sqrt{E_G/E}}.$$

We can think of the target size of a nucleus as **nuclear reaction cross-section**, which is proportional to the three probabilities,

$$\sigma \propto \lambda_{\mathrm{dB}}^2 \mathcal{P}_{\mathrm{nuc}} e^{-\sqrt{E_G/E}} \propto \frac{1}{E}\mathcal{P}_{\mathrm{nuc}} e^{-\sqrt{E_G/E}}.$$

Traditionally, the cross-section is written as

$$\sigma(E) = \frac{S(E)}{E} e^{-\sqrt{E_G/G}}.$$

where $S(E)$ is called the **astrophysical S-factor**. It includes the probability of certain nuclear reaction and all other constants of proportionality. The S-factor has a dimension of [energy $\times$ area]. It is usually a constant independent of energy for non-resonant reactions, but it has a huge peak like a delta function for resonant reactions. Resonant reactions refer to reactions with the total energy (kinetic plus rest mass) of the reaction being one of the quasi-stationary states of the compound nucleus. It is analogous to electron excitation, where you have the energy of a photon just right at the required energy. The probability or rate is much higher than non-resonant cases.

3.2.2 Reaction Rate

Assume that the nuclei are non-relativistic, so $v_r = \sqrt{2E/m_r}$. Nuclear fusion in stars are always taken to be non-relativistic. If they are relativistic, the high energies will cause the nuclei to disintegrate into protons and electrons, or neutrons at high density. The nuclear reaction rate for *one* nucleus of species A with energy E and a target species B is given by

$$\text{rate} = n_B v_r(E)\sigma(E).$$

Plugging in $\sigma(E)$ and $v_r(E)$, and integrate over all energy, we obtain the total rate:

$$\text{rate} = n_B \int_0^\infty \sqrt{\frac{2E}{m_r}} \frac{S(E)}{E} \mathcal{P}(E) e^{-\sqrt{E_G/E}}\, dE,$$

where $\mathcal{P}(E)$ is the probability of having energy E. This is the rate per nucleus A, so the rate per unit volume acquires an additional factor of the number density n_A,

$$R_{AB} = \frac{n_A n_B}{1+\delta_{AB}} \int_0^\infty \sqrt{\frac{2E}{m_r}} \frac{S(E)}{E} \mathcal{P}(E) e^{-\sqrt{E_G/E}}\, dE.$$

Here R_{AB} denotes the rate per volume of nuclear reaction between species A and B. δ_{AB} is the Kronecker delta to prevent double-counting. Now we know everything except the probability of having energy E. In a classical system in thermal equilibrium, this probability is given by the Maxwell-Boltzmann distribution (see Appendix A.4)

$$\mathcal{P}(E) = \frac{2\pi\sqrt{E}}{(\pi kT)^{3/2}} e^{-E/kT}.$$

Putting everything together,

$$R_{AB} = \frac{n_A n_B}{1+\delta_{AB}} \left(\frac{8}{\pi m_r}\right)^{1/2} \left(\frac{1}{kT}\right)^{3/2} \int_0^\infty S(E) e^{-E/kT - \sqrt{E_G/E}}\, dE. \tag{3.2}$$

The integrand has a peak at energy

$$E_{\text{peak}} = \left[\frac{E_G (kT)^2}{4}\right]^{1/3}.$$

Taylor expand the exponent around E_{peak} gives

$$e^{-E/kT - \sqrt{E_G/E}} \approx e^{-3(E_G/4kT)^{1/3}} e^{-[(E-E_{\text{peak}})/(\Delta/2)]^2},$$

where the width of this Gaussian is

$$\Delta = \frac{4}{3^{1/2} 2^{1/3}} E_G^{1/6} (kT)^{5/6} \approx \frac{4}{3^{1/2}} (E_{\text{peak}} kT)^{1/2}.$$

For non-resonant reactions, $S(E)$ is constant and we can pull it out from the integral of R_{AB},

$$\begin{aligned}
\int_0^\infty S(E) e^{-E/kT - \sqrt{E_G/E}}\, dE &\approx S(E_{\text{peak}}) e^{-3(E_G/4kT)^{1/3}} \int_0^\infty e^{-[(E-E_{\text{peak}})/(\Delta/2)]^2}\, dE \\
&\approx S(E_{\text{peak}}) e^{-3(E_G/4kT)^{1/3}} \int_{-\infty}^\infty e^{-[(E-E_{\text{peak}})/(\Delta/2)]^2}\, dE \\
&\approx S(E_{\text{peak}}) e^{-3(E_G/4kT)^{1/3}} \frac{\Delta\sqrt{\pi}}{2}.
\end{aligned}$$

In the second approximation, the lower limit of the integral is extended to negative infinity to evaluate the Gaussian integral exactly. The final step is to put the definition of Δ, E_G, and $m_r = A_r m_p$ into the reaction rate,

$$\begin{aligned}
R_{AB} &= \frac{n_A n_B}{1+\delta_{AB}} \left(\frac{8}{\pi m_r}\right)^{1/2} \left(\frac{1}{kT}\right)^{3/2} \int_0^\infty S(E) e^{-E/kT - \sqrt{E_G/E}}\, dE \\
&= \frac{n_A n_B}{1+\delta_{AB}} \left(\frac{8}{\pi m_r}\right)^{1/2} \left(\frac{1}{kT}\right) S(E_{\text{peak}}) e^{-3(E_G/4kT)^{1/3}} \frac{\Delta\sqrt{\pi}}{2} \\
&= \frac{n_A n_B}{1+\delta_{AB}} \frac{8}{\sqrt{3}\pi\alpha c} \frac{1}{m_r Z_A Z_B} \left(\frac{E_G}{4kT}\right)^{2/3} S(E_{\text{peak}}) e^{-3(E_G/4kT)^{1/3}} \\
&\approx 6.48 \times 10^{-24} \frac{n_A n_B}{(1+\delta_{AB}) A_r Z_A Z_B} \left(\frac{E_G}{4kT}\right)^{2/3} \frac{S(E_{\text{peak}})}{\text{keV} \cdot \text{barns}} e^{-3(E_G/4kT)^{1/3}} \text{m}^{-3}\,\text{s}^{-1}.
\end{aligned}$$

3.2.3 Scaling with Temperature

Long ago we mentioned that the rate of fusion is a strong function of temperature. Here we will discuss how this rate scales with temperature. To do this, we will focus on the logarithmic derivative,

$$\frac{d\ln R_{AB}}{d\ln T} = \frac{T}{R_{AB}}\frac{dR_{AB}}{dT}.$$

The temperature dependent terms in the fusion rate is

$$R_{AB} \propto T^{-2/3} e^{-3(E_G/4kT)^{1/3}}.$$

Taking the logarithm of both sides,

$$\ln R_{AB} = -\frac{2}{3}\ln T - 3\left(\frac{E_G}{4kT}\right)^{1/3} + \text{const}.$$

Then differentiate with respect to $\ln T$, and using the fact that $dT/d\ln T = T$,

$$\frac{d\ln R_{AB}}{d\ln T} = -\frac{2}{3} + \left(\frac{E_G}{4k}\right)^{1/3}\frac{1}{T^{4/3}}\frac{dT}{d\ln T} = -\frac{2}{3} + \left(\frac{E_G}{4k}\right)^{1/3}\frac{T}{T^{4/3}} = -\frac{2}{3} + \left(\frac{E_G}{4kT}\right)^{1/3}. \tag{3.3}$$

If the temperature is 10^7 K and $E_G = 493\,\text{keV}$ for fusing two protons, this logarithmic derivative is about $d\ln R_{AB}/d\ln T \approx 4$. This is the slowest step in the most important reaction chain in the Sun. As the Sun gets hotter, the reaction rate is growing to the fourth power of temperature.

3.3 Hydrogen Fusion

3.3.1 Proton-Proton Chain

Neglecting constant of proportionality, the fusion rate is a function of temperature, the S-factor $S(E)$, and the Gamow energy,

$$R_{AB} \propto \frac{n_A n_B}{(1+\delta_{AB})m_r Z_A Z_B} S(E_{\text{peak}}) \left(\frac{E_G}{4kT}\right)^{2/3} e^{-3(E_G/4kT)^{1/3}}.$$

Define $\zeta \equiv (E_G/4kT)^{1/3}$, we can write

$$R_{AB} \propto \frac{n_A n_B}{(1+\delta_{AB})m_r Z_A Z_B} S(E_{\text{peak}})\zeta^2 e^{-3\zeta}.$$

In the Sun, the **proton-proton chain** is the most important hydrogen fusion chain. There are three reactions in this chain:

$$\begin{aligned} p + p &\longrightarrow {}^2\text{D} + e^+ + \nu_e \\ p + {}^2\text{D} &\longrightarrow {}^3\text{He} + \gamma \\ {}^3\text{He} + {}^3\text{He} &\longrightarrow {}^4\text{He} + 2p \end{aligned}$$

Some other reactions can replace the third step. Like in chemistry, there is always some rate-limiting step that proceeds more slowly than other steps. It is this step that controls the over all reaction rate of the chain. Here are the S-factor and Gamow energy of each reaction step (reference: `https://arxiv.org/pdf/1004.2318.pdf`).

$$\begin{aligned} p + p &\longrightarrow {}^2\text{D} + e^+ + \nu_e & S(E_{\text{peak}}) &\approx 4\times 10^{-22}\,\text{keV barns} & E_G &\approx 493\,\text{keV} \\ p + {}^2\text{D} &\longrightarrow {}^3\text{He} + \gamma & S(E_{\text{peak}}) &\approx 2\times 10^{-4}\,\text{keV barns} & E_G &\approx 740\,\text{keV} \\ {}^3\text{He} + {}^3\text{He} &\longrightarrow {}^4\text{He} + 2p & S(E_{\text{peak}}) &\approx 5\times 10^{3}\,\text{keV barns} & E_G &\approx 23.7\,\text{MeV} \end{aligned}$$

Notice that the S-factor of the first step is extremely small compared to the other two. This is because there is a weak nuclear reaction (this "weak" is the fundamental interaction, not strength) which has a relatively small cross-section. Thus, a low probability incorporated in $S(E)$. Plug in E_G, $S(E_{\text{peak}})$, and other constants into the reaction rate,

$$R_{pp} \propto \frac{n_p^2/2}{m_p/2} \times 3\times 10^{-27}, \quad R_{p\text{D}} \propto \frac{n_p n_{\text{D}}}{2m_p/3} \times 2\times 10^{-10}, \quad R_{^3\text{He}^3\text{He}} \propto \frac{n_{^3\text{He}}^2/2}{3m_p/2\times 2^2} \times 7\times 10^{-22}.$$

The high Gamow energy causes the last step to have much lower rate than the second step, but a further small S-factor of the first step makes it the rate-limiting step. Therefore, the rate of the first step determines the rate of the proton-proton chain. At $T \approx 1.5 \times 10^7\,\mathrm{K}$, $kT \approx 1.3\,\mathrm{keV}$,

$$\begin{aligned}
R_{pp} &\approx 6.48 \times 10^{-24} \frac{n_A n_B}{A_r Z_A Z_B} \left(\frac{E_G}{4kT}\right)^{2/3} \frac{S(E_{\text{peak}})}{\text{keV} \cdot \text{barns}} e^{-3(E_G/4kT)^{1/3}} \\
&\approx 6.48 \times 10^{-24} \frac{(X\rho/m_p)^2/2}{0.5 \cdot 1 \cdot 1} \left(\frac{493}{4 \cdot 1.3}\right)^{2/3} (4 \times 10^{-22}) e^{-3(493/4\cdot 1.3)^{1/3}} \\
&\approx 2.21 \times 10^4 X^2 \rho^2
\end{aligned}$$

We already know the temperature dependence of the proton-proton chain is $\propto T^4$, so the reaction rate can be expressed as

$$R_{pp} \approx 4.37 \times 10^{-25} X^2 \rho^2 \left(\frac{T}{1.5 \times 10^7\,\mathrm{K}}\right)^4.$$

In the chain, each ^{4}He made produces $Q = 26.7\,\mathrm{MeV}$. It takes two proton-proton fusion (the first step) to produce one ^{4}He. Thus, the net energy production per proton-proton fusion is actually $Q/2$. The rate of energy generation per volume is

$$\epsilon_{pp} = \frac{1}{2} Q R_{pp} \approx 10^{-37} X^2 \rho^2 \left(\frac{T}{1.5 \times 10^7\,\mathrm{K}}\right)^4 \mathrm{J/m^3\,s}.$$

Note that since $E_G \propto Z_A^2 Z_B^2$ and proton-proton fusion has the lowest Z_A and Z_B, the Gamow energy for any nuclear reaction is at least 493 keV. The logarithmic derivative $d\ln R_{AB}/d\ln T$ is proportional to $\sqrt[3]{E_G}$, so the temperature dependence of other nuclear reactions are much steeper. This will have significant impact on stellar structure and evolution.

3.3.2 CNO Cycle

The **CNO** cycle consists of the CN cycle (left) and the NO cycle (right):

$$\begin{aligned}
{}^{12}\mathrm{C} + p &\longrightarrow {}^{13}\mathrm{N} + \gamma \\
{}^{13}\mathrm{N} &\longrightarrow {}^{13}\mathrm{C} + e^+ + \nu_e \\
{}^{13}\mathrm{C} + p &\longrightarrow {}^{14}\mathrm{N} + \gamma \\
{}^{14}\mathrm{N} + p &\longrightarrow {}^{15}\mathrm{O} + \gamma \\
{}^{15}\mathrm{O} &\longrightarrow {}^{15}\mathrm{N} + e^+ + \nu_e \\
{}^{15}\mathrm{N} + p &\longrightarrow {}^{12}\mathrm{C} + {}^4\mathrm{He}
\end{aligned}
\qquad\qquad
\begin{aligned}
{}^{14}\mathrm{N} + p &\longrightarrow {}^{15}\mathrm{O} + \gamma \\
{}^{15}\mathrm{O} &\longrightarrow {}^{15}\mathrm{N} + e^+ + \nu_e \\
{}^{15}\mathrm{N} + p &\longrightarrow {}^{16}\mathrm{O} + \gamma \\
{}^{16}\mathrm{O} + p &\longrightarrow {}^{17}\mathrm{F} + \gamma \\
{}^{17}\mathrm{F} &\longrightarrow {}^{17}\mathrm{O} + e^+ + \nu_e \\
{}^{17}\mathrm{O} + p &\longrightarrow {}^{14}\mathrm{N} + {}^4\mathrm{He}
\end{aligned}$$

Note that the net production of both cycles is a ^{4}He, while ^{12}C and ^{14}N are recycled. This means these nuclei act like catalysts to fuse hydrogen into helium. Due to a higher charge of O ($Z = 8$) than C ($Z = 6$) and N ($Z = 7$), it has a higher Coulomb barrier and thus the NO cycle requires a higher temperature. At $T \lesssim 3 \times 10^7\,\mathrm{K}$, only the CN cycle is present. Above this temperature, both cycle are present. The slowest step in the CNO cycle is

$${}^{14}\mathrm{N} + p \longrightarrow {}^{15}\mathrm{O} + \gamma,$$

which is the fourth step of the CN cycle and the first step of the NO cycle. Since this step is the slowest, most of the initially present C, N, and O will end up as ^{14}N. An over-abundance of ^{14}N then indicates that a star has undergone the CNO cycle. It can be shown that the energy generation rate per unit volume of the CNO cycle is

$$\epsilon_{\text{CNO}} \propto X_1 X_{\text{CNO}} \rho^2 T^{18}.$$

where X_1 is the hydrogen abundance. The power of the temeprature may vary from 16 to 20. This reflects a large Gamow energy according to (3.3), and hence a large Coulomb barrier in the cycle.

For main sequence stars with $M \lesssim 1.2\,M_\odot$, hydrogen burning is dominated by the proton-proton chain. For $M \gtrsim 1.2\,M_\odot$, those stars have a central temperature $T \gtrsim 1.8 \times 10^7\,\mathrm{K}$, and hydrogen burning is dominated by the CNO cycle. Because of the steep temperature dependence, the CNO cycle is only present in a small region of a star: it release energy in a small volume in the core of a star. In other words, $L(r)/m(r)$ is large, making convective energy transport possible in the core.

3.4 Further Fusion Reactions

Before we proceed to further fusion reactions, we need to know the binding energy per nucleon as a function of number of nucleons in the nucleus. The binding energy is the energy needed to break the nucleus, so a higher binding energy means a more stable nucleus. The general trend of binding energy vs number of nucleons is the following: it increases from ^{1}H to ^{56}Fe, and decreases from then on. Thus, the isotopes around ^{56}Fe are the most stable ones. Lighter elements tends to fuse into heavier elements until ^{56}Fe if the temperature is high enough. Fusion among these elements are exothermic.

Another interesting feature about nuclear binding energy is the concept of **magic numbers**. They are similar to number of electrons in a filled shell in an electron configuration. In nuclear physics, thee numbers are $2, 8, 20, 28, 50, 82, 126$. ^{4}He is a *doubly* magic nucleus because both its proton number and neutron number are 2. This gives ^{4}He a particularly stable configuration, so its binding energy is much higher than its neighboring nucleus (such as hydrogen or lithium). Another useful nucleus that is doubly magic is ^{16}O, with 8 protons and 8 neutrons.

3.4.1 Helium Burning (Triple-Alpha Reaction)

The **triple-alpha reaction** fuses three ^{4}He (also called α particles) into one ^{12}C. It involves three steps, two of which are reversible:

$$\begin{aligned} {}^4\mathrm{He} + {}^4\mathrm{He} &\longleftrightarrow {}^8\mathrm{Be} \\ {}^8\mathrm{Be} + {}^4\mathrm{He} &\longleftrightarrow {}^{12}\mathrm{C}^* \\ {}^{12}\mathrm{C}^* &\longrightarrow {}^{12}\mathrm{C} + 2\gamma \end{aligned}$$

Three ^{4}He fuses into a ^{12}C. The first step is reversible because ^{8}Be is unstable ($\tau \approx 1.2 \times 10^{-16}\,\mathrm{s}$) and it may decay back to two ^{4}He. Fortunately, the lifetime is long enough for it to encounter another ^{4}He and undergo the second step. Moreover, the first step is endothermic, which needs an energy of $92\,\mathrm{keV}$. This energy falls within the range of $E_{\mathrm{peak}} \pm \Delta$ at $T \approx 10^8$. To see that, compute the Gamow energy and E_{peak}:

$$E_G({}^4\mathrm{He} + {}^4\mathrm{He}) = \left(\frac{Z_{\mathrm{He}} Z_{\mathrm{He}}}{Z_p Z_p}\right)^2 \frac{m_{r,\mathrm{He}}}{m_{r,p}} E_G(p+p) = 2^2 \cdot 2^2 \cdot \frac{2}{0.5} E_G(p+p) \approx 64 \times 493\,\mathrm{keV} \approx 31.6\,\mathrm{MeV}.$$

At $T \approx 10^8\,\mathrm{K}$, $kT \approx 10\,\mathrm{keV}$, the peak energy is

$$E_{\mathrm{peak}} = \left[\frac{E_G (kT)^2}{4}\right]^{1/3} \approx 92\,\mathrm{keV}.$$

This is great, because at $T \approx 10^8\,\mathrm{K}$, there is a *resonance* between the peak energy and the reaction energy. Recall that the S-factor is significantly enhanced, and so does the fusion rate. The equilibrium number density of ^{8}Be can be obtained by the Saha equation. The equilibrium condition is

$$2\mu_{\mathrm{He}} = \mu_{\mathrm{Be}}.$$

Using the chemical potential and quantum concentration

$$\mu = mc^2 + kT \ln\left(\frac{n}{g_s n_Q}\right), \quad n_Q = \left(\frac{2\pi m kT}{h^2}\right)^{3/2},$$

we have

$$n_{\mathrm{Be}} = n_{\mathrm{He}}^2 \left(\frac{h^2}{\pi m_{\mathrm{He}} kT}\right)^{3/2} e^{-(m_{\mathrm{Be}} - 2m_{\mathrm{He}})c^2/kT} = n_{\mathrm{He}}^2 \left(\frac{h^2}{\pi m_{\mathrm{He}} kT}\right)^{3/2} e^{-92\,\mathrm{keV}/kT}.$$

There is also a resonance in the second step. The second step fuses ^{8}Be and ^{4}He into an excited state (called the Hoyle state) of ^{12}C, denoted as ^{12}C*. It has $288\,\mathrm{keV}$ more energy than $^4\mathrm{He} + {}^8\mathrm{Be}$. The Gamow energy for the two nuclei is

$$E_G({}^4\mathrm{He} + {}^8\mathrm{Be}) = \left(\frac{Z_{\mathrm{He}} Z_{\mathrm{Be}}}{Z_p Z_p}\right)^2 \frac{m_{r,\mathrm{He,Be}}}{m_{r,p}} E_G(p+p) = 2^2 \cdot 4^2 \cdot \frac{(4 \cdot 8/12)}{0.5} \times E_G(p+p) \approx 341 \cdot 493\,\mathrm{keV} \approx 168\,\mathrm{MeV}.$$

At $T \approx 2 \times 10^8\,\mathrm{K}$, $kT \approx 17\,\mathrm{keV}$, the peak energy is

$$E_{\mathrm{peak}} = \left[\frac{E_G (kT)^2}{4}\right]^{1/3} \approx 230\,\mathrm{keV}.$$

The required energy 288 keV lies within the $1\Delta \approx 144\,\mathrm{keV}$ range. We can also find the equilibrium number density of $^{12}\mathrm{C}^*$ for this step by another Saha equation:

$$\begin{aligned} n_{\mathrm{C}^*} &= n_{\mathrm{He}} n_{\mathrm{Be}} \left(\frac{3}{2}\right)^{3/2} \left(\frac{h^2}{2\pi m_{\mathrm{He}} kT}\right)^{3/2} e^{-(m_{\mathrm{C}^*} - m_{\mathrm{He}} - m_{\mathrm{Be}})c^2/kT} \\ &= n_{\mathrm{He}} n_{\mathrm{Be}} \left(\frac{3}{2}\right)^{3/2} \left(\frac{h^2}{2\pi m_{\mathrm{He}} kT}\right)^{3/2} e^{-278\,\mathrm{keV}/kT}. \end{aligned}$$

From the Saha equation for the first step, we can plug in n_{Be} into this equation,

$$n_{\mathrm{C}^*} = n_{\mathrm{He}}^3 3^{3/2} \left(\frac{h^2}{2\pi m_{\mathrm{He}} kT}\right)^3 e^{-(92+278)\,\mathrm{keV}/kT} = n_{\mathrm{He}}^3 3^{3/2} \left(\frac{h^2}{2\pi m_{\mathrm{He}} kT}\right)^3 e^{-370\,\mathrm{keV}/kT}.$$

Through these reactions, most $^{12}\mathrm{C}^*$ will decay back to $^4\mathrm{He} + {}^8\mathrm{Be}$, but some of them will decay ($\tau \approx 1.8 \times 10^{-13}\,\mathrm{s}$) to the ground state $^{12}\mathrm{C}$, emitting 2γ. The rate of $^{12}\mathrm{C}$ formation is

$$R_{3\alpha} = \frac{dn_{\mathrm{C}}}{dt} = -\frac{dn_{\mathrm{C}^*}}{dt} = \frac{n_{\mathrm{C}^*}}{\tau} = n_{\mathrm{He}}^3 \frac{e^{3/2}}{\tau} \left(\frac{h^2}{2\pi m_{\mathrm{He}} kT}\right)^3 e^{-370\,\mathrm{keV}/kT}.$$

The total energy released per triple-alpha reaction is $Q = (m_{\mathrm{C}} - 3m_{\mathrm{He}})c^2 \approx 7.28\,\mathrm{MeV}$. This gives the energy generation rate:

$$\epsilon_{3\alpha} = QR_{3\alpha} \approx 51\, Y^3 \rho^3 \left(\frac{10^8\,\mathrm{K}}{T}\right)^3 e^{-43\times 10^8\,\mathrm{K}/T},$$

where Y is the mass fraction ($m_{\mathrm{He}} = Y\rho/4m_p$). The reference is set at $T = 10^8\,\mathrm{K}$. Finally, let's see how the triple-alpha reaction is sensitive to temperature. The logarithmic derivative is

$$\frac{d\ln R_{3\alpha}}{d\ln T} = -3 + 43\left(\frac{10^8\,\mathrm{K}}{T}\right).$$

At $T \approx 10^8$, this is approximately 40, so

$$R_{3\alpha} \propto Y^3 \rho^3 T^{40}.$$

Once $^{12}\mathrm{C}$ is formed, a further reaction can be done by

$$^{12}\mathrm{C} + {}^4\mathrm{He} \longrightarrow {}^{16}\mathrm{O} + \gamma.$$

The rate is comparable to the triple-alpha burning. Thus, we can say that the primary products of helium burning are $^{12}\mathrm{C}$ and $^{16}\mathrm{O}$.

3.4.2 Advanced Stages of Nuclear Burning

Carbon Burning Carbon burning requires a temperature of $T \gtrsim 6 \times 10^8\,\mathrm{K}$. There are three ways to fuse two carbon nuclei:

$$\begin{aligned} {}^{12}\mathrm{C} + {}^{12}\mathrm{C} &\longrightarrow {}^{20}\mathrm{Ne} + {}^4\mathrm{He} \\ &\longrightarrow {}^{23}\mathrm{Na} + p \\ &\longrightarrow {}^{24}\mathrm{Mg} + \gamma. \end{aligned}$$

The first one dominates over the other two, so the main product of carbon burning is $^{20}\mathrm{Ne}$, with minor products $^{23}\mathrm{Na}$ and $^{24}\mathrm{Mg}$.

Neon Burning Neon burning starts at $T \approx 1.2 \times 10^9\,\mathrm{K}$. Photons at this temperature are able to photodisintegrate $^{20}\mathrm{Ne}$ into $^{16}\mathrm{O}$ and $^4\mathrm{He}$. This $^4\mathrm{He}$ can fuse with another $^{20}\mathrm{Ne}$:

$$\begin{aligned} {}^{20}\mathrm{Ne} + \gamma &\longrightarrow {}^{16}\mathrm{O} + {}^4\mathrm{He} \\ {}^{20}\mathrm{Ne} + {}^4\mathrm{He} &\longrightarrow {}^{24}\mathrm{Mg} + \gamma. \end{aligned}$$

The major products are $^{16}\mathrm{O}$ and $^{24}\mathrm{Mg}$.

Oxygen Burning When $T \approx 1.5 \times 10^9$, oxygen burning starts. There are two ways to fuse oxygen, one is to directly produce ^{28}Si, the other is to produce ^{31}P. The latter proceeds to produce ^{32}S:

$$\begin{aligned} {}^{16}\mathrm{O} + {}^{16}\mathrm{O} &\longrightarrow {}^{28}\mathrm{Si} + {}^{4}\mathrm{Ne} \\ &\longrightarrow {}^{31}\mathrm{P} + p \\ {}^{31}\mathrm{P} + p &\longrightarrow {}^{32}\mathrm{S} + \gamma. \end{aligned}$$

THe first reaction dominates. The major products are ^{28}Si, with minor products ^{32}S and ^{31}P.

Silicon burning Silicon burning occurs at $T \approx 3 \times 10^9$. A ^{28}Si can be photodisintegrated and produce a ^{4}He, which can be fused with other elements, and get to further photodisintegration. The rate of photodisintegration and fusion almost reaches an equilibrium, but the overall trend is still towards the iron group:

$$\begin{aligned} {}^{28}\mathrm{Si} + \gamma &\longrightarrow {}^{24}\mathrm{Mg} + {}^{4}\mathrm{He} \\ {}^{28}\mathrm{Si} + {}^{4}\mathrm{He} &\longrightarrow {}^{32}\mathrm{S} + \gamma \\ {}^{32}\mathrm{S} + {}^{4}\mathrm{He} &\longrightarrow {}^{36}\mathrm{Ar} + \gamma \end{aligned}$$

followed by ^{40}Ca, ^{44}Ti, ^{48}Cr, ^{52}Fe, and ^{56}Ni, each requiring one ^{4}He. The final product ^{56}Ni can decay into ^{56}Fe.

Table 3.1 lists all types of nuclear fusion discussed above. Deuterium burning is separated from the intermediate step in hydrogen burning because it is available in even lower temperatures. These deuterium nuclei as reactants were produced by Big Bang nucleosynthesis.

Burning stage	Reactions	Typical T (K)	Products
Deuterium burning	$^{2}\mathrm{D} + p \longrightarrow {}^{3}\mathrm{He} + \gamma$	10^6	^{3}He
Hydrogen burning	$4p \longrightarrow {}^{4}\mathrm{He} + 2\gamma + 2e^+ + 2\nu_e$	10^7	^{4}He
	CN cycle	$\lesssim 3 \times 10^7$	
	NO cycle	$\gtrsim 3 \times 10^7$	
Helium burning	$3\alpha \longrightarrow {}^{12}\mathrm{C} + 2\gamma$	10^8	^{12}C, ^{16}O
	$^{12}\mathrm{C} + \alpha \longrightarrow {}^{16}\mathrm{O} + \gamma$		
Carbon burning	$^{12}\mathrm{C} + {}^{12}\mathrm{C} \longrightarrow {}^{20}\mathrm{Ne} + \alpha$	6×10^8	^{20}Ne
	$^{12}\mathrm{C} + {}^{12}\mathrm{C} \longrightarrow {}^{23}\mathrm{Na} + p$		some ^{23}Na and ^{24}Mg
	$^{12}\mathrm{C} + {}^{12}\mathrm{C} \longrightarrow {}^{24}\mathrm{Mg} + \gamma$		
Neon burning*	$^{20}\mathrm{Ne} + \gamma \longrightarrow {}^{16}\mathrm{O} + \alpha$	1.2×10^9	^{16}O, ^{24}Mg
	$^{20}\mathrm{Ne} + \alpha \longrightarrow {}^{24}\mathrm{Mg} + \alpha$		
Oxygen burning	$^{16}\mathrm{O} + {}^{16}\mathrm{O} \longrightarrow {}^{28}\mathrm{Si} + \alpha$	1.5×10^9	^{28}Si
	$^{16}\mathrm{O} + {}^{16}\mathrm{O} \longrightarrow {}^{32}\mathrm{S} + \gamma$		some ^{31}P and ^{32}S
Silicon burning*	$^{28}\mathrm{Si} + \gamma \longrightarrow {}^{24}\mathrm{Mg} + \alpha$	3×10^9	^{32}S, ^{36}Ar, ^{40}Ca,
	$^{28}\mathrm{Si} + \alpha \longrightarrow {}^{32}\mathrm{S} + \gamma$		^{44}Ti, ^{48}Cr, ^{52}Fe,
	$^{32}\mathrm{S} + \alpha \longrightarrow {}^{36}\mathrm{Ar} + \gamma$		^{54}Fe, ^{56}Fe, ^{56}Ni,
	α-captures up to ^{56}Ni		(gradual shift to iron-peak)

Table 3.1: A list of nuclear burning until iron-56. ^{4}He is denoted by α, and $*$ indicates a photodisintegration.

Stars with initial mass $M \lesssim 0.5\, M_\odot$ cannot undergo helium burning. At the end of their lives they will leave behind a helium white dwarf, but these low-mass stars have main sequence lifetime greater than the age of the universe. Helium white dwarfs observed now are results of binary interactions. Stars with initial mass $0.5 \lesssim M \lesssim 8\, M_\odot$ cannot undergo carbon burning. They will end up as carbon-oxygen white dwarfs. Stars with initial mass $8 \lesssim M \lesssim 10\, M_\odot$ will undergo up to carbon burning, leaving an oxygen-neon white dwarf or a neutron star. For $M \gtrsim 10\, M_\odot$, they will go all the way to silicon burning.

4 Stellar Structure and Evolution

We now have obtained many equations from previous chapters. There are four main equations for stellar structure and evolution:

- Mass conservation:
$$\frac{dm}{dr} = 4\pi r^2 \rho(r).$$
- Equation of hydrostatic equilibrium:
$$\frac{dP}{dr} = -\frac{Gm(r)\rho(r)}{r^2}.$$
- Methods of energy generation include nuclear burning, expansion work, etc. In this notes, we only consider nuclear burning:
$$\frac{dL_{\rm nuc}}{dr} = 4\pi r^2 \epsilon_{\rm nuc}(r).$$
- Equation of energy transport, particularly radiative diffusion:
$$F_{\rm rad}(r) = -\frac{4acT^3}{3\kappa\rho}\frac{dT}{dr}.$$

4.1 Scaling Relations

First, let's make order-of-magnitude estimates on how $L_{\rm rad} = 4\pi r^2 F_{\rm rad}$ and $L_{\rm nuc}$ depends on stellar mass M, radius R, or mean molecular weight μ.

4.1.1 Scaling Relations for Low-Mass Stars

Assumptions:

1. Assume the star is composed of classical ideal gas, $T \propto \mu M/R$, and the composition is uniform throughout the star.
2. Energy is transported solely by radiative diffusion.
3. For low-mass stars, its temperature should be low enough so that the opacity is dominated by bound-free opacity, which is given by Kramer's law: $\kappa \propto \rho T^{-3.5}$.
4. Energy is generated via proton-proton chain: $\epsilon_{pp} \propto \rho^2 T^4$.

Starting with the equation of radiative diffusion,
$$F(r) \propto -\frac{1}{\kappa\rho} T^3 \frac{dT}{dr}.$$
There are several components to be approximated:
$$r \sim R, \quad \rho \sim \frac{M}{R^3}, \quad \frac{dT}{dr} \sim -\frac{T_c}{R},$$
where T_c is the central temperature. (The subscript will be dropped later). Using Kramer's law, $\kappa \propto \rho T^{-3.5}$, the radiation luminosity is
$$L_{\rm rad} = 4\pi r^2 F \propto -\frac{r^2 T^3}{\kappa\rho}\frac{dT}{dr} \propto -\frac{r^2 T^{6.5}}{\rho^2}\frac{dT}{dr} \propto \frac{R^2 T^{6.5}}{(M/R^3)^2}\frac{T}{R} \propto \frac{R^7}{M^2} T^{7.5}.$$

By the assumption of classical ideal gas, $T \propto \mu M/R$, so

$$L_{\rm rad} \propto \frac{\mu^{7.5} M^{5.5}}{R^{0.5}}. \tag{4.1}$$

Total power generated by nuclear burning,

$$L_{\rm nuc}(r) = \int_0^r \epsilon_{\rm nuc}(r) 4\pi r^2\, dr \sim \epsilon_{\rm nuc} R^3 \propto \rho^2 T^4 R^3,$$

where we have used the assumption of energy generation by proton-proton chain, $\epsilon_{\rm nuc} = \epsilon_{pp} \propto \rho^2 T^4$. Again, plugging in $\rho \sim M/R^3$ and $T \sim \mu M/R$, the nuclear luminosity is

$$L_{\rm nuc} \propto \frac{\mu^4 M^6}{R^7}. \tag{4.2}$$

Main sequence stars are in quasi-steady state, where $L_{\rm nuc} \approx L_{\rm rad}$. Setting them equal and solve for R gives

$$\frac{\mu^4 M^6}{R^7} \propto \frac{\mu^{7.5} M^{5.5}}{R^{0.5}} \quad \Longrightarrow \quad R \propto \mu^{-7/13} M^{1/13}.$$

Substitute the mass-radius relation into luminosity,

$$L_{\rm rad} \propto \frac{\mu^{7.5} M^{5.5}}{R^{0.5}} \propto \mu^{101/13} M^{71/13}.$$

This is somewhat steeper from the observed results $L \propto M^{4.5}$ if we exclude fully convective stars with $M \lesssim 0.4\, M_\odot$. From Stefan-Boltzmann law (1.2), $L = 4\pi R^2 \sigma T_{\rm eff}^4$, it is also possible to derive the effective temperature:

$$L = 4\pi R^2 \sigma T_{\rm eff}^4 \quad \Longrightarrow \quad T_{\rm eff} \propto \mu^{115/52} M^{69/52}.$$

Many scalings above are not very accurate. This is because many assumptions made have limitations: we completely ignored convection, and the composition is usually not uniform throughout the star. But the main point is that stars with higher mass have higher luminosity, and as a star ages its luminosity also increases with μ.

4.1.2 Scaling Relations for High-Mass Stars

Assumptions:

1. Assume the star is composed of classical ideal gas ($T \propto \mu M/R$) with energy transported by radiative diffusion.
2. The opacity is dominated by Thomson scattering opacity, which is roughly a constant throughout the star.
3. Energy is generated by the CNO cycle, $\epsilon_{\rm CNO} \propto \rho^2 T^{18}$.

$$L_{\rm rad} \propto \mu^4 M^3. \tag{4.3}$$

$$L_{\rm nuc} \propto \frac{\mu^{18} M^{20}}{R^{21}}. \tag{4.4}$$

In steady state, $L_{\rm nuc} \approx L_{\rm rad}$, we can solve for R,

$$R \propto \mu^{2/3} M^{17/21}.$$

By Stefan-Boltzmann law,

$$L = 4\pi R^2 \sigma T_{\rm eff}^4 \quad \Longrightarrow \quad T_{\rm eff} \propto \mu^{2/3} M^{29/84}.$$

4.2 Polytropic Models

4.2.1 Lane-Emden equation

The **polytropic model** combines the equation of hydrostatic equilibrium with mass conservation

$$\frac{dm}{dr} = 4\pi r^2 \rho, \quad \frac{dP}{dr} = -\frac{Gm\rho}{r^2}.$$

If P is only a function of density, then these two equations are independent from the equation of radiative diffusion and equation of energy generation. To combine the two equations, first multiply the equation of hydrostatic equilirium by r^2/ρ and differentiate with respect to r,

$$\frac{d}{dr}\left(\frac{r^2}{\rho}\frac{dP}{dr}\right) = -G\frac{dm}{dr}.$$

The RHS is just the mass conservation, so

$$\frac{1}{r^2}\frac{d}{dr}\left(\frac{r^2}{\rho}\frac{dP}{dr}\right) = -4\pi G\rho. \tag{4.5}$$

This can also be derived directly from Poisson's equation, $\nabla^2\Phi = 4\pi G\rho$, where Φ is the gravitational potential. The gravitational acceleration is given by $g = d\Phi/dr$. Using hydrostatic equilibrium, $d\Phi/dr = g = -(1/\rho)(dP/dr)$, then

$$4\pi G\rho = \nabla^2\Phi = \frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d\Phi}{dr}\right) \implies \frac{1}{r^2}\frac{d}{dr}\left(\frac{r^2}{\rho}\frac{dP}{dr}\right) = -4\pi G\rho.$$

In polytropic models, the idea is to assume a polytropic equation of state:

$$P(r) = K(\rho(r))^{1+\frac{1}{n}}.$$

where K and γ are constants, and n is known as the **polytropic index**. Plugging this equation of state into (4.5) yields

$$\frac{(n+1)K}{4\pi Gn}\frac{1}{r^2}\frac{d}{dr}\left(r^2\rho^{-1+\frac{1}{n}}\frac{d\rho}{dr}\right) = -\rho.$$

Now we want to *non-dimensionalize* this equation by defining $\rho \equiv \rho_c\theta^n$, where ρ_c is the central density and θ is some dimensionless variable. This gives

$$\frac{(n+1)K\rho_c^{-1+\frac{1}{n}}}{4\pi G}\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d\theta}{dr}\right) = -\theta^n.$$

The coefficient on the LHS has dimensions of length squared. Thus, we can also define a dimensionless variable for length,

$$r = \alpha\xi = \left[\frac{(n+1)K\rho_c^{-1+\frac{1}{n}}}{4\pi G}\right]^{1/2}\xi.$$

The final nice and neat equation is the following:

$$\frac{1}{\xi^2}\frac{d}{d\xi}\left(\xi^2\frac{d\theta}{d\xi}\right) = -\theta^n. \tag{4.6}$$

This equation is known as the **Lane-Emden equation**. It is a second order differential equation. Only when $n = 0$, 1, and 5 gives analytical solutions. The boundary conditions are $\theta = 1$ and $d\theta/d\xi = 0$ at $\xi = 0$. Physically, this means the central density is $\rho = \rho_c$, and the derivative of density at $r = 0$ is zero. $\theta(\xi)$ is a monotonically decreasing function until $\theta(\xi) = 0$ at $\xi = \xi_1$. Physically ξ_1 corresponds to the surface of the star because obviously the density outside the star should be zero. With the polytropic model, one can find some relations dependent on n in the form of some "polytropic constants" such as R_n, M_n, etc. These constants can be determined numerically (or analytically for analytical solutions).

4.2.2 Polytropic Constants

Stellar Radius The stellar radius is denoted by

$$R = \alpha\xi_1 = \alpha R_n, \tag{4.7}$$

where $R_n = \xi_1$ is where θ first reaches zero.

Stellar Mass The stellar mass is the integral

$$M=\int_0^R 4\pi r^2\rho\,dr=4\pi\rho_c\alpha^3\int_0^R \xi^2\theta^n\,d\xi=-4\pi\rho_c\alpha^3\int_0^{\xi_1}\frac{d}{d\xi}\left(\xi^2\frac{d\theta}{d\xi}\right)d\xi=-4\pi\rho_c\alpha^3\xi_1^2\frac{d\theta}{d\xi}\bigg|_{\xi_1}.$$

In the second line, we substitute θ^n using the Lane-Emden equation 4.6. The constant M_n is defined as

$$M=4\pi\rho_c\alpha^3 M_n\quad\text{or}\quad M_n=-\xi_1^2\frac{d\theta}{d\xi}\bigg|_{\xi_1}. \tag{4.8}$$

For a given n, if the mass and K in the polytropic equation of state are known, we can calculate the central density ρ_c.

Average Density The average density is given by

$$\bar{\rho}=\frac{M}{4\pi R^3/3}=-4\pi\rho_c\alpha^3\xi_1^2\frac{d\theta}{d\xi}\bigg|_{\xi_1}\bigg/\frac{4\pi}{3}\alpha^3\xi_1^3=-\frac{3(d\theta/d\xi)_{\xi_1}}{\xi_1}\rho_c.$$

The central density and the average density are in linear relation. We shall define D_n to be the constant of proportionality:

$$\rho_c=-\frac{\xi_1}{3(d\theta/d\xi)_{\xi_1}}\bar{\rho}\equiv D_n\bar{\rho}. \tag{4.9}$$

Numerically, D_n increases with n, so the star becomes more centrally concentrated for higher n.

Central Pressure The final polytropic constant relate the central pressure P_c to the central density ρ_c. Starting with the polytropic equation of state, $P_c=K\rho_c^{1+\frac{1}{n}}$, substituting K by α:

$$P_c=K\rho_c^{1+\frac{1}{n}}=\alpha^2\frac{4\pi G}{n+1}\rho_c^{1-\frac{1}{n}}\cdot\rho_c^{1+\frac{1}{n}}=\alpha^2\frac{4\pi G}{n+1}\rho_c^2=\left(\frac{R}{R_n}\right)^2\frac{4\pi G}{n+1}\rho_c^2.$$

Using (4.9), we can rewrite the stellar radius

$$\bar{\rho}=\frac{M}{4\pi R^3/3}=\frac{\rho_c}{D_n}\quad\Longrightarrow\quad R=\left(\frac{3D_nM}{4\pi\rho_c}\right)^{1/3}.$$

The central pressure is then

$$P_c=\frac{1}{R_n^2}\left(\frac{3D_nM}{4\pi\rho_c}\right)^{2/3}\frac{4\pi G}{n+1}\rho_c^2=\frac{(3D_n)^{2/3}}{R_n^2(n+1)}(4\pi)^{1/3}GM^{2/3}\rho_c^{4/3}.$$

Combining all n-dependent quantities into a single constant B_n, we have:

$$P_c=B_n(4\pi)^{1/3}GM^{2/3}\rho_c^{4/3}. \tag{4.10}$$

Numerically, B_n varies very slowly with n.

Mass-Radius Relation To derive the mass-radius relation, first rewrite ρ_c in terms of α from the definition of α,

$$\alpha=\left[\frac{(n+1)K\rho_c^{-1+\frac{1}{n}}}{4\pi G}\right]^{1/2}\quad\Longrightarrow\quad\rho_c=\left[\frac{4\pi G}{(n+1)K}\alpha^2\right]^{\frac{n}{1-n}}.$$

Substitute this into (4.8),

$$M=4\pi\alpha^3\left[\frac{4\pi G}{(n+1)K}\alpha^2\right]^{\frac{n}{1-n}}M_n\quad\Longrightarrow\quad\left(\frac{M}{M_n}\right)^{1-n}=4\pi\left[\frac{G}{(n+1)K}\right]^n\alpha^{3-n}.$$

Then by $R=\alpha R_n$, the mass-relation is

$$\left(\frac{M}{M_n}\right)^{1-n}=4\pi\left[\frac{G}{(n+1)K}\right]^n\left(\frac{R}{R_n}\right)^{3-n}. \tag{4.11}$$

For $1 < n < 3$, the radius decreases when the mass increases,

$$R^{3-n} \propto \frac{1}{M^{n-1}}.$$

If $n = 1$, there is a unique R determined by K which is independent of M:

$$R = \left(\frac{K}{2\pi G}\right)^{1/2} R_1,$$

where $R_1 = R_{n=1}$. For $n = 3$, there is a unique M determined by K which is independent of R:

$$M = \frac{M_3}{\sqrt{4\pi}} \left(\frac{4K}{G}\right)^{3/2},$$

where $M_3 = M_{n=3}$. This means there is only one mass that satisfies hydrostatic equilibrium for a given K. We will analyze the $n = 3$ case in more detail later.

Summary The following are useful polytropic relations and relevant constants:

$$R = \alpha R_n, \qquad M = 4\pi\rho_c\alpha^3 M_n,$$
$$\rho_c = D_n\bar{\rho}, \qquad P_c = B_n(4\pi)^{1/3} G M^{2/3} \rho_c^{4/3}.$$

Table 4.1 shows some numerical values of polytropic constants.

n	R_n	M_n	D_n	B_n
1.0	π	π	3.290	0.233
1.5	3.65	2.71	5.991	0.206
2.0	4.35	2.41	11.40	0.185
2.5	5.36	2.19	23.41	0.170
3.0	6.90	2.02	54.18	0.157
1.5	9.54	1.89	152.9	0.145

Table 4.1: Polytropic Constants.

4.2.3 $n = 1.5$ Polytrope

When $n = 1.5$, the equation of state is

$$P = K\rho^{1+\frac{1}{1.5}} = K\rho^{5/3}.$$

We already know that this is the non-relativistic degeneracy pressure.

Low-Mass White Dwarfs The degeneracy pressure from electrons in a white dwarf star is given by (2.8)

$$P_{nr} = \frac{h^2}{5m_e}\left(\frac{3}{8\pi}\right)^{2/3} n_e^{5/3} = \frac{h^2}{5m_e}\left(\frac{3}{8\pi}\right)^{2/3}\left(\frac{\rho}{\mu_e m_p}\right)^{5/3} \propto \rho^{5/3}.$$

It corresponds to $n = 1.5$ with

$$K = \frac{h^2}{5m_e}\left(\frac{3}{8\pi}\right)^{2/3}\left(\frac{1}{\mu_e m_p}\right)^{5/3} \approx 9.94 \times 10^6 \mu_e^{-5/3}$$

in SI units. Typical white dwarfs are made up of helium, carbon, and oxygen, which have mean molecular mass per electron $\mu_e = 2$. Directly substituting $n = 3/2$ into the mass-radius relation (4.11) we just derived, we find a familiar relation:

$$R^{3-n} \propto \frac{1}{M^{n-1}} \quad \Longrightarrow \quad R \propto M^{-1/3}.$$

In previous chapters we just stop here, but with the Lane-Emden equation, we can find the constant of proportionality. From (4.11),

$$\left(\frac{M}{M_n}\right)^{1-n} = 4\pi\left[\frac{G}{(n+1)K}\right]^n\left(\frac{R}{R_n}\right)^{3-n}.$$

Solving for R,

$$R = (4\pi)^{-\frac{1}{3-n}}\left[\frac{(n+1)K}{G}\right]^{\frac{n}{3-n}}\left(\frac{M}{M_n}\right)^{\frac{1-n}{3-n}} R_n.$$

We can find $M_{1.5}$ and $R_{1.5}$ in Table 4.1. Subtituting these and $K = 9.94\times 10^6\mu_e^{-5/3}$,

$$R = 0.0148\left(\frac{M}{0.6\,M_\odot}\right)^{-1/3}\left(\frac{2}{\mu_e}\right)^{5/3} R_\odot.$$

Let $\mu_e = 2$ and $M = 0.6\,M_\odot$, a typical white dwarf radius is about $0.015\,R_\odot$, a bit larger than the Earth's radius.

Convective Stars In the discussion of convection, we said that the pressure-density relation is *very* close to the adiabatic relation,

$$P \propto \rho^\gamma.$$

The monatomic ideal gas has $\gamma = 5/3$, which corresponds to $n = 1.5$ polytrope. Fully convective stars and red giants that typically have a significant convective envelope are often described well by $n = 1.5$ polytrope.

4.2.4 $n = 3$ Polytrope

Ultra-Relativistic Degeneracy Pressure The relativistic degeneracy pressure is given by (2.9),

$$P_r = \frac{hc}{4}\left(\frac{3}{8\pi}\right)^{1/3} n_e^{4/3} = \frac{hc}{4}\left(\frac{3}{8\pi}\right)^{1/3}\left(\frac{\rho}{\mu_e m_p}\right)^{4/3} \propto \rho^{4/3}.$$

This corresponds to an $n = 3$ polytrope, where we already know that only one mass can support hydrostatic equilibrium (unstable) for a given K,

$$K = \frac{hc}{4}\left(\frac{3}{8\pi}\right)^{1/3}\left(\frac{1}{\mu_e m_p}\right)^{4/3} \approx 1.23\times 10^{10}\mu_e^{-4/3}$$

in SI units. The polytropic mass-radius relation (4.11) says that

$$M = \frac{M_3}{\sqrt{4\pi}}\left(\frac{4K}{G}\right)^{3/2} \approx 1.44\left(\frac{2}{\mu_e}\right)^2 M_\odot \equiv M_{\rm Ch},$$

where $M_3 = 2.02$ and $\mu_e = 2$. This mass is defined as the Chandrasekhar mass $M_{\rm Ch}$, the maximum mass a white dwarf can have.

Eddington Standard Model Recall that the Eddington standard model assumes that the ideal gas pressure $P_{\rm gas}$ is some constant fraction β of the total pressure P,

$$P_{\rm gas} = \beta P, \quad P_{\rm rad} = (1-\beta)P.$$

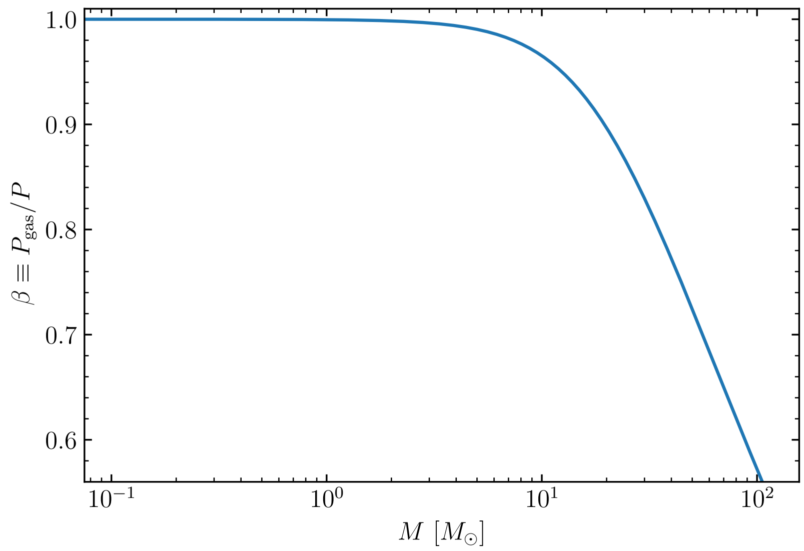
According to (2.19)

$$P_{\rm tot} = \left(\frac{1-\beta}{\beta^4}\right)^{1/3}\left(\frac{3}{a}\right)^{1/3}\left(\frac{k}{\mu m_p}\right)^{4/3}\rho^{4/3}.$$

With β constant throughout the star, $P \propto \rho^{4/3}$. Combining the two equations

$$K = \left(\frac{1-\beta}{\beta^4}\right)^{1/3}\left(\frac{3}{a}\right)^{1/3}\left(\frac{k}{\mu m_p}\right)^{4/3} \quad \text{and} \quad M = \frac{M_3}{\sqrt{4\pi}}\left(\frac{4K}{G}\right)^{3/2},$$

we can show that there is a unique (positive and real) β for a given stellar mass M:



A more massive star is more radiation pressure dominated. The Sun has $\beta \approx 0.9996$, so it is ideal gas pressure dominated. With $\bar{\rho} = 1.4\,\mathrm{g/cm^3}$ and $D_3 \approx 54.18$, the central density of the Sun is $\rho_c = D_3\bar{\rho} \approx 76.4\,\mathrm{g/cm^3}$, about a factor of 2 lower than the observed value. The central temperature can be obtained by the ideal gas law, giving $T_c \approx 1.2 \times 10^7\,\mathrm{K}$, very close to the observed value $1.5 \times 10^7\,\mathrm{K}$. Thus, the $n = 3$ polytrope fits to the Sun's structure well.

4.3 Stellar Evolution

From now on, many statements made about stellar evolution will be rather qualitative. This is because stellar evolution is a three-dimensional time-dependent many-body problems. The actual evolution relies heavily on numerical simulations. MESA model is a good website to simulate stellar evolution process.

4.3.1 Star Formation

A cold molecular cloud has a typical temperature of $15\,\mathrm{K}$, a typical density of $\rho \sim 10^{-18}\,\mathrm{kg/m^3}$, and a typical mass of $10^4 - 10^7\,M_\odot$. It is prevented from collapse by ideal gas pressure, rotation, magnetic pressure, and turbulent pressure. Perturbations in density may trigger collapse of small regions.

Consider a region of mass M and radius R with density $\rho \sim M/R^3$. The collapse starts if the gravitational potential energy is greater than the kinetic energy,

$$|E_g| \gtrsim E_k.$$

If $E_g \sim -GM^2/R$ and $E_k = \frac{3}{2}NkT \sim MkT/\bar{m}$, then we can find the critical mass for collapse:

$$\frac{GM^2}{R} \gtrsim \frac{M}{\bar{m}}kT \quad \Longrightarrow \quad GM^{5/3}\rho^{1/3} \gtrsim \frac{M}{\bar{m}}kT.$$

Solving for M, the critical mass is known as the **Jeans mass**:

$$M \gtrsim M_{\rm J} \equiv \left(\frac{kT}{G\mu m_p}\right)^{3/2} \rho^{-1/2}. \tag{4.12}$$

If we plug in $T \approx 15\,\mathrm{K}$ and $\rho \sim 10^{-18}\,\mathrm{kg/m^3}$, we find that $M_{\rm J} \sim 40\,M_\odot$, off by some factors of unity. The collapse can be seen as a free fall, so its timescale is given by the dynamical timescale:

$$t_{\rm dyn} \sim \frac{1}{\sqrt{G\rho}} \sim 10^6\,\mathrm{yr}.$$

Protostar The density increases during the collapse of a region of gases. This process is isothermal because gas molecules are constantly heating up by cosmic rays but cooling down by dust grain emission. Since the Jeans mass scales like $M_J \propto \rho^{-1/2}$, it decreases during the collapse. A large cloud can **fragment** into smaller clouds, which ends up with fragments of typical mass $\sim 0.006\, M_\odot$.

Recall that the mean free path of photons is given by

$$l = \frac{1}{\kappa\rho}.$$

Now the cloud fragment is denser, and denser clouds also give a higher opacity, so the mean free path of photons decreases. The fragment is no longer transparent to its radiation, so it heats up and gain pressure. There exists a core of the fragment that has a state nearly in hydrostatic equilibrium, while the fragment itself is accreting gas from surrounding gases and gaining mass. This forms a **protostar**. For a protostellar core of mass M_{pro} and R_{pro}, the luminosity from accretion is

$$L_{\text{acc}} = \eta \frac{GM_{\text{pro}}\dot{M}}{R_{\text{pro}}}.$$

The qualitative interpretation is the following: a small mass dm near the protostar is accreted by it. The gravitational potential energy is converted into kinetic energy. At the surface of the protostellar core, the kinetic energy gained by the mass is $GM_{\text{pro}}\, dm/R_{\text{pro}}$. The core is already at (near) hydrostatic equilibrium, so the mass must stop somewhere near or inside the core. Hence the kinetic energy is radiated away in the form of thermal energy at rate L_{acc}, where η (typically between 0.1 and 0.001) is the efficiency of this conversion. Finally, the protostar grows at a rate $\dot{M} \sim 1 - 10\, M_\odot/\text{Myr}$ for a time about 0.1 Myr.

4.3.2 Pre-Main Sequence Contraction

When the gas around the protostar are cleared, it is now a **pre-main sequence star** that is isolated in space. In hydrostatic equilibrium, it will contracts on a Kelvin-Helmholtz timescale. The contraction will stop when one of the two things happen:

1. Electrons in the core becomes degenerate, and the collapse is prevented by the degeneracy pressure. The pre-main sequence star becomes a **brown dwarf**.

2. The central temperature is high enough to trigger nuclear fusion, and the collapse is prevented by ideal gas and radiation pressure. The pre-main sequence star becomes an actual star.

By the virial theorem, the typical temperature of a star is

$$kT \sim \frac{GM\mu m_p}{R} \quad \Longrightarrow \quad kT \sim GM^{2/3} m_p \rho^{1/3}, \tag{4.13}$$

where ρ is the average density. During contraction, both ρ_c, ρ, and T will increase. The center will become degenerate when the electron number density n_e is comparable to the quantum concentration,

$$\frac{\rho}{\mu_e m_p} \sim \left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} \quad \Longrightarrow \quad \rho \sim m_p \left(\frac{m_e kT}{h^2}\right)^{3/2}. \tag{4.14}$$

Once the center becomes degenerate, the pressure is independent of temperature as $P_{\text{deg}} \propto \rho^{5/3}$. By approximating $P \sim GM^2/R^4$ and $\rho \sim M/R^3$, we have

$$\rho \propto M^2.$$

This means when the core becomes degenerate, the density of the star is fixed by mass. There will be no contraction and thus no heating, unless M is increased for no reason.

Hydrogen fusion starts when the central temperature reaches $T \sim 10^7$ K. This should be achieved before the core becomes degenerate, so we will set the threshold using (4.13) and (4.14):

$$kT \sim GM^{2/3} m_p \rho^{1/3} \sim GM^{2/3} m_p \left[m_p \left(\frac{m_e kT}{h^2}\right)^{3/2}\right]^{1/3} \sim GM^{2/3} m_p^{3/2} \left(\frac{mkT}{h^2}\right)^{1/2}.$$

Solving for M gives the minimum mass for a star:

$$M_{\min} \sim \left(\frac{kT}{m_e}\right)^{3/4} \left(\frac{h}{G}\right)^{3/2} \frac{1}{m_p^2} \approx 0.24\, M_\odot.$$

It is modeled that the minimum mass for a star is about $0.08\, M_\odot$ or 84 Jupiter masses. The analysis above is a factor of 3 off. If the mass of an object is less than $0.08\, M_\odot$, then it is a brown dwarf or a planet. Note that deuterium burning in brown dwarfs only requires $T \sim 10^6\,\mathrm{K}$, while planets cannot undergo deuterium burning. Thus, it is defined that an object with mass less than 13 Jupiter masses is a **planet**.

Figure 4.1 shows the evolution of stars with different masses on the Hertzsprung-Russell diagram.

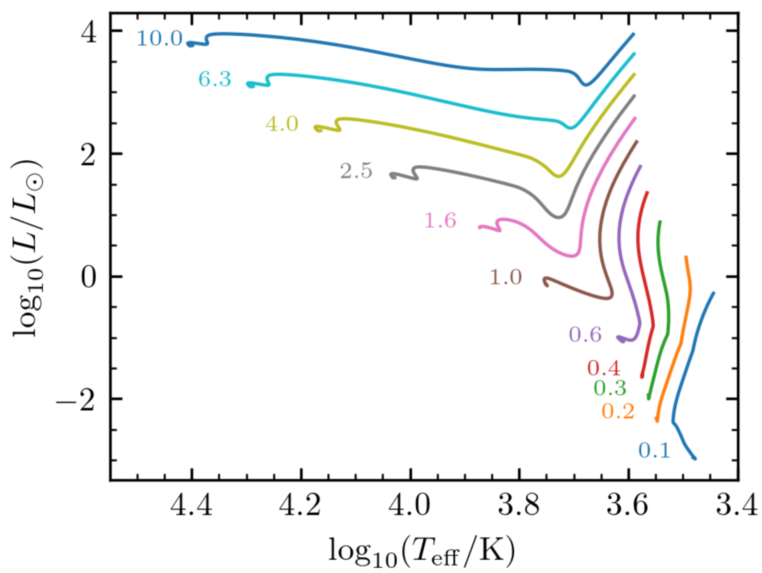


Figure 4.1: The evolution of stars with different masses in pre-main sequence phase. Some always evolve vertically down at roughly the same T_{eff}; some has increasing T_{eff} later; some always evolve towards higher T_{eff}.

The vertical tracks are called the **Hayashi track**, while the horizontal tracks are the **Henyey track**. When a star is on the hayashi track, they are fully convective. A star on the Henyey track develops a radiative core.

Hayashi Track On the Hayashi track, the star's opacity contains all kinds (bound-bound, bound-free, H^-, etc) of opacities. The opacity is large at low temperatures, so convection is the main way of energy transport. Above the photosphere, the star is radiative or else it is not glowing.

In hydrostatic equilibrium, the central temperature and central pressure of an ideal gas are like

$$P_c \propto \frac{M^2}{R^4} \quad \text{and} \quad T_c \propto \frac{M}{R}.$$

We know the photosphere pressure is given by (2.23),

$$P_{\mathrm{phot}} = \frac{2g}{3\kappa} \sim \frac{2GM}{3R^2\kappa}.$$

On the Hayashi track, the temperature is cool enough that H^- opacity dominates, $\kappa \propto \rho^{0.5} T^9$, so

$$P_{\mathrm{phot}} \propto \frac{M}{R^2} \frac{1}{\rho_{\mathrm{phot}}^{0.5} T_{\mathrm{phot}}^9}.$$

Meanwhile, the classical ideal gas law says that

$$P_{\mathrm{phot}} = \frac{\rho_{\mathrm{phot}}}{\bar{m}} k T_{\mathrm{phot}} \propto \rho_{\mathrm{phot}} T_{\mathrm{phot}}.$$

Finally, convection over all stars give the adiabatic relation between T and P until the photosphere,

$$\left(\frac{d\ln T}{d\ln P}\right) = 1 - \frac{1}{\gamma} = \frac{2}{5} \quad \Longrightarrow \quad \frac{T_c}{P_c^{2/5}} = \frac{T_{\rm phot}}{P_{\rm phot}^{2/5}}.$$

Combining all the following equations derived above,

$$P_c \propto \frac{M^2}{R^4}, \quad T_c \propto \frac{M}{R}, \quad P_{\rm phot} \propto \frac{M}{R^2}\frac{1}{\rho_{\rm phot}T_{\rm phot}^9}, \quad P_{\rm phot} \propto \rho_{\rm phot}T_{\rm phot}, \quad \frac{T_c}{P_c^{2/5}} = \frac{T_{\rm phot}}{P_{\rm phot}^{2/5}},$$

we can find that the radius of a star on the Hayashi track scales as

$$R \propto M^{-14}T_{\rm phot}^{49}.$$

The temperature of the photosphere $T_{\rm phot}$ is just the effective temperature $T_{\rm eff}$ by definition. Relating it to the surface luminosity $L = 4\pi R^2\sigma T_{\rm eff}^4$ gives

$$T_{\rm eff} \propto L^{1/102}M^{-7/51}.$$

The shallow dependence of L is consistent with the Hayashi track in Figure 4.1. Even though L varies over several orders of magnitude, $T_{\rm eff}$ varies little.

Henyey Track Once the temperature of the star is high enough that the core has radiative diffusion and the opacity obeys Kramers' law, the luminosity scales like (4.1),

$$L_{\rm rad} \propto \frac{\mu^{7.5}M^{5.5}}{R^{0.5}}.$$

As the star contracts, its luminosity and effective temperature both increases slighly. This is shown on the Hanyey track in Figure 4.1.

4.3.3 Main Sequence

When a star starts core hydrogen burning, it is at the **zero-age main sequence (ZAMS)**. Hydrogen is converted into helium in the core. We already know most of the characteristics of main sequence stars.

Stars with $M \gtrsim 2\,M_\odot$:

1. The core hydrogen burning is by the CNO cycle, $\epsilon_{\rm CNO} \propto T^{18}$.
2. Since $\epsilon_{\rm CNO}$ has a steep temperature dependence, there is only a limited range of mass coordinate that has temperatures enough for the CNO cycle. This gives a very large energy flux $L(r)/m(r)$. That is, the core of a star is convective. The energy transport in the envelope is dominated by radiative diffusion, with opacity dominated by Thomson scattering opacity.
3. Turbulent motion by convection also mix up elements in the core. The size of the convective core determines the mass fraction burned from hydrogen to helium.
4. A star dominated by radiative diffusion has $L_{\rm rad} \propto \mu^4 M^3$ with Thomson scattering opacity. As hydrogen is burned into helium, μ increases over the stellar lifetime. Hence the star becomes brighter during the main sequence.

Stars with $M \lesssim 2\,M_\odot$:

1. The core hydrogen burning is by the proton-proton chain, $\epsilon_{pp} \propto T^4$.
2. Stars with $0.5 \lesssim M \lesssim 1.2\,M_\odot$ have radiative cores and convective envelopes. The opacity is dominated by bound-free and free-free opacity, obeying Kramers' law. Stars with $M \lesssim 0.5\,M_\odot$ are fully convective.
3. Similar to high-mass stars, as μ increases, the star becomes brighter during the main sequence.

At the end of the main sequence, a star exhausts all core hydrogen. This stage before helium burning is called the **terminal age main sequence (TAMS)**.

4.3.4 Post-Main Sequence

When the main sequence ends, the mass of the star determines whether there are further nuclear fusion.

Stars with $M \gtrsim 2\,M_\odot$:

1. The helium core is inert until T rises to $\sim 10^8\,\text{K}$. More hydrogen burning will occur in a **shell** around the helium core. There is a **Henyey hook** at the time of core hydrogen exhaustion. This is because the convective core is well-mixed, and the transition from core hydrogen burning to shell hydrogen burning is abrupt.

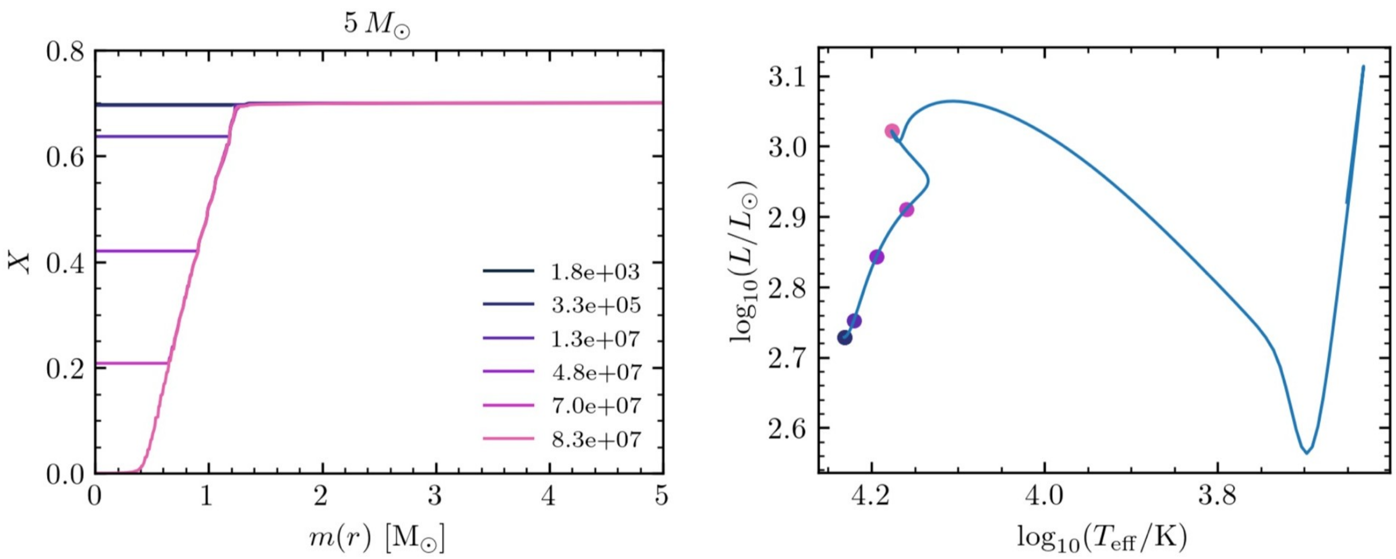


Figure 4.2: The evolution of a $5\,M_\odot$ main sequence star. The left plot shows the hydrogen mass fraction as a function of mass coordinate. (Different colors indicates the age.) The right plot shows the HR diagram of the evolution with corresponding colors.

   The mass of the helium core in this case will exceed the **Schönberg-Chandrasekhar limit** ($M_{\text{core}} \gtrsim 0.08M$). This limit is the maximum isothermal core mass that can support the weight of the overlying envelope while maintaining thermal equilibrium. In other words, if the core exceeds this limit, then it will contract on a Kelvin-Helmholtz timescale, heating up the core. The basic argument of the Schönberg-Chandrasekhar limit requires a *classical ideal gas* pressure.

   The core heats up as it contracts. It also heats up the hydrogen burning shell. The rate of energy generation increases ($\epsilon_{\text{CNO}} \propto T^{18}$). The virial theorem for a classical ideal gas says that the envelope will respond by expanding and cooling. This process is sometimes called the **mirror principle** (i.e. core contracts while envelope expands). The star will eventually expand into a **red giant** or **red supergiant**. Since the Kelvin-Helmholtz timescale is much shorter than the hydrogen nuclear timescale, it is rare to find a star in this phase obeying the mirror principle. Thus, there is a gap known as the **Hertzsprung gap** between the upper main sequence and the red giant branch.

   On the red giant branch, the opacity in the envelope is large because of its low temperature. The envelope is now convective, so the star evolve backwards along the Hayashi track with T_{eff} approximately constant.

   Finally, the core will have enough temperature to ignite helium. Once that happens, the core expands while the envelope contracts a little (another case of mirror principle).

2. When the core helium is exhausted, a star will have a carbon-oxygen core. There will be helium burning in the shell around the core, and hydrogen burning in an outer shell. Again, the core contracts and the envelope expands as the star proceeds along the Hayashi track. This phase is called the **asymptotic giant branch (AGB)**. The star is very luminous and has a radiation pressure dominated surface, so there is a strong mass loss from the star.

   For stars with initial mass $\lesssim 8\,M_\odot$, they will not undergo carbon burning. The electron becomes degenerate in the core. The hydrogen envelope is either burned into helium, then carbon and oxygen, or is lost through stellar winds. Eventually, there will be only a carbon-oxygen white dwarf left. A new-born white dwarf has a temperature of $T_{\text{eff}} \sim 10^5 - 10^6\,\text{K}$. It can radiate energetic photons to the environment and excite the gases from the envelope just losed. These gases will further emit photons, forming a **planetary nebula**.

3. Stars with initial mass between 8 and $10\,M_\odot$ can undergo carbon burning. One possible product is an oxygen-neon white dwarf. Another possible result (for stars near the upper limit of mass range) is an **electron-capture supernova**: the products of carbon burning include ^{20}Ne and ^{24}Mg. When the degenerate oxygen-neon core grows near the Chandrasekhar limit, the density may be high enough for electron capture reactions to occur:

$$\begin{aligned} {}^{24}\mathrm{Mg} + e^- &\longrightarrow {}^{24}\mathrm{Na} + \nu_e \\ {}^{24}\mathrm{Na} + e^- &\longrightarrow {}^{24}\mathrm{Ne} + \nu_e. \end{aligned}$$

But the hydrostatic equilibrium is supported by electron degeneracy pressure. Now electron number is reduced, the core starts to contract. As the density increases, further electron capture reactions may occur:

$$\begin{aligned} {}^{20}\mathrm{Ne} + e^- &\longrightarrow {}^{20}\mathrm{F} + \nu_e \\ {}^{20}\mathrm{F} + e^- &\longrightarrow {}^{20}\mathrm{O} + \nu_e, \end{aligned}$$

accelerating the contraction. Eventually, the core may implodes into a neutron star or explodes if oxygen burning at the center occurs. The mass range and the fate of an oxygen-neon white dwarf are still debated today.

4. Stars with initial mass more than $10\,M_\odot$ can undergo nuclear burning all the way to silicon burning. Each burning stage proceeds from the core to the surrounding shell, and each stage is shorter than the previous one (silicon burning may only take several days). Section 3.4.2 discusses how to proceed to an iron core in details. When the iron core is formed, it starts to contract and electrons become relativistically degenerate. During the contraction, the iron core heats up to a temperature where photons are so energetic that

$${}^{56}\mathrm{Fe} + \gamma \longrightarrow 13{}^{4}\mathrm{He} + 4n$$

is favored. This reaction is endothermic (absorbing $2\,\mathrm{MeV/nucleon}$) and reduces the pressure (because it reduces the adiabatic index to less than 4/3), accelerating the contraction. The contraction now is a free fall, with a timescale of

$$t_{\mathrm{dyn}} \sim \frac{1}{\sqrt{G\rho}} \sim 0.1\,\mathrm{s},$$

where the typical density of an iron core is $\rho \sim 10^{12}\,\mathrm{kg/m^3}$. Then temperature will increase to a stage where the photodisintegration of ^{4}He occurs:

$${}^{4}\mathrm{He} + \gamma \longrightarrow 2p + 2n.$$

It is endothermic (absorbing $6\,\mathrm{MeV/nucleon}$) and pressure reducing.

5. Electron capture reactions begin when the density is $\sim 10^{15}\,\mathrm{kg/m^3}$,

$$p + e^- \to n + \nu_e.$$

The iron core collapses into a neutron star. The dynamical timescale is to the order of ms. The energy of the collapse released by neutrinos is around 10^{42} J. This event is a **core-collapse supernova**.

The pressure of a neutron star is dominated by the neutron degeneracy pressure. One may think that all we need to do is to replace m_e by m_n in the expression of degeneracy pressure,

$$P_{\mathrm{deg}} \sim \frac{h^2}{m_n} n_n^{5/3}.$$

However, neutron stars are general relativistic because of its high density and strong gravitational fields. There are also strong nuclear force between neutrons. Therefore, the equation of state of neutron stars is still unknown. The equivalent of the Chandrasekhar mass, known as the **TOV limit** is also unknown. We know that it is around $2-3\,M_\odot$ because no neutron star with a higher mass has been observed. When the TOV limit is exceeded, the neutron star collapse into a black hole.

Stars with $M \lesssim 2\,M_\odot$:

1. Stars $\lesssim 1.5\,M_\odot$ without convective cores at the end of the main sequence will gradually transition from core hydrogen burning to shell hydrogen burning. Thus, there is no Henyey hook.

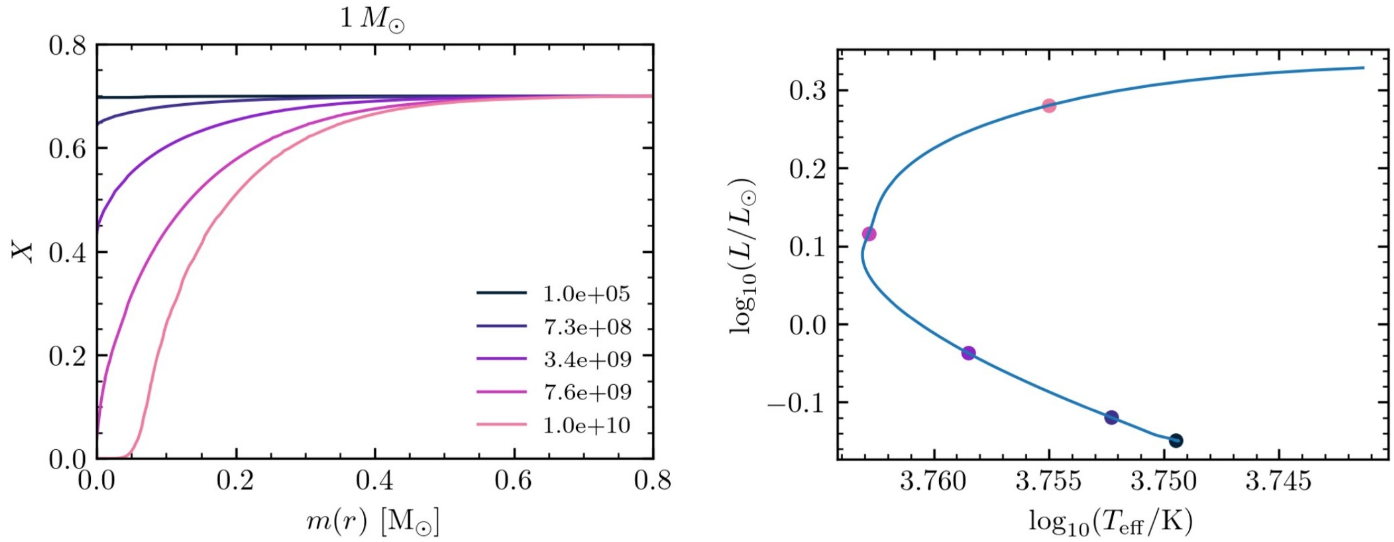


Figure 4.3: The evolution of a $1\,M_\odot$ star on the main sequence: the hydrogen mass fraction as a function of mass coordinate and ages (left) and the HR diagram.

Stars with $M \lesssim 2\,M_\odot$ will have a degenerate helium core. Since Schönberg-Chandrasekhar limit assumes that the core is an ideal gas, this limit does not apply here. This kind of stars transitions from the main sequence to the red giant branch. This phase is known as the **subgiant branch**.

2. The helium core still contracts and heat up even it is degenerate. This is because the hydrogen shell-burning increase the mass of the degenerate helium core, and $R \propto M^{-1/3}$. The shell is an ideal gas at nearly thermal equilibrium with the core, so its temperature $T \propto M_{\mathrm{core}}/R_{\mathrm{core}}$ rises. The core is isothermal because of the high efficiency of energy transport by electron conduction. As the surface luminosity is given by the hydrogen burning shell, the red giant is brighter when the core mass increases.

   Numerical models show that helium fusion starts when the core mass reaches $0.48\,M_\odot$, giving $L \sim 10^{3.5}\,L_\odot$, *regardless* of the total mass. This is known as the **tip of the red giant branch**. Since the luminosity of the tip is fixed, we can measure the distance to the nearby galaxy that contains this red giant.

3. In a degenerate core, nuclear burning is explosive. This is because the degeneracy pressure is independent of temperature. There is no expansion work that releases energy, so the total energy keeps growing in the core. At some point, the temperature will be high enough to lift the degeneracy to an ideal gas, leading to a **helium flash** that releases tons of energy. The nuclear burning luminosity can reach $10^{11}\,L_\odot$ for a few seconds. Once the core becomes an ideal gas, helium burning is in a regular mannor. The core expands and the envelope contracts after the helium flash. Stars at solar metallicity undergoing core helium burning are called **red clump stars**. Those with lower metallicity are **horizontal branch stars**.

4. Because the initial mass of these stars is below $8\,M_\odot$, they will end up being an carbon-oxygen white dwarf.

Special thanks to Sunny Wong, my instructor of the course Phys 132: Stellar Structure and Evolution.

References:

Physics of Stars by Phillips.

Stellar Structure and Evolution by Pinsonneault and Ryden.

Principles of Stellar Evolution and Nucleosynthesis by Clayton.

Stellar Structure and Evolution by Kippenhahn, Weigert and Weiss.

An introduction to the Theory of Stellar Structure and Evolution by Prialnik.

Part II

Galaxies

To be written.

Part III

Cosmology

*Notation:

E/U: total energy/internal energy.
E_g: gravitational potential energy.
E_k: kinetic energy.
Q: heat.
P: pressure.
V: volume
T: temperature.
F: energy flux.
L: luminosity.
N: particle number.
n: number density.
ϵ/u: energy density
ρ: mass density.
$\bar{m}$: average mass.
z: redshift.
a: scale factor.
$a_{\rm rad}$: radiation constant.
H: Hubble parameter.
Λ: cosmological constant.
κ: curvature parameter.
σ: Stefan-Boltzmann constant.
γ: adiabatic constant, photon, or Lorentz factor, depending on the context.
Ω: density parameter or solid angle, depending on the context.
ν: frequency or neutrino, depending on the context.

5 Introduction

Cosmology is the study of the universe as a whole. As we all know, the universe contains structures with various scales: planets, stars, galaxies, clusters of galaxies, superclusters, and so on. In cosmology, the details of these structures are negligible and the universe is idealized as a uniform and smooth volume of matter, radiation, and mysterious contents. By then we can study its past, the present, and predict the future.

*In this notes about undergraduate level cosmology, we will look at equations from various perspectives instead of deriving them rigorously. Detailed derivation of equations, such as the Friedmann equation, requires a considerable knowledge of general relativity.

5.1 Scales and Units

On the extreme scale of the universe, we need special units.

- **Length.** The astronomical unit (AU) is the mean distance between the Earth and Sun: $1\,\mathrm{AU} = 1.50 \times 10^{11}\,\mathrm{m}$. This is only for our solar system. A larger unit that is useful for interstellar distances is the parsec (pc), the distance at which 1 AU subtends an angle of 1 arcsecond, $1\,\mathrm{pc} = 3.09 \times 10^{16}\,\mathrm{m}$. The distance from the Sun to the nearest star Proxima Centauri is about 1.30 pc. To measure intergalctic distances, we often use the magaparsec (Mpc), $1\,\mathrm{Mpc} = 10^6\,\mathrm{pc} = 3.09 \times 10^{22}\,\mathrm{m}$. The distance between the Milky Way galaxy and the nearest galaxy (M31, or the Andromeda galaxy) is about 0.76 Mpc.
- **Mass and Luminosity.** The standard unit of mass and luminosity in astrophysics and cosmology are the mass and luminosity of the sun: $1\,M_\odot = 1.998 \times 10^{30}\,\mathrm{kg}$ and $1\,L_\odot = 3.85 \times 10^{24}\,\mathrm{W}$. The mass of the Milky Way galaxy is of the order $M_{\mathrm{gal}} \sim 10^{12}\,M_\odot$, and its luminosity is about $L_{\mathrm{gal}} \approx 3 \times 10^{10}\,L_\odot$.
- **Time.** For evolution of the universe, one year is too short. We often use megayears (Myr), $1\,\mathrm{Myr} = 10^6\,\mathrm{yr} = 3.16 \times 10^{13}\,\mathrm{s}$, or even gigayears (Gyr), $1\,\mathrm{Gyr} = 3.16 \times 10^{16}\,\mathrm{s}$. The age of the Earth is about 4.57 Gyr.

5.2 Fundamental Observations

5.2.1 Isotropy and Homogeneity

The universe is isotropic and homogeneous *on a large scale*, that is, on scales of $\gtrsim 100\,\mathrm{Mpc}$. This is known as the **cosmological principle**. **Isotropy** means that there is no preferred direction. **Homogeneity** means that there is no preferred location. Based on observation, the cosmological principle is a good approximation. Imagine spheres of radius r in the universe. Most spheres with $r = 3\,\mathrm{AU}$ do not contain a star. Most spheres with $r = 3\,\mathrm{Mpc}$ do not contain a pair of bright galaxies. However, most spheres of $r \gtrsim 100\,\mathrm{Mpc}$ contains about the same pattern of superclusters and voids statistically. Moreover, the cosmological principle assumes that physics laws are independent of space and time. If physics laws are different in the past or at some other locations in the universe, then using our knowledge on Earth now simply can't study the universe as a whole.

5.2.2 Olbers' Paradox

Around 16 centuries, astronomers thought that the universe was an infinite space. There are stars like the Sun that scatters throughout this space. **Olbers' paradox** says that the sky should be uniformly bright with starlight, but evidently the night sky is dark at visible wavelength. To see how Heinrich Olbers made this statement, let's compute the flux received on Earth from stars. Let $n_\star$ be the number density of stars in the universe. If the universe were infinite, isotropic, and homogeneous, we should see a star at *any* angle in the sky. Assume that all stars have a typical radius $R_\odot$ and luminosity $L_\odot$ of the Sun. Then for a star at distance r from the Earth, its solid angle (in steradian) seen from the Earth will be

$$\Omega = \frac{\pi R_\odot^2}{r^2}.$$

Its flux measured at a distance r will be

$$F = \frac{L_\odot}{4\pi r^2}.$$

Thus, the surface brightness (the power per square meter per steradian) received on Earth is

$$\Sigma_\star = \frac{F}{\Omega} = \frac{L_\odot}{\pi R_\odot^2},$$

which is independent of r. In conclusion, for an infinite universe, the entire sky at all times should be as bright as the Sun! From daily observations, this is obviously wrong. The universe can neither be infinitely large nor infinitely old.

One would suggest that the universe may not be transparent enough that we can see stars arbitrarily far away. This is certainly right, because the universe are full of dusts that absorbs light. However, as dusts absorb light and get heated up, its temperature increases and spontaneously radiates photons via blackbody radiation. These photons are typically in the infrared spectrum. Therefore, the effect of dusts is essentially turning visible lights to infrared. The flux F integrates over all frequency range, so the dusts do not do much. The conclusion of a finite universe still holds.

5.2.3 Hubble's Law

The universe is dynamic. There are galaxies moving away and towards us. When we look at a galaxy, we are basically observing the spectrum at some certain wavelength from emission lines of elements. These emission lines are universal in the lab, but they are different when seen from galaxies. Let the wavelength of a certain line measured in lab be $\lambda_{\rm em}$, and the wavelength of the "same" line from a distant galaxy be $\lambda_{\rm ob}$. They are in general not equal, but differed by the fraction called the redshift z:

$$z \equiv \frac{\lambda_{\rm ob} - \lambda_{\rm em}}{\lambda_{\rm em}} \quad \Longleftrightarrow \quad 1 + z \equiv \frac{\lambda_{\rm ob}}{\lambda_{\rm em}}. \tag{5.1}$$

Then we say that the galaxy has a **blueshift** when $z < 0$, and a **redshift** when $z > 0$. The light from most galaxies are redshifted—their wavelengths are stretched, either because the galaxies are moving away from us, or because of the expansion of the universe, or both. Redshifts is a result of the Doppler effect: the non-relativistic Doppler shift says that $z = v/c$, so we can measure the velocity v at which the galaxy is moving. The fact that the galaxies are moving away tells that the universe is expanding. In 1929, Edwin Hubble noted that more distant galaxies had higher redshifts. The relation is linear, and is known as **Hubble's law**:

$$z = \frac{H_0 r}{c} \quad \Longleftrightarrow \quad v = H_0 r \tag{5.2}$$

where $H_0 = 68 \pm 2\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ is the Hubble constant. Note that there are two types of redshift of a galaxy. One results from the expansion of the universe or space, called the cosmological redshift. The other results from the relative motion the galaxy. Take the closest galaxy (the Andromeda galaxy) for example, it experiences cosmological redshift, but the gravitational attraction between the Milky way and the Andromeda galaxy dominates over the cosmological redshift. Thus, lights from the Andromeda galaxy is blueshifted. More distant galaxies have much larger cosmological redshift but much smaller gravitational interaction with the Milky Way. This is why most galaxies appear redshifted.

Now we know that the galaxies are mostly fleeing away. Then does it mean that we are the preferred location in the universe (i.e. the universe is not homogeneous and isotropic?) The answer is no. In fact, a homogeneous and isotropic expansion of universe is completely consistent with Hubble's law. Consider three galaxies forming a triangle with relative distance r_{12}, r_{23}, and r_{31}. A homogeneous and isotropic expansion means that after some time t as the galaxies move away from *each other*, the final triangle and the initial one are similar triangles (neglect gravity for now). Mathematically,

$$r_{12}(t) = a(t) r_{12}(t_0), \quad r_{23}(t) = a(t) r_{23}(t_0), \quad r_{31}(t) = a(t) r_{31}(t_0).$$

The function $a(t)$ is called the **scale factor**. It is independent of location or direction. Since the scale factor is relative in time, it is convenient to set the normalization condition to be $a(t_0) = 1$, where t_0 means "now". For an observer in galaxy 1, the galaxy 2 and 3 are receding with a speed

$$v_{12}(t) = \frac{dr_{12}}{dt} = \dot{a} r_{12}(t_0) = \frac{\dot{a}}{a} r_{12}(t), \quad v_{31}(t) = \frac{dr_{31}}{dt} = \dot{a} r_{31}(t_0) = \frac{\dot{a}}{a} r_{31}(t).$$

Similar linear relations will be found by observers in galaxy 2 or galaxy 3. In conclusion, to all galaxies, if the universe expansion is free, homogeneous, and isotropic, they will find that other galaxies are receding. Comparing with Hubble's law (5.2), we find that

$H(t) = \dot{a}/a$. This quantity is usually called the **Hubble parameter**.

If the galaxies are moving away from each other, then they must be once close to each other in the past. From this we can deduce the time at which they are at the same place,

$$t_0 \sim \frac{r}{v} = \frac{r}{H_0 r} = H_0^{-1} = (14.38 \pm 0.42)\,\text{Gyr}.$$

This is known as the **Hubble time**. It also indicates that one Hubble time ago, all galaxies are compressed in some volume. This Hubble time has the same order of the age of the universe. It is not exactly the age of the universe because we set $H(t)$ to H_0, while $H(t)$ is actually changing with time. Nonetheless, all reasoning above suggests (though not proves) the **Big Bang** model.

5.2.4 Cosmic Microwave Background

The grandest photon gas in the universe is the cosmic microwave background (CMB). It is nearly *perfectly* fitted by a blackbody spectrum (2.24):

$$\epsilon(\nu) = \frac{8\pi h}{c^3}\frac{\nu^3}{e^{h\nu/kT} - 1}$$

at $T_0 \approx (2.7255 \pm 0.0006)\,\text{K}$. By (2.10) and (2.12), the energy density and number density of the CMB are, respectively,

$$u = a_{\text{rad}}T^4 \approx 4.175 \times 10^{-14}\,\text{J/m}^3 = 0.2606\,\text{MeV/m}^3 \quad \text{and} \quad n = bT^3 = 4.107 \times 10^8\,\text{m}^{-3}.$$

(Here since the scale factor a is much useful in cosmology, we will denote the radiation constant by a_{rad}.) The mean energy of CMB photons is then

$$\langle \epsilon_\gamma \rangle = \frac{u}{n} \approx 6.344 \times 10^{-4}\,\text{eV}.$$

The CMB is a strong indicator of the Big Bang model. The early universe was a ball of very hot matter, full of photons and other elementary particles. As the universe expands, mass-energy is conserved, so the energy density and temperature must decrease.

We can analyze the universe as a whole thermodynamically. The first law of thermodynamics says that $dQ = dE = P\,dV$. Since there is no other system exchanging energy with the universe, $dQ = 0$. The change in internal energy is exactly $dE = -P\,dV$. We can use the photon energy $E = uV = a_{\text{rad}}VT^4$ and radiation pressure (2.11), $P = aT^4/3$,

$$\frac{dE}{dt} = -P(t)\frac{dV}{dt} \quad \Longrightarrow \quad a_{\text{rad}}\left(4T^3\frac{dT}{dt}V + T^4\frac{dV}{dt}\right) = -\frac{1}{3}a_{\text{rad}}T^4\frac{dV}{dt}.$$

Rearranging some terms,

$$\frac{1}{T}\frac{dT}{dt} = -\frac{1}{3V}\frac{dV}{dt} \quad \Longrightarrow \quad \frac{d\ln T}{dt} = -\frac{1}{3}\frac{d\ln V}{dt}.$$

We do not know the exact volume of the universe, but we know that $V \propto a(t)^3$, so

$$\frac{d\ln T}{dt} = -\frac{d\ln a}{dt}.$$

This implies a simple relation between the temperature of the CMB and the expansion of the universe:

$$T_{\text{CMB}}(t) \propto a(t)^{-1} \tag{5.3}$$

6 Cosmological Dynamics

6.1 Newton vs. Einstein

6.1.1 Newtonian Gravity

Newtonian mechanics assumes that we live in a 3-dimensional Euclidean flat space with universal time rate. In a Euclidean space, the shortest distance between two points is a straight line, with the distance given by the Pythagorean theorem; the sum of interior angles in a triangle is π; the ratio of the circumference of a circle to its radius is 2π. With Euclid's axioms, all other theorems about geometry can be proved. In particular, the invariant (infinitesimal) line element in Newtonian mechanics is given by

$$ds^2 = dx^2 + dy^2 + dz^2.$$

Newton's law of motion says that an object in an inertial (non-accelerating) frame experiencing no net force will travel along a straight line with constant velocity. When there is a net force, Newton proposed his second law of motion:

$$\mathbf{F} = m_i \mathbf{a},$$

where we will call m_i the inertial mass for now. Newton also found that celestial objects are moving around in the sky with velocities changing. He then proposed the law of universal gravitation between two objects with masses M_g and m_g separated by a displacement $\mathbf{r}$,

$$\mathbf{F} = -\frac{GM_g m_g}{r^2}\hat{\mathbf{r}}.$$

We will call m_g the gravitational mass. The **equivalence principle** states that the gravitational mass and the inertial mass are in fact equivalent. Then one is able to define other quantities of interest for Newtonian gravity:

1. Field: In Newtonian gravity, the gravitational field (or acceleration) $\mathbf{g}$ is defined as the force per unit mass:

   $$\mathbf{g} = -\frac{GM}{r^2}\hat{\mathbf{r}}.$$

2. Potential: Gravity is a conservative force. The gravitational potential Φ is defined as

   $$\Phi = -\frac{GM}{r},$$

   It is related to the gravitational field by

   $$\mathbf{g} = -\nabla\Phi \quad \Longleftrightarrow \quad \Phi(\mathbf{r}_b) - \Phi(\mathbf{r}_a) = -\int_{\mathbf{r}_a}^{\mathbf{r}_b} \mathbf{g}\, d\boldsymbol{\ell}.$$

   Since gravity is conservative, the path integral of Φ (which is the work done on the particle) from $\mathbf{r}_a$ to $\mathbf{r}_b$ is path independent.

3. Linearity: Newtonian gravity is linear, so the gravitational field obeys the principle of superposition. The total potential due to a mass distribution is equal to the sum of the potentials from individual small masses:

   $$\Phi(\mathbf{r}) = -\sum_i \frac{GM}{|\mathbf{r} - \mathbf{r}_i|} \quad \text{and} \quad \Phi(\mathbf{r}) = -\int \frac{G\rho(\mathbf{r}')\, d^3\mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|},$$

   where ρ stands for the mass density of the distribution.

4. Gauss's law and Poisson's equation: these are the gravitational counterpart of Gauss's law and Poisson's equation for electrostatics. Gauss's law:

   $$\nabla \cdot \mathbf{g} = -4\pi G\rho \quad \Longleftrightarrow \quad \oint_S \mathbf{g} \cdot d\mathbf{S} = -4\pi GM. \tag{6.1}$$

Poisson's eqution:

$$\nabla^2\Phi = 4\pi G\rho. \tag{6.2}$$

6.1.2 Einstein's Theory of Relativity

In 1905, Einstein published his special theory of relativity. In special relativity, there are two postulates:

1. All laws of physics are the same in all inrtial frames.
2. The speed of light in vacuum is c in all inertial frames.

The constancy of the speed of light is supported by the Michelson and Morley experiment in 1887. With these two postulates, one can derive the **Lorentz transformation** between two inertial frames S and S'.

$$\begin{aligned} ct' &= \gamma\left(ct - \frac{vx}{c}\right), \\ x' &= \gamma(x - vt), \\ y' &= y, \\ z' &= z, \end{aligned} \tag{6.3}$$

where $\gamma = 1/\sqrt{1 - v^2/c^2}$ is the **Lorentz factor**. The Lorentz transformation says that space and time are mixed up when we make coordinate transformations. Therefore in relativity, we consider space and time equally, combining them into a 4-dimensional spacetime. The line element, according to Lorentz transformation, is specified by the **Minkowski metric**,

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2.$$

Again, the line element is invariant under all coordinate transformations. However, even though this space is flat, it is no longer Euclidean. Because of the minus sign before c^2dt^2, the distance measured in this space can be positive (spacelike), negative (timelike), or zero (null or lightlike). For cosmology, all we need to keep in mind is that light travels along **null geodesics**—light travels along the "shortest path" in this space with $ds^2 = 0$,

$$ds^2 = 0 \quad \Longrightarrow \quad c^2dt^2 = d\mathbf{x}^2 \quad \Longrightarrow \quad \left|\frac{d\mathbf{x}}{dt}\right| = c.$$

Ten years later, Einstein worked out his general theory of relativity, incorporating gravity into special relativity. Consider two frames, each of which has an identical particle in it:

1. A frame with a downward gravitational field g, e.g. the surface of the Earth.
2. A frame with an upward acceleration with magnitude $a = g$, e.g. a rocket in free space accelerating.

The second frame is a non-inertial frame. If we stick to this frame, we will see the particle accelerating downward, as if it is in a gravitational field. Einstein argues that by the equivalence principle, one in the frame cannot tell whether the frame is accelerated or is in a gravitational field. This led Einstein to general relativity.

Now consider the same two frame, but now there is a laser beam on the left wall that emits light horizontally. If you are in the second frame, light bends downward because the frame is accelerating upward. By the equivalence principle, the light beam in the frame with gravity should also bend downward. Since the two frames are equivalent, we arrive at a very astonishing conclusion: light in a gravitational field follows a curved path. Because light travels along null geodesics, this is saying that gravity curves space or spacetime.

Einstein said that the presence of mass and energy (which are interchangeable via $E = mc^2$) produce gravity. Meanwhile, gravity curves spacetime. Therefore, Einstein concluded that there is no forces acting on an object in a gravitational field; it is just folloing its geodesic in curved spacetime, independent of its composition. Einstein's idea can be summarized into two sentences by John Wheeler:

Spacetime tells matter how to move; matter tells spacetime how to curve.

6.1.3 The Metric and Curvature

Now we somewhat understand gravity. We need a tool to describe the curvature of spacetime. First, by the cosmological principle, the universe on a large scale is isotropic and homogeneous. Thus, the curvature of the whole universe should be either flat, or uniformly positive, or uniformly negative.

In a 3-dimensional Euclidean space, we are able to visualize 2-dimensional curved surface. (An N-dimensional surface is a surface that is parametrized by N independent variables.) A 3-dimensional curved surface need to be embedded in a 4-dimensional space, which is impossible to visualize in 3 dimensions. But we can still get some intuition of what are positively curved and negatively curved space using 2-dimensional curved surface. A uniformly, positively curved 2D surface is the surface of a sphere. The sum of the interior angles of triangles on a sphere is always greater than π. If the sphere has a radius R, then its line element can be expressed in spherical coordinates:

$$ds^2 = R^2 d\theta^2 + R^2 \sin^2\theta\, d\phi^2,$$

where θ is the polar angle and ϕ is the azimuth. A surface of uniform negative curvature cannot be constructed in a 3D Euclidean space. A typical example of a negatively, non-uniformly curved surface is a hyperboloid.

It is straightforward to write down the line element of a three-dimensional flat space:

$$ds^2 = dx^2 + dy^2 + dz^2 = dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2).$$

A three-dimensional space with uniform positive curvature has the line element

$$ds^2 = dr^2 + R^2 \sin^2(r/R)[d\theta^2 + \sin^2\theta\, d\phi^2],$$

and that with uniform negative curvature has line element

$$ds^2 = dr^2 + R^2 \sinh^2(r/R)[d\theta^2 + \sin^2\theta\, d\phi^2].$$

These three cases can be summarized into one compact form:

$$ds^2 = dr^2 + S_\kappa(r)^2 d\Omega^2, \tag{6.4}$$

where

$$d\Omega^2 \equiv d\theta^2 + \sin^2\theta\, d\phi^2 \quad \text{and} \quad S_\kappa(r) = \begin{cases} R\sin(r/R) & (\kappa = +1) \\ r & (\kappa = 0) \\ R\sinh(r/R) & (\kappa = -1) \end{cases}$$

The parameter R is known as the **radius of curvature**.

6.2 The Robertson-Walker Metric

To study the universe with relativity, one need to incorporate the time component of spacetime. In the 1930s, the physicists Howard Robertson and Arthur Walker derived (independently) a metric that is spatially homogeneous and isotropic at all time:

$$\boxed{ds^2 = -c^2 dt^2 + a(t)^2 \left[dr^2 + S_\kappa(r)^2 d\Omega^2\right].} \tag{6.5}$$

This is known as the **Robertson-Walker metric**. It is the only metric we will study in this notes. Note that according to this metric, the distance are allowed to expand or contract as a function of time. The factor before the spatial line element is the scale factor $a(t)$ squared. Also, this metric does not guarantee that the curvature is constant over all time. The curvature is constant in space at any *particular* time.

The time coordinate is known as the cosmological proper time, or the **cosmic time**. It is the time measured by an observer who sees the universe expandnig uniformly around them. The spatial variables (r, θ, ϕ) are the **comoving coordinates** of a point in space. The comoving coordinates of any point is constant if the expansion of the universe is perfectly homogeneous and isotropic. In other words, the comoving distance factors out $a(t)$ and is only related to the objects own motion in the universe. For example, a galaxies velocity without accounting for the expansion of the universe will change its comoving distance to us.

6.2.1 Proper Distance

The **proper distance** is the distance between two sets of comoving coordinates at the *same* cosmic time. The proper distance *do* take the scale factor into account. For example, suppose we set our position to be the origin and observe a galaxy at a comoving coordinate (r, θ, ϕ). Due to the expanding (or contracting) universe, the galaxy's proper distance is

$$ds^2 = a(t)^2[dr^2 + S_\kappa(r)^2 d\Omega^2].$$

Usually θ and ϕ will be constant, so we have $ds = a(t)\, dr$. The proper distance to the galaxy is given by

$$d_p(t) = a(t) \int_0^r dr = a(t)r. \tag{6.6}$$

Assume that the comoving coordinate r is constant. The rate of change of the proper distance to the galaxy is

$$\dot{d}_p = \dot{a}r = \frac{\dot{a}}{a} d_p.$$

At current time $t = t_0$, this gives a linear relation between the proper distance to a galaxy and its recession speed:

$$v_p(t_0) = H_0 d_p(t_0),$$

where $H_0 = \dot{a}(t_0)/a(t_0)$ is the Hubble constant now. As the scale factor changes, so does the Hubble parameter. Hence the Hubble parameter is defined for all t as

$$\boxed{H(t) = \frac{\dot{a}(t)}{a(t)}.} \tag{6.7}$$

6.2.2 Cosmological Redshift

In general, as one look at a galaxy, it is only possible to know its angular position (θ, ϕ). To measure the proper distance, we need to make use of redshifts. Recall that the light travels along the null geodesic, $ds^2 = 0$. By symmetry, in a homogeneous and isotropic universe, its trajectory should be along constant θ and ϕ lines. This reduces (6.5) to

$$c^2 dt^2 = a(t)^2 dr^2 \quad \Longrightarrow \quad c\frac{dt}{a(t)} = dr.$$

Suppose the galaxy emits light with a wavelength λ_e. The wave crest of the emitted light is emitted at a time t_e and observed at a time t_0. Then

$$c\int_{t_e}^{t_0} \frac{dt}{a(t)} = \int_0^r dr = r. \tag{6.8}$$

The next wave crest is emitted at a time $t_e + \lambda_e/c$, and is observed at $t_0 + \lambda_0/c$, where $\lambda_0 \neq \lambda_e$ in general. The second wave crest obeys the equation

$$c\int_{t_e+\lambda_e/c}^{t_0+\lambda_0/c} \frac{dt}{a(t)} = \int_0^r dr = r.$$

(The r does not change because it represents the comoving distance.) Comparing the two integral equations, we find that

$$\int_{t_e}^{t_0} \frac{dt}{a(t)} = \int_{t_e+\lambda_e/c}^{t_0+\lambda_0/c} \frac{dt}{a(t)} \quad \Longrightarrow \quad \left(\int_{t_e}^{t_e+\lambda_e/c} + \int_{t_e+\lambda_e/c}^{t_0}\right)\frac{dt}{a(t)} = \left(\int_{t_e+\lambda_e/c}^{t_0} + \int_{t_0}^{t_0+\lambda_0/c}\right)\frac{dt}{a(t)}.$$

Canceling the same integrals on both sides,

$$\int_{t_e}^{t_e+\lambda_e/c} \frac{dt}{a(t)} = \int_{t_0}^{t_0+\lambda_0/c} \frac{dt}{a(t)}.$$

The integral of $dt/a(t)$ between the emission of successive wave crests is equal to the integral of $dt/a(t)$ between the obsevation of the same two wave crests. Evidently, the universe cannot expand much for a time interval λ_c/c, so $a(t)$ can be treated as a constant in such an interval. We can safely pull $a(t)$ out of the integrals and evaluate it at emission and observation times:

$$\frac{1}{a(t_e)}\int_{t_e}^{t_e+\lambda_e/c} dt = \frac{1}{a(t_0)}\int_{t_0}^{t_0+\lambda_0/c} dt \quad \Longrightarrow \quad \frac{\lambda_e}{a(t_e)} = \frac{\lambda_0}{a(t_0)}.$$

By the definition of redshift (5.1), the redshift of light from a distant galaxy is related to the scale factor *at the time of emission* by

$$\boxed{1 + z = \frac{1}{a(t_e)},} \tag{6.9}$$

where we used the convention $a(t_0) = 1$. The redshift does not depend on how $a(t)$ changes with time between $a(t_e)$ and $a(t)$. Only the endpoints matter. This simple relation (6.9) will be one of the most useful equation in this notes.

6.2.3 Angular Distance

In a constantly curved space, the sum of the interior angles α, β, and γ of a triangle is related to its area A by

$$\alpha + \beta + \gamma = \pi + \frac{\kappa A}{R_0^2},$$

where R_0 is the radius of the curvature. In principle, if we want to measure the curvature of the universe, we just draw a very large triangle and compair $\alpha + \beta + \gamma$ with π. If $\alpha + \beta + \gamma > \pi$, then the universe is positively curved, or if $\alpha + \beta + \gamma < \pi$, it is negatively curved. In reality, it is impossible to draw and measure such a large triangle because the radius of curvature R_0 is much larger than we can imagine.

In fact, R_0 cannot be much smaller than the current Hubble distance $c/H_0 \approx 4380\,\text{Mpc}$. Suppose we measure two ends of a galaxy to have a diameter D and a distance r from the Earth. Moreover, assume that this galaxy is at constant ϕ lines ($d\phi = 0$) and constant r lines ($dr = 0$). Then according to the spatial part of Robertson-Walker metric (6.5),

$$ds^2 = a(t_0)[dr^2 + S_\kappa(r)^2 d\Omega^2] \quad \Longrightarrow \quad D^2 = S_\kappa(r)^2\theta^2.$$

In the limit of $D \ll r$, the angular size θ of the galaxy should be $\theta = D/r$ in flat space. In a positvely curved universe, the angular size is

$$\theta_+ = \frac{D}{R_0 \sin(r/R_0)}.$$

When $r < \pi R_0$, the angular size of the galaxy appears larger than that in flat space, $\alpha_+ > D/r$. At $r = \pi R_0$, the angular size blows up. Equivalently, this is saying that the galaxy is everywhere in the sky. Evidently there is no such measurement made, which means $r = \pi R_0 \gtrsim c/H_0$. The radius of curvature must be comparable or larger than the Hubble distance. If the universe is negatively curved, then the observed angular size should be

$$\theta_- = \frac{D}{R_0 \sinh(r/R_0)} < \frac{D}{r}.$$

At $r \gg R_0$, $\sinh(r/R_0) \approx e^{r/R_0}/2$, so

$$\theta_- \approx \frac{2D}{R_0} e^{-r/R_0}.$$

Thus, in a negatively curved universe, galaxies that are much farther than R_0 have angular size exponentially suppressed. This is also not observed from galaxies at $r \sim c/H_0$, which means the radius of curvature must have $R_0 \gtrsim c/H_0$.

6.3 The Friedmann Equation

6.3.1 Einstein's Field Equation

In 1915, Einstein published his general theory of relativity, relating the curvature of spacetime to mass-energy. In other words, the curvature and mass-energy density of the universe can infer each other. The key equation is Einstein's field equation:

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}. \tag{6.10}$$

The LHS is the Einstein tensor, a 4×4 symmetric tensor $G_{\mu\nu} = G_{\nu\mu}$ with ten independent components. It describes the curvature of spacetime. Basically, it is some form of second derivatives of the metric. On the RHS is the stress-energy tensor or the energy momentum tensor $T_{\mu\nu}$. It is also a 4×4 symmetric tensor, describing the mass-energy density and pressure. Einstein's field equation is extremely hard to solve because it consists of ten coupled nonlinear second-order partial differential equations.

In a homogeneous and isotropic universe, the stress-energy tensor $T_{\mu\nu}$ is relatively simple. It is related to the energy density $\epsilon(t)$ and pressure $P(t)$ of materials in the universe. If we use the Robertson-Walker metric (6.5) to describe the curvature of spacetime, then the solution to Einstein's equation is the **Friedmann equation**:

$$\boxed{\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\epsilon}{3c^2} - \frac{\kappa c^2}{R_0^2 a^2}.} \tag{6.11}$$

The goal of (6.11) is to find the scale factor $a(t)$ and its relation to κ, ϵ, and R_0.

6.3.2 A Cheat by Newtonian Gravity

Though the Friedmann equation is completely derived from Einstein's field equation, we can still have a good intuition of it from a Newtonian perspective. We will start from Newtonian gravity to derive the non-relativistic version of the Friedmann equation, and then add relativity to make it correct.

Consider a homogeneous sphere of matter of total mass M_s that is constant with time. The sphere can expand or contract isotropically, so its radius $R_s(t)$ depends on time. The gravitational force experienced by a mass m at the surface of the sphere is given by

$$F = -\frac{GM_s m}{R_s^2}.$$

By Newton's second law, the acceleration at the surface of the sphere is

$$\frac{d^2R_s}{dt^2} = -\frac{GM_s}{R_s^2}.$$

Multiplying both sides by dR_s/dt and integrating,

$$\frac{d^2R_s}{dt^2}\frac{dR_s}{dt} = -\frac{GM_s}{R_s^2}\frac{dR_s}{dt} \quad\Longrightarrow\quad \frac{1}{2}\frac{d}{dt}\left(\frac{dR_s}{dt}\right)^2 = \frac{d}{dt}\left(\frac{GM_s}{R_s}\right) \quad\Longrightarrow\quad \frac{1}{2}\left(\frac{dR_s}{dt}\right)^2 = \frac{GM_s}{R_s} + U,$$

where U is some integration constant. Then multiply both sides by m, the equation becomes much clearer,

$$\frac{1}{2}m\dot{R}_s^2 - \frac{GM_s m}{R_s} = U. \tag{6.12}$$

The LHS is the sum of the kinetic energy and potential energy of mass m, and the sum is the total energy. Now write M_s in terms of density ρ and radius $R_s(t)$, $M_s = 4\pi\rho R_s^3/3$. The radius $R_s(t)$ can be written as $R_s(t) = a(t)r_s$, where $a(t)$ is the scale factor and r_s is the comoving radius of the sphere. We can do this because the expansion or contraction is isotropic and the universe is homogeneous.

Substitute M_s and R_s into (6.12),

$$\frac{1}{2}r_s^2\dot{a}^2 = \frac{4\pi}{3}Gr_s^2\rho a^2 + U.$$

Dividing both sides by $r_s^2a^2/2$ gives

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3} + \frac{2U}{a^2r_s^2}.$$

This equation resembles the Friedmann equation (6.11) a lot already. Before discussing the flaws of this equation, we shall first analyze this equation first. Assume that the equation is expanding, just like our universe. There are three cases:

1. $U > 0$: the RHS of the equation is always positive. Then $\dot{a}^2$ is always positive as there is always a positive U on the RHS—the expansion never stops.

2. $U < 0$: the RHS starts out positive. Once ρ is small enough so that the RHS becomes zero, the scale factor reaches some maximum,

$$a_{\max} = -\frac{GM_s}{Ur_s}.$$

The sphere then contracts as $\ddot{a} < 0$ at maximum.

3. $U = 0$: this is the boundary case. Since $\rho \propto a^{-3}$, it goes to zero as a becomes larger, and $\dot{a}$ goes to zero as well.

Incorporating Relativity. Comparing Newtonian form of the Friedmann equation with (6.11), some c's are missing because we neglect relativistic effects. The mass density ρ can be converted into the energy density ϵ by $\epsilon = \rho c^2$ in the low velocity limit $v \ll c$. The only difference now is the constant term U. We need to have

$$\frac{2U}{r_s^2} = -\frac{\kappa c^2}{R_0^2}.$$

Though we cannot obtain this term using any method, it makes sense in the following cases. We know that U is proportional to the total energy of the test mass from (6.12). When $U < 0$, the universe is *bounded* in classical sense. Meanwhile, $U < 0$ means that $\kappa = +1$, a positively curved space is also bounded (think about the sphere). Similarly, when $U < 0$, the universe is *unbounded* and $\kappa = -1$, where the curved space is not bounded (think about the hyperboloid). Finally, $U = 0$ represents a perfectly flat space, $\kappa = 0$, which is the boundary of the two cases above.

Despite all those intuitive reasoning, the Friedmann equation in Newtonian form cannot represent the universe: A sphere of finite radius has a special point (its center). It breaks the assumption of homogeneity and isotropy at the start while we keep using it throughout the analysis. If we consider the sphere of radius R_s inside an infinite, homogeneous, isotropic universe, then Newtonian gravity tells that a test mass experiences no force inside such a system. Moreover, when we convert ρ into ϵ, we are neglecting energy from photons as they are massless.

6.3.3 Critical Density

To link the Friedmann equation to some measurable quantities describing the expansion of the universe, we can replace $\dot{a}/a$ by $H(t)$, the Hubble parameter. The Friedmann equation then can be written in the form

$$H(t)^2 = \frac{8\pi G\epsilon}{3c^2} - \frac{\kappa c^2}{R_0^2 a^2}. \tag{6.13}$$

At the present, the Friedmann equation is

$$H_0^2 = \frac{8\pi G\epsilon_0}{3c^2} - \frac{\kappa c^2}{R_0^2},$$

where $H_0 = (68 \pm 2)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ is the usual Hubble constant. At spatially flat universe with $\kappa = 0$, the Friedmann equation has the form

$$H(t)^2 = \frac{8\pi G\epsilon}{3c^2} \quad \Longrightarrow \quad \epsilon_c(t) \equiv \frac{3c^2}{8\pi G} H(t)^2. \tag{6.14}$$

This energy density $\epsilon(t)$ is assigned as the **critical density** $\epsilon_c(t)$ of the universe for a given Hubble parameter $H(t)$. At the present, the critical density is at

$$\epsilon_{c,0} = \frac{3c^2}{8\pi G} H_0^2 = (7.8 \pm 0.5) \times 10^{-10}\,\mathrm{J/m^3} = (4870 \pm 290)\,\mathrm{MeV/m^3}.$$

The equivalent mass density (at the present) is

$$\rho_{c,0} \equiv \frac{\epsilon_{c,0}}{c^2} = (8.7 \pm 0.5) \times 10^{-27}\,\mathrm{kg/m^3} = (1.28 \pm 0.08) \times 10^{11} M_\odot/\mathrm{Mpc}^3.$$

The critical density is roughly equivalent to one proton per $0.2\,\mathrm{m}^3$. This is much lower than the density of the interstellar space within the Milky Way galaxy. We should remember that such a low density is averaged over the whole universe with intergalactic voids.

To see why ϵ_c is called the critical density, we define a dimensionless parameter called the **density parameter**,

$$\Omega(t) \equiv \frac{\epsilon(t)}{\epsilon_c(t)}. \tag{6.15}$$

We can write the Friedmann equation in terms of Ω. Using (6.13),

$$H(t)^2 = \frac{8\pi G\epsilon}{3c^2} - \frac{\kappa c^2}{R_0^2 a^2} \quad \Longrightarrow \quad 1 = \frac{8\pi G\epsilon}{3c^2 H^2} - \frac{\kappa c^2}{R_0^2 a^2 H^2}.$$

Now use the definition of the critical density (6.14),

$$1 = \frac{\epsilon}{\epsilon_c} - \frac{\kappa c^2}{R_0^2 a^2 H^2}$$

and the definition of the density parameter (6.15),

$$\boxed{1 - \Omega(t) = -\frac{\kappa c^2}{R_0^2 a^2 H^2}.} \tag{6.16}$$

Note that the RHS of (6.16) cannot change sign once κ is set. This means if $\Omega < 1$ (or $\epsilon < \epsilon_c$) at any time, it remains less than 1 at any time; if $\Omega > 1$ at any time, it is greater than 1 at any time, and similar for $\Omega = 1$ at all times. The current value of Ω_0 is between $0.995 < \Omega_0 < 1.005$. That is, we do not know whether the curvature is positive, negative, or zero yet, but it is very close to flat.

6.4 The Fluid and Acceleration Equations

The Friedmann equation have two unknowns: the scale factor $a(t)$ and the energy density $\epsilon(t)$. We need another equation to model the universe.

The first law of thermodynamics says

$$dQ = dE + P\,dV \tag{6.17}$$

If the universe is homogeneous, then there is no heat flow for any volume, $dQ = 0$. In an adiabatic ($dQ = 0$) process, the entropy within the volume does not change, $dS = dQ/T = 0$. Hence a homogeneous, isotropic expansion of the universe does not increase its entropy. Taking the time-derivative on both sides of (6.17) gives

$$\dot{E} + P\dot{V} = 0. \tag{6.18}$$

Consider a sphere of comoving radius r_s, so its proper radius is $R_s(t) = r_s a(t)$. The volume and rate of change of the volume are

$$V(t) = \frac{4\pi}{3} r_s^3 a(t)^3 \quad \implies \quad \dot{V} = \frac{4\pi}{3} r_s^3 (3a^2 \dot{a}) = 3V\frac{\dot{a}}{a}.$$

The internal energy within the sphere is related to energy density by $E(t) = V(t)\epsilon(t)$, so the rate of change of the internal energy is

$$\dot{E} = V\dot{\epsilon} + \dot{V}\epsilon = V\left(\dot{\epsilon} + 3\epsilon\frac{\dot{a}}{a}\right).$$

Substituting $\dot{V}$ and $\dot{E}$ into (6.18) yields

$$\boxed{\dot{\epsilon} + 3(\epsilon + P)\frac{\dot{a}}{a} = 0.} \tag{6.19}$$

This is the **fluid equation** or the **continuity equation**. It tells how the energy density of the universe changes with time.

Now we can combine the fluid equation and the Friedmann equation. Multiplying the Friedmann equation (6.11) by a^2 on both sides gives

$$\dot{a}^2 = \frac{8\pi G}{3c^2}\epsilon a^2 - \frac{\kappa c^2}{R_0^2}.$$

Taking the time-derivative and dividing by $2\dot{a}a$,

$$2\dot{a}\ddot{a} = \frac{8\pi G}{3c^2}(\dot{\epsilon}a^2 + 2\epsilon a\dot{a}) \quad \implies \quad \frac{\ddot{a}}{a} = \frac{4\pi G}{3c^2}\left(\dot{\epsilon}\frac{a}{\dot{a}} + 2\epsilon\right).$$

The fluid equation (6.19) says that

$$\dot{\epsilon}\frac{a}{\dot{a}} = -3(\epsilon + P).$$

Subtituting in, we get the **acceleration equation** (also called the **second Friedmann equation**):

$$\boxed{\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(\epsilon + 3P).} \tag{6.20}$$

Since a is always positive, the sign of $\ddot{a}$ depends only on the RHS, or explicitly, the energy density and pressure of components of the universe. There are now three equations: the Friedmann equation, fluid equation, and acceleration equation. The first two are independent, and the third can be derived from them. However, we also introduced a new unknown in addition to $a(t)$ and $\epsilon(t)$, which is the pressure $P(t)$. Different components in the universe (photons, baryonic matter, dark matter, etc.) have different contributions to pressure. The relations of pressure and energy density are the equations of state.

6.5 Equations of State

In general, the equations of state $P(\epsilon)$ can be a complicated nonlinear function of the energy density ϵ. Fortunately, for substances important to cosmology, the equations of state are of the form

$$P = w\epsilon,$$

where w is a dimensionless number.

6.5.1 Matter and Radiation

The most familiar equation of state should be the ideal gas law

$$P = \frac{\rho}{\bar{m}}kT \quad \Longleftrightarrow \quad P = \frac{kT}{\bar{m}c^2}\epsilon,$$

which models a low-density gas of non-relativistic particles. Here ρ and $\bar{m}$ are the mass density and the average mass of gas particles, respectively. For non-relativistic particles, the energy is almost entirely contributed by the rest mass energy, so we can safely write $\rho = \epsilon/c^2$. By the equipartition theorem (A.1.1), the temperature T is related to $\langle v^2 \rangle$ by $3kT = \bar{m}\langle v^2 \rangle$. Therefore, we have

$$P_{\text{ideal}} = w\epsilon_{\text{ideal}} \quad \text{where} \quad w \approx \frac{\langle v^2 \rangle}{3c^2} \ll 1.$$

Since the ideal gas is non-relativistic, $\langle v^2 \rangle \ll c^2$. For example, a low-density electron gas requires $T \sim 10^9\,\text{K}$ to be relativistic. A low-density proton gas requires even higher temperature, $T \sim 10^{13}\,\text{K}$. At the present moment, the universe does not have such a high temperature as a whole.

Another common type of gas is the photon gas. The equation of state of a photon gas is given by the radiation pressure (2.11),

$$P_{\text{rad}} = \frac{1}{3}\epsilon_{\text{rad}}.$$

This result is exact, so $w = 1/3$ for a photon gas (or other ultra-relativistic gases such as a neutrino gas). The ideal gas law and radiation pressure sets the range of w for all ordinary matter: $0 < w < 1/3$. For simplicity, we call non-relativistic particles "matter" and relativistic particles "radiation".

6.5.2 Dark Energy and the Cosmological Constant

In principle, there are other cases of w that do not lie within this range. The case $w < -1/3$ is interesting because according to the acceleration equation (6.20) it predicts a positive acceleration, $\ddot{a} > 0$. Components of the universe that have $w < -1/3$ are referred to as **dark energy**.

By 1915, after Einstein published his general theory of relativity, he wanted to apply general relativity to model the universe. He observed that the universe was full of matter and radiation. However, nobody, including Einstein, was aware of the cosmic microwave background at that time. He thought that most radiation comes from starlight, and the rest mass energy of matter dominates over radiation. Thus, Einstein concluded that the universe is pressureless, as matter has $w \ll 1$.

At that time, astronomers are unable to observe the recession distant galaxies. There was no evidence that suggests whether the universe is expanding or contracting. Thus, Einstein believed that the universe is static, with a positive energy density but negligible pressure. But the acceleration equation says that if $\epsilon > 0$ and $P \approx 0$, then the universe must not be static. It either expand forever or reach a maximum radius and the contract.

To solve this problem, Einstein introduces the **cosmological constant** Λ in his Field equation,

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu} + \Lambda g_{\mu\nu}, \tag{6.21}$$

where $g_{\mu\nu}$ is the **metric tensor**, a tensor that defines the line element of a (curved) spacetime. With this cosmological constant, the Friedmann equation (6.11) and acceleration equation (6.20) become

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\epsilon}{3c^2} - \frac{\kappa c^2}{R_0^2 a^2} + \frac{\Lambda c^2}{3} \quad \text{and} \quad \frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(\epsilon + 3P) + \frac{\Lambda c^2}{3}.$$

The fluid equation is unaffected because it comes from the first law of thermodynamics, the same in both Newtonian gravity and Einstein gravity. Adding the cosmological constant to the Friedmann equation is equivalent to adding a new component to the universe with energy density

$$\epsilon_\Lambda \equiv \frac{c^4}{8\pi G}\Lambda. \tag{6.22}$$

(We can see this from Einstein's field equation (6.21) directly. Since $T_{\mu\nu}$ represent energy density, adding a Λ is equivalent to adding additional energy density ϵ_Λ.) If Λ is constant over time, then so is ϵ_Λ. According to the fluid equation, to have $\dot{\epsilon}_\Lambda = 0$, the cosmological constant must have an associated pressure

$$P_\Lambda = -\epsilon_\Lambda = -\frac{c^4}{8\pi G}\Lambda.$$

With the cosmological constant, Einstein got the static model universe with $\dot{a} = 0$ and $\ddot{a} = 0$. If $\ddot{a} = 0$ and $P = 0$ for a matter dominated static universe, the acceleration equation says that

$$0 = -\frac{4\pi G\epsilon}{3c^2} + \frac{\Lambda c^2}{3} \quad \Longrightarrow \quad \Lambda = \frac{4\pi G\rho}{c^2},$$

where ρ is the mass density. If $\dot{a} = 0$, then the Friedmann equation becomes

$$0 = \frac{8\pi G\rho}{3} - \frac{\kappa c^2}{R_0^2} + \frac{\Lambda c^2}{3} = 4\pi G\rho - \frac{\kappa c^2}{R_0^2}.$$

Therefore, Einstein's static universe must be positively curved ($\kappa = +1$), with a radius of curvature

$$R_0 = \frac{c}{\sqrt{4\pi G\rho}} = \Lambda^{-1/2}. \tag{6.23}$$

Einstein published his static model of the universe in 1917, but this model has a significant shortcoming. This model is an unstable equilibrium with the repulsive force of Λ balancing the attractive force of ρ. We know that ϵ_Λ (or Λ itself) is constant. If Einstein's static universe expands a little, the energy density of matter lowers. Cosmological constant dominates over matter, and the universe expands further. Similarly, compressing Einstein's static universe causes a runaway collapse.

Modern observation shows that the expansion of the universe has a positive acceleration, so Einstein's static universe was abandoned. Nonetheless, the cosmological constant Λ was kept, and many physicists were wondering what Λ represents physically. The most popular answer is the **vacuum energy**. The vacuum energy is calculated in quantum field theory (which is not unified with general relativity) to be

$$\epsilon_{\text{vac}} \sim 10^{132}\,\text{eV/m}^3.$$

Recall that the critical density of the universe at the present is $\epsilon_{c,0} \sim 10^9\,\text{eV/m}^3$. The vacuum energy predicted by theory is 123 orders of magnitude larger! Evidently, we do not have much understanding on vacuum energy yet.

7 Model Universes

We are now equipped with three equations to model the Universe:

- The Friedmann equation:
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\epsilon}{3c^2} - \frac{\kappa c^2}{R_0^2 a^2}.$$
- The fluid equation:
$$\dot{\epsilon} + 3(\epsilon + P)\frac{\dot{a}}{a} = 0.$$
- The acceleration equation:
$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(\epsilon + 3P).$$

The Friedmann equation and fluid equation are independent. We derived the acceleration equation from them. In addition, we have the equation of state $P = w\epsilon$ for different components of the universe. With appropriate initial conditions, we can solve for $\epsilon(t)$, $P(t)$, and $a(t)$ for all times in principle.

7.1 Evolution of Energy Density

First, we need to know that the energy density and pressure for different components are additive,

$$\epsilon = \sum_i \epsilon_i \quad \text{and} \quad P = \sum_i w_i \epsilon_i.$$

If so, the fluid equation must hold for each component separately,

$$\dot{\epsilon}_i + 3(\epsilon_i + P_i)\frac{\dot{a}}{a} = 0 \quad \Longrightarrow \quad \dot{\epsilon}_i + 3\epsilon_i(1 + w_i)\frac{\dot{a}}{a} = 0.$$

Rearranging some term, we can get a differential equation relating ϵ_i and a,

$$\frac{d\epsilon_i}{\epsilon_i} = -3(1 + w_i)\frac{da}{a}.$$

If we assume w_i is constant for each component, then the evolution of the energy density is

$$\epsilon_i(a) = \epsilon_{i,0} a^{-3(1+w_i)}. \tag{7.1}$$

where the integration constant $\epsilon_{i,0}$ is the energy density of the ith component today. This result is solely from the fluid equation and the equation of state. It has nothing to do with the Friedmann equation yet. With (7.1), we can obtain the energy density of matter and radiation using $w = 0$ for matter and $w = 1/3$ for radiation:

$$\epsilon_m = \epsilon_{m,0} a^{-3}, \quad \epsilon_r = \epsilon_{r,0} a^{-4}. \tag{7.2}$$

These equations certainly make sense. Let $\bar{E}$ be the mean energy of particles and n be the number density. Then we can write $\epsilon = n\bar{E}$. For non-relativistic matter particles, assume that they cannot be created or destroyed. As the universe expands, the number density scales like $n \propto 1/V \propto a^{-3}$. The mean energy of these particles does not change because the energy is contributed by the rest mass energy $\bar{E} = \bar{m}c^2$. Hence the energy density also scales like $\epsilon \propto a^{-3}$.

For relativistic particles like photons, their number density dependence on the scale factor, $n \propto a^{-3}$. But unlike matter particles, the individual energy of photons does change throughout the expansion because of the redshift. By (6.9), we know that $\lambda \propto a$, so the photon energy scales like $\bar{E} = hc/\lambda \propto a^{-1}$. Altogether, $\epsilon_r = n\bar{E} = n(hc/\lambda) \propto a^{-3}a^{-1} \propto a^{-4}$. However, this is based on the assumption that photons cannot be created or destroyed. This is certainly wrong, because photons are constantly absorbed or emitted by particles. The resolution is that the cosmic microwave background has the energy density larger than all photos emitted by all stars in the history of the universe. The energy density of the CMB at the present is

$$\epsilon_{\text{CMB},0} = a_{\text{rad}} T^4 = 0.2606\,\text{MeV/m}^3.$$

In terms of the density parameter, $\Omega_{\text{CMB},0} = 5.35 \times 10^{-5}$. This is small compared to the critical density, but let's see what is the energy of stars. The observed present luminosity density of galaxies is

$$\Psi \approx 1.7 \times 10^8\, L_\odot/\text{Mpc}^{-3} \approx 2.2 \times 10^{-33}\,\text{W/m}^3.$$

Assume that the galaxies has been emitting light at this rate for the age of the universe $t_0 \approx H_0^{-1} \approx 4.5 \times 10^{17}\,\text{s}$. The energy density in starlight is roughly $\epsilon_{\text{starlight},0} \sim \Psi t_0 \sim 0.006\,\text{MeV/m}^3$, which is much smaller than that of the CMB.

The cosmic microwave background is a remnant of photons that lived in a time when the universe was hot and dense enough to be opaque to photons. Far earlier, there should also be a time when the unverse was hot and dense enough to be opaque to neutrinos. There should be a cosmic neutrino background today. The energy density of this background calculated for each neutrino flavor is calculated to be

$$\epsilon = \frac{7}{8}\left(\frac{4}{11}\right)^{4/3} \epsilon_{\text{CMB}} = 0.227\epsilon_{\text{CMB}}.$$

Three flavors of neutrino gives $\epsilon_\nu = 0.681\epsilon_{\text{CMB}}$, or $\Omega_\nu = 0.681\Omega_{\text{CMB}}$. Thus, the density paramter in radiation should be $\Omega_{\text{rad},0} = \Omega_{\text{CMB},0} + \Omega_{\nu,0} = 9.00 \times 10^{-5}$. The energy density of matter and that of dark energy are not well determined. Currently, we believed that $\Omega_{\text{mat}} \approx 0.31$, and $\Omega_{\Lambda,0} \approx 0.69$. A model that fit these observations well is the Benchmark Model, which will be discussed in Section 7.4 and 7.5.

As we can see, the energy density of radiation is negligible compared to that of matter and dark energy. The ratio of the energy density in Λ to that in matter today is

$$\frac{\epsilon_{\Lambda,0}}{\epsilon_{m,0}} = \frac{\Omega_{\Lambda,0}}{\Omega_{m,0}} \approx 2.23.$$

We say that the cosmological constant is dominant over matter today. However, if we put in the scale factor dependence according to (7.2),

$$\frac{\epsilon_\Lambda(a)}{\epsilon_m(a)} = \frac{\epsilon_{\Lambda,0}}{\epsilon_{m,0}/a^3} = \frac{\Omega_{\Lambda,0}}{\Omega_{m,0}} a^3.$$

If the universe was expanded from a very dense state, then there must be some time in the past when $\epsilon_\Lambda = \epsilon_{\text{mat}}$. This is called the **matter-Λ equality**, which occured when the scale factor was $a_{m\Lambda} = (\Omega_{\text{mat},0}/\Omega_{\Lambda,0})^{1/3} \approx 0.766$. Similarly, there is a **radiation-matter equality**:

$$1 = \frac{\epsilon_m(a)}{\epsilon_r(a)} = \frac{\epsilon_{m,0}}{\epsilon_{r,0}} a \approx 3400a \quad\Longrightarrow\quad a_{rm} \approx 2.9 \times 10^{-4}.$$

In general, if the universe was expanding and will expand in the future, then in the limit $a \to 0$, the component with the largest value of w is dominant; as $a \to \infty$, the component with the smallest value of w is dominant. The scale factor a is a monotonically increasing function of t in a continuously expanding universe, so the scale factor can be used to indicate t. And since $1 + z = 1/a(t_e)$, the redshift is also an indicator of time. For example, we can say "Matter-Lambda equality happened at a redshift $z_{m\Lambda} \approx 0.31$". This means the emitted wavelength of light at the time of matter-Λ equality is observed to be stretched by a factor of $1 + z_{\text{m}\Lambda} \approx 1.31$.

The following sections introduce different models of the universe, and we will find $a(t)$ by solving the Friedmann equation. Most of those models are wrong, but we will get some intuition from them. For each model, we will compute the age of the universe, the cosmological redshift, and the proper distance of objects.

7.2 Empty Universe

The simplest universe is an empty universe with no ϵ or Λ. The Friedmann equation for this universe is

$$\dot{a}^2 = -\frac{\kappa c^2}{R_0^2}. \tag{7.3}$$

We can immediately see one solution: an empty, static, spatially flat universe ($\dot{a} = 0$ and $\kappa = 0$). In this universe, Minkowski metric holds true. Since the LHS is nonnegative, it cannot permit positively curved universe ($\kappa = +1$). The negatively curved universe has solutions

$$\dot{a} = \pm\frac{c}{R_0} \quad\Longrightarrow\quad a(t) = \frac{t}{t_0}, \tag{7.4}$$

where $t_0 = R_0/c$ and we consider the positive (expanding) solution. For an empty universe with no acceleration, the age of the universe is exactly equal to the Hubble time, $t_0 = H_0^{-1}$. By (6.9), the redshift of a light source is

$$1 + z = \frac{1}{a(t_e)} = \frac{t_0}{t_e} \quad \Longleftrightarrow \quad t_e = \frac{t_0}{1+z} = \frac{H_0^{-1}}{1+z}.$$

Recall that the current proper distance is given by (6.6),

$$d_p(t_0) = a(t_0) \int_0^r dr = r.$$

To determine r, which is also the comoving distance, suppose light is emitted by the galaxy at t_e and observed at t_0. Light follows the null geodesic with geometry described by the Robertson-Walker metric (6.5), so

$$c^2 dt^2 = a(t)^2 dr^2 \quad \Longrightarrow \quad r = \int_0^r dr = c \int_{t_e}^{t_0} \frac{dt}{a(t)}$$

This gives the current proper distance from us to the galaxy:

$$d_p(t_0) = c \int_{t_e}^{t_0} \frac{dt}{a(t)}. \tag{7.5}$$

This is true for any universe with a Robertson-Walker metric. For empty expanding universe with $a = t/t_0$,

$$d_p(t_0) = ct_0 \int_{t_e}^{t_0} \frac{dt}{t} = ct_0 \ln\left(\frac{t_0}{t_e}\right) = \frac{c}{H_0} \ln(1+z). \tag{7.6}$$

The age of the universe is only $1/H_0$, but in this universe, we can see objects currently at an arbitrarily large distance as $d_p(t_0)$ can be much larger than c/H_0, depending on the redshift. This does not mean we can see a light emitted earlier than the birth of the universe, because the distance light actually travels is the proper distance calculated at the time of emission. The proper distance of the light source at the time of emission, $d_p(t_e)$, is scaled by a factor of $a(t_e)/a(t_0) = 1/(1+z)$. For an empty expanding universe,

$$d_p(t_e) = \frac{c}{H_0} \frac{\ln(1+z)}{1+z}.$$

Note that $d_p(t_e)$ has maximum $d_{p,\max}(t_e) = (1/e)c/H_0$ at a redshift $z = e - 1$. This means objects with higher redshift has a smaller proper distance at emission from the observer.

7.3 Single-Component Universes

A single-component universe is spatially flat ($\kappa = 0$), with only a single component with a single value of w. The Friedmann equation (6.11) reduces to the form

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G \epsilon}{3c^2}.$$

Plugging in the evolution of energy density ϵ as a function of a, we get the Friedmann equation for a flat single-component universe:

$$\dot{a}^2 = \frac{8\pi G \epsilon_0}{3c^2} a^{-(1+3w)}. \tag{7.7}$$

This equation is easy to solve by setting $a(t) = At^\xi$, i.e. we are guessing that $a(t)$ is some power of t. Substituting $a(t) = At^\xi$ into (7.7),

$$A^2 \xi^2 t^{2\xi - 2} = \frac{8\pi G \epsilon_0}{3c^2} A^{-(1+3w)} t^{-\xi(1+3w)}.$$

Both the powers of t and coefficients should match on both sides, so there are two equations: one for A and one for ξ,

$$2\xi - 2 = -\xi(1+3w), \quad A^2\xi^2 = \frac{8\pi G \epsilon_0}{3c^2} A^{-(1+3w)}.$$

If $w \neq -1$, the solutions are

$$\xi = \frac{2}{3+3w}, \quad A = \left(\frac{8\pi G \epsilon_0}{3\xi^2 c^2}\right)^{1/(3+3w)} = \left[\frac{6\pi G \epsilon_0 (1+w)^2}{c^2}\right]^{1/(3+3w)}.$$

The scale factor is then

$$a(t) = \left[\frac{6\pi G\epsilon_0(1+w)^2}{c^2}\right]^{1/(3+3w)} t^{2/(3+3w)} = \left[\sqrt{\frac{6\pi G\epsilon_0(1+w)}{c^2}}\, t\right]^{2/(3+3w)},$$

or

$$a(t) = \left(\frac{t}{t_0}\right)^{2/(3+3w)} \quad \text{where} \quad t_0 = \frac{1}{1+w}\left(\frac{c^2}{6\pi G\epsilon_0}\right)^{1/2}. \tag{7.8}$$

The Hubble constant in such a universe is

$$H_0 \equiv \left(\frac{\dot{a}}{a}\right)\bigg|_{t=t_0} = \frac{2}{3t_0(1+w)}\left(\frac{t}{t_0}\right)^{2/(3+3w)-1}\left(\frac{t}{t_0}\right)^{-2/(3+3w)}\bigg|_{t=t_0} = \frac{2}{3t_0(1+w)}.$$

Equivalently, the age of the universe in terms of the Hubble constant is

$$t_0 = \frac{2}{3(1+w)}H_0^{-1}. \tag{7.9}$$

In this 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. This make sense because recall that any component with $w < -1/3$ is considered dark energy. Dark energy accelerates the expansion of the universe, so $\dot{a}(t_0) > \dot{a}(t < t_0)$. We are underestimating the age of the universe because we are using $\dot{a}(t_0)$ as the uniform expansion rate to calculate the age of the universe. (This is the meaning of Hubble time.)

Once we get $a(t)$, we can substitute it back into the evolution equation (7.1) of energy density and get

$$\epsilon = \epsilon_0 a^{-3(1+w)} = \epsilon_0\left(\frac{t}{t_0}\right)^{-2}. \tag{7.10}$$

Since the universe is flat, we can set $\epsilon_0 = \epsilon_{c,0}$, the critical density,

$$\epsilon_0 = \epsilon_{c,0} = \frac{3c^2}{8\pi G}H_0^2 = \frac{c^2}{6\pi(1+w)^2}t_0^{-2} \quad\Longrightarrow\quad \epsilon(t) = \frac{c^2}{6\pi G(1+w)^2}t^{-2}.$$

The cosmological redshift (6.9) helps compute the emission time of light from a distance galaxy at a redshift z,

$$1+z = \frac{a(t_0)}{a(t_e)} = \left(\frac{t_0}{t_e}\right)^{2/(3+3w)} \quad\Longrightarrow\quad t_e = \frac{t_0}{(1+z)^{3(1+w)/2}} = \frac{2(1+z)^{-3(1+w)/2}}{3(1+w)H_0}.$$

Finally, the proper distance to the galaxy is (by 7.5)

$$d_p(t_0) = c\int_{t_e}^{t_0}\frac{dt}{a(t)} = \frac{3ct_0(1+w)}{1+3w}\left[1-\left(\frac{t_e}{t_0}\right)^{(1+3w)/(3+3w)}\right] = \frac{2c}{H_0(1+3w)}\left[1-\frac{1}{(1+z)^{(1+3w)/2}}\right]. \tag{7.11}$$

In a single-component universe, there is a maximum distance that we can observe at t_0. This is known as the **horizon distance**, the proper distance of the most distant object which emitted light at $t = 0$. The portion of the universe within the horizon distance is called the **visible universe**. All points in the visible universe are **causally connected** to the observer. At time t, the proper distance to a point is given by (6.6),

$$d_p(t) = a(t)\int_0^{r(t)} dr = a(t)c\int_{t_e}^{t}\frac{dt}{a(t)}.$$

Extending t_e to zero gives the horizon distance $d_{\text{hor}}(t)$ at time t:

$$d_{\text{hor}}(t) = a(t)c\int_0^t \frac{dt}{a(t)}. \tag{7.12}$$

Thus, for all universes described by the Robertson-Walker metric, the current horizon distance is

$$d_{\text{hor}}(t_0) = c\int_0^{t_0}\frac{dt}{a(t)}. \tag{7.13}$$

An empty universe has an infinite horizon distance according to (7.6). A flat, single-component has a horizon distance of

$$d_{\text{hor}}(t_0) = \frac{3ct_0(1+w)}{1+3w} = \frac{2c}{H_0(1+3w)}, \tag{7.14}$$

provided that $w > -1/3$. For $w \leq -1/3$, the horizon distance is still infinite.

The above analyses are for all single-component universes. We will look at some specific examples below, including matter-only, radiation-only, and Lambda-only universe.

7.3.1 Matter-Only Universe

Consider a universe containing only non-relativistic matter, $w = 0$. By (7.9), (7.14), and (7.8), we obtain its age, horizon distance, and the scale factor as a function of time:

$$t_0 = \frac{2}{3H_0}, \quad d_{\text{hor}}(t_0) = 3ct_0 = \frac{2c}{H_0}, \quad a(t) = \left(\frac{t}{t_0}\right)^{2/3}.$$

If one observe a galaxy with redshift z, the proper distance at the time of observation is given by (7.11) and at the time of emission by dividing $(1+z)$:

$$d_p(t_0) = 3ct_0 \left[1 - \left(\frac{t_e}{t_0}\right)^{1/3}\right] = \frac{2c}{H_0}\left[1 - \frac{1}{\sqrt{1+z}}\right] \quad \Longrightarrow \quad d_p(t_e) = \frac{2c}{H_0(1+z)}\left[1 - \frac{1}{\sqrt{1+z}}\right].$$

In this universe, $d_p(t_e)$ has a maximum at $z = 5/4$, where $d_p(t_e) = (8/27)c/H_0$.

7.3.2 Radiation-Only Universe

In flat, radiation-only universe, its age, horizon distance, and the scale factor is

$$t_0 = \frac{1}{2H_0}, \quad d_{\text{hor}}(t_0) = 2ct_0 = \frac{c}{H_0}, \quad a(t) = \left(\frac{t}{t_0}\right)^{1/2}.$$

Note that the horizon distance in such a universe is exactly equal to the Hubble distance. The proper distance as a function of redshift z is

$$d_p(t_0) = 2ct_0 \left[1 - \left(\frac{t_e}{t_0}\right)^{1/2}\right] = \frac{c}{H_0}\frac{z}{1+z} \quad \Longrightarrow \quad d_p(t_e) = \frac{c}{H_0}\frac{z}{(1+z)^2}.$$

In this universe, $d_p(t_e)$ has a maximum at $z = 1$, where $d_p(t_e) = 0.25c/H_0$. A flat, radiation-only universe is particularly useful, because the early stage of the universe before radiation-matter equality is radiation dominated. It can model to a time as early as about the Planck time $t_{\text{P}} \sim 10^{-43}$ s. The universe at an age comparable or smaller than the Planck time is extremely dense such that we must take quantum effects into account. This requires a theory of quantum gravity, which does not exist yet.

7.3.3 Lambda-Only Universe

For a flat, Lambda-only universe, above analyses breaks down because the cosmological constant Λ has $w = -1$. Luckily, the Friedmann equation is even easier to solve. The Friedmann equation takes the form

$$\dot{a}^2 = \frac{8\pi G\epsilon_\Lambda}{3c^2}a^2, \tag{7.15}$$

where ϵ_Λ is constant with time. Alternatively, it can be written as

$$\dot{a} = H_0 a \quad \text{where} \quad H_0 = \left(\frac{8\pi G\epsilon_\Lambda}{3c^2}\right)^{1/2}.$$

The solution is an exponential,

$$a(t) = e^{H_0(t-t_0)}. \tag{7.16}$$

A universe with only a cosmological constant is infinitely old, since there is no time at which $a(t) = 0$. Hence it has an infinite horizon distance d_{hor}. The proper distance as a function of redshift is

$$d_p(t_0) = c\int_{t_e}^{t_0} e^{H_0(t_0-t)}\,dt = \frac{c}{H_0}\left[e^{H_0(t_0-t_e)} - 1\right] = \frac{c}{H_0}z \quad \Longrightarrow \quad d_p(t_e) = \frac{c}{H_0}\frac{z}{1+z}. \tag{7.17}$$

The final equality of $d_p(t_0)$ comes from the cosmological redshift relation (6.9),

$$1 + z = \frac{1}{a(t_e)} = \frac{1}{e^{H_0(t_e - t_0)}} = e^{H_0(t_0 - t_e)} \quad \Longrightarrow \quad z = e^{H_0(t_0 - t_e)} - 1.$$

An exponentially growing universe is the only universe with $d_p(t_0)$ linearly proportional to z. Other universes only has this relation at $z \ll 1$. There is no maximum $d_p(t_e)$. At $z \gg 1$, the proper distance of the light source at the time of emission goes to c/H_0. This means if the light source is more than a Hubble distance from us, $d_p(t_e) > c/H_0$, we can no longer observe that light source. Its recession velocity is greater than the speed of light.

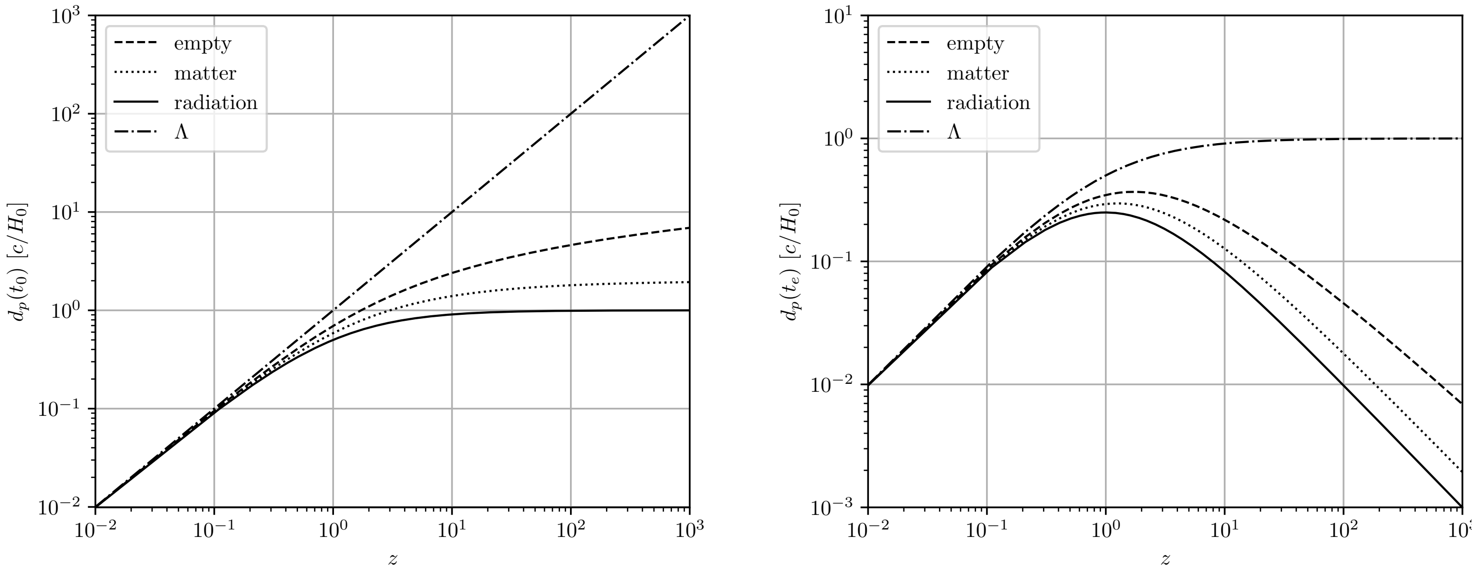


Figure 7.1: Proper distance $d_p(t_0)$ (left) and $d_p(t_e)$ (right) of empty (curvature-only) universe, matter-only universe, radiation only universe, and Lambda-only universe.

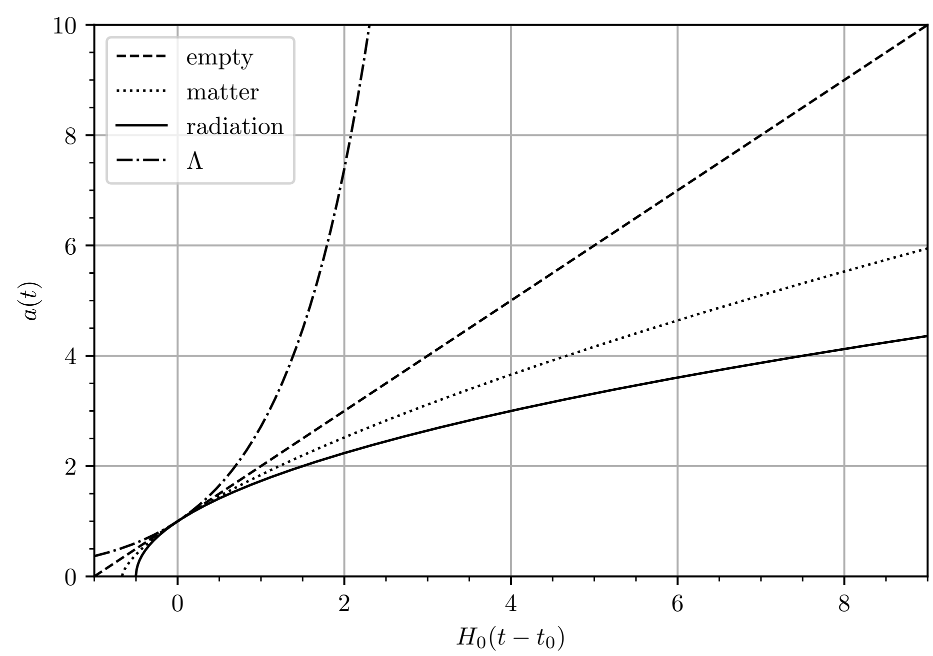


Figure 7.2: Scale factor $a(t)$ of empty (curvature-only) universe, matter-only universe, radiation only universe, and Lambda-only universe.

7.4 Multiple-Component Universes

Recall that the Friedmann equation can be written in terms of the density parameter, (6.16):

$$1 - \Omega(t) = -\frac{\kappa c^2}{R_0^2 a^2 H^2}.$$

We express the curvature as all quantities measured today, in terms of R_0, H_0, Ω_0, and $a(t_0) = 1$,

$$\frac{\kappa}{R_0^2} = \frac{H_0^2}{c^2}(\Omega_0 - 1).$$

Then the Friedmann equation (6.13) will have the form

$$H(t)^2 = \frac{8\pi G\epsilon}{3c^2} - \frac{H_0^2}{a(t)^2}(\Omega_0 - 1) \quad \Longrightarrow \quad \frac{H(t)^2}{H_0^2} = \frac{\epsilon(t)}{\epsilon_{c,0}} + \frac{1-\Omega_0}{a(t)^2},$$

where $\epsilon_{c,0} \equiv 3c^2 H_0^2/8\pi G$ is the critical density. From (7.2), we know the evolution of matter energy density $\epsilon_m = \epsilon_{m,0} a^{-3}$ and radiation energy density $\epsilon_r = \epsilon_{r,0} a^{-4}$. The energy density of the cosmological constant is $\epsilon_\Lambda = \epsilon_{\Lambda,0} = \text{const}$. These energy densities contribute to the total energy density ϵ. Writing all of them in terms of the density parameter, the Friedmann equation takes the form

$$\frac{H^2}{H_0^2} = \frac{\Omega_{r,0}}{a^4} + \frac{\Omega_{m,0}}{a^3} + \Omega_{\Lambda,0} + \frac{1-\Omega_0}{a^2}. \tag{7.18}$$

Note that in this equation, the curvature term (the last term) can also be interpreted as a component of the universe. The sum of all density parameters is $\Omega_{r,0} + \Omega_{m,0} + \Omega_{\Lambda,0} = \Omega_0$. Since the Hubble parameter is $H = \dot{a}/a$, multiplying (7.18) by a^2 and taking the square root on both sides gives

$$H_0^{-1}\dot{a} = \left[\frac{\Omega_{r,0}}{a^2} + \frac{\Omega_{m,0}}{a} + \Omega_{\Lambda,0} a^2 + (1-\Omega_0)\right]^{1/2}.$$

Hence we get the integral to calculate the cosmic time,

$$H_0 t = \int_0^a \frac{da}{[\Omega_{r,0}/a^2 + \Omega_{m,0}/a + \Omega_{\Lambda,0} a^2 + (1-\Omega_0)]^{1/2}}. \tag{7.19}$$

If the upper limit of the integral is $a = 1$, then this integral is the age of the universe. In general, this integral does not have an analytic solution. In the Benchmark model where $\Omega_0 = 1$ (a flat universe), it can be divided into several era where a simple, analytic solution is a good approximation. For example, if $a \ll a_{rm} \approx 2.9 \times 10^{-4}$, the radiation-matter equality, then the Benchmark model can be approximated as a flat, radiation-only universe; if $a \gg a_{m\Lambda} \approx 0.77$, the matter-$\Lambda$ equality, the Benchmark model can be approximated as a flat, lambda-only universe. For scale factors close to those equalities, we must use two-component model of the universe: e.g. radiation + matter universe or matter + Lambda universe.

7.4.1 Matter + Curvature

The matter + curvature model of the universe is historically significant in the 20th century. The Friedmann equation is

$$\frac{H^2}{H_0^2} = \frac{\Omega_0}{a^3} + \frac{1-\Omega_0}{a^2}. \tag{7.20}$$

Since there is only one component except the curvature, $\Omega_{m,0} = \Omega_0$. We already know that for a flat, matter-only universe ($\Omega_0 = 1$), the scale factor is given by $a(t) = (t/t_0)^{2/3}$. The universe expands outward forever. Its fate is usually referred to as a *Big Chill*. If the curvature is nonzero, then the fate of the universe will be quite different.

There can be a universe that have $H(t) = 0$ (or $\dot{a} = 0$), which cease to expand at some point. The first term on the RHS of (7.20) is always positive, so $H(t) = 0$ at some time requires $\Omega_0 > 1$ (or $\kappa = +1$) so that the second term is negative. At the maximum expansion,

$$0 = \frac{\Omega_0}{a_{\max}^3} + \frac{1-\Omega_0}{a_{\max}^2} \quad \Longrightarrow \quad a_{\max} = \frac{\Omega_0}{\Omega_0 - 1}.$$

After reaching $a_{\max}$, the universe will start to contract. Because the LHS of (7.20) depends only on H^2 or $\dot{a}^2$, the contraction a time reversal of the expansion. In fact, an exact time reversal will happen if the expansion of the universe is perfectly homogeneous and adiabatic so that entropy does not build up. Otherwise, it will violate the second law of thermodynamics. The ultimate fate of such a universe is a *Big Crunch*. If $\Omega_0 < 1$, the universe is a negatively curved with only matter. Then both terms on the RHS of (7.20) are positive. If this universe is expanding at some time $t = t_0$, then it will expand forever. In the end, when the universe is large enough ($a \to \infty$) that the density of matter is diluted to nearly an empty space, the scale factor will grow like $a \propto t$.

The Friedmann equation (7.20) can be solved exactly. They are usually written in a parametric form. For $\Omega_0 > 1$, the solution is

$$a(\theta) = \frac{\Omega_0}{2(\Omega_0 - 1)}(1 - \cos\theta), \quad t(\theta) = \frac{\Omega_0}{2H_0(\Omega_0 - 1)^{3/2}}(\theta - \sin\theta), \tag{7.21}$$

where $\theta \in [0, 2\pi]$ from the Big bang at $\theta = 0$ and the Big Crunch at $\theta = 2\pi$. The time of the Big Crunch is at

$$t_{\text{crunch}} = \frac{\pi\Omega_0}{H_0(\Omega_0 - 1)^{3/2}}.$$

For $\Omega_0 < 1$, the solution is

$$a(\eta) = \frac{\Omega_0}{2(1 - \Omega_0)}(\cosh\eta - 1), \quad t(\eta) = \frac{\Omega_0}{2H_0(1 - \Omega_0)^{3/2}}(\sinh\eta - \eta), \tag{7.22}$$

where $\eta \in [0, \infty)$. The scale factor $a(t)$ for for matter + curvature universes is plotted in Figure 7.3.

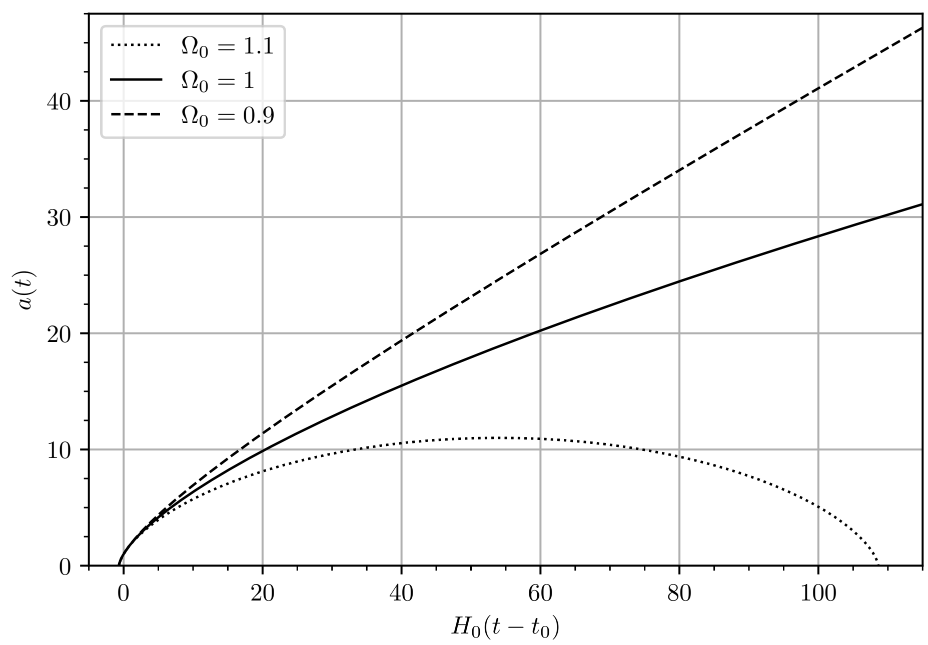


Figure 7.3: The scale factor as a function of time $a(t)$ for matter + curvature universe, plotted with three current density parameters: $\Omega_0 = 0.9$, $\Omega_0 = 1$, and $\Omega_0 = 1.1$. Note the Big Crunch in the case $\Omega_0 = 1.1$.

7.4.2 Matter + Lambda

Our universe today is mostly matter + lambda and nearly spatially flat. It is useful to discuss a flat universe with both matter and a cosmological constant. For a flat universe, $\Omega_{m,0} + \Omega_{\Lambda,0} = 1$, or $\Omega_{\Lambda,0} = 1 - \Omega_{m,0}$. The Friedmann equation is

$$\frac{H^2}{H_0^2} = \frac{\Omega_{m,0}}{a^3} + (1 - \Omega_{m,0}). \tag{7.23}$$

The density parameter of matter is always positive. The second term on the RHS depends on the value of $\Omega_{m,0}$. If $\Omega_{m,0} < 1$, then both terms on the RHS are positive. The universe will expand forever if it is expanding at $t = t_0$, resulting in a Big Chill. The matter-Lambda equality occurs at

$$a_{m\Lambda} = \left(\frac{\Omega_{m,0}}{\Omega_{\Lambda,0}}\right)^{1/3} = \left(\frac{\Omega_{m,0}}{1 - \Omega_{m,0}}\right)^{1/3}.$$

The analytic solution of (7.23) can be obtained by direct integration of (7.19),

$$H_0 t = \frac{2}{3\sqrt{1-\Omega_{m,0}}} \ln\left[\left(\frac{a}{a_{m\Lambda}}\right)^{3/2} + \sqrt{1+\left(\frac{a}{a_{m\Lambda}}\right)^3}\right]. \tag{7.24}$$

When $a \ll a_{m\Lambda}$ and $a \gg a_{m\Lambda}$, we have

$$a(t) \approx \begin{cases} \left(\frac{3}{2}\sqrt{\Omega_{m,0}}H_0 t\right)^{2/3} \propto t^{2/3}, & a \ll a_{m\Lambda}, \\ a_{m\Lambda} e^{\sqrt{1-\Omega_{m,0}}H_0 t} \propto e^{Kt}, & a \gg a_{m\Lambda}, \end{cases}$$

for some constant K. The age of the universe at time t_0 is give by (7.24),

$$t_0 = \frac{2}{3H_0\sqrt{1-\Omega_{m,0}}} \ln\left(\frac{\sqrt{1-\Omega_{m,0}}+1}{\sqrt{\Omega_{m,0}}}\right).$$

Our universe has $\Omega_{m,0} = 0.31$ and $\Omega_{\Lambda,0} = 0.69$. The age is approximately $t_0 = 0.955 H_0^{-1} = (13.74 \pm 0.40)\,\text{Gyr}$. The matter-Lambda equality will occur at $t_{m\Lambda} = 2\ln\big(1+\sqrt{2}\big)/3H_0\sqrt{1-\Omega_{m,0}} = 0.707 H_0^{-1} = (10.17 \pm 0.30)\,\text{Gyr}$. This is about 3.6 billion years ago.

If $\Omega_{m,0} > 1$, at a certain time the RHS of (7.23) will be zero, meaning $H = 0$ at a maximum scale factor,

$$a_{\max} = \left(\frac{\Omega_{m,0}}{\Omega_{m,0}-1}\right)^{1/3}.$$

The cosmological constant acts like an attraction force in this case. The solution to (7.23) is

$$H_0 t = \frac{2}{3\sqrt{\Omega_{m,0}-1}} \arcsin\left[\left(\frac{a}{a_{\max}}\right)^{3/2}\right]. \tag{7.25}$$

There will be a Big Crunch at $a = 0$, at which $\arcsin^{-1}(a/a_{\max})^{3/2} = \pi$, or

$$t_{\text{crunch}} = \frac{2\pi}{3H_0\sqrt{\Omega_{m,0}-1}}.$$

The scale factor $a(t)$ for for matter + Lambda universes is plotted in Figure 7.4.

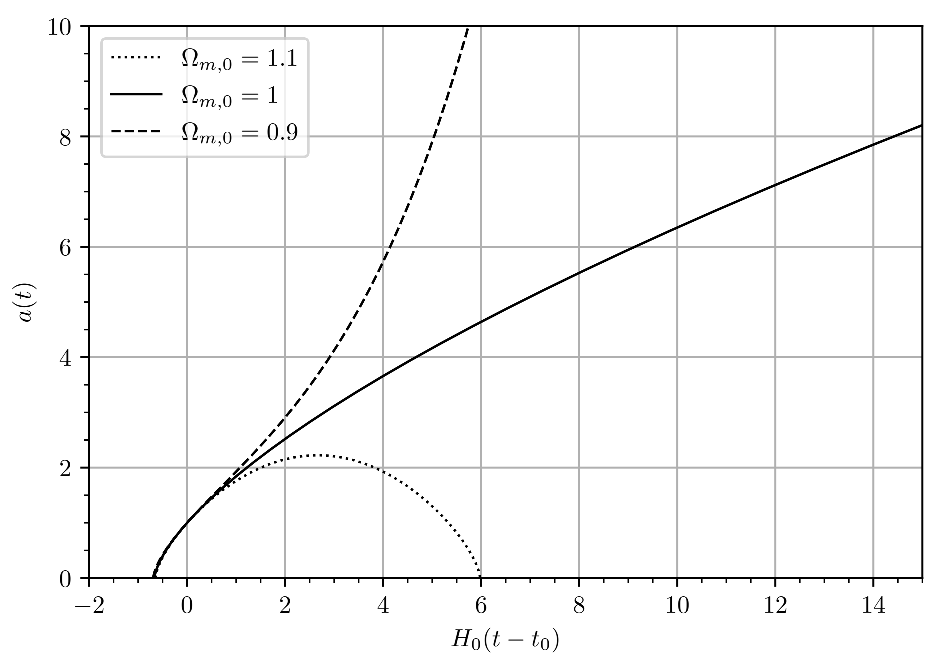


Figure 7.4: The scale factor as a function of time $a(t)$ for flat, matter + Lambda universe, plotted with three matter density parameters: $\Omega_{m,0} = 0.9$, $\Omega_{m,0} = 1$, and $\Omega_{m,0} = 1.1$. Note the Big Crunch in the case $\Omega_{m,0} = 1.1$ and $\Omega_{\Lambda,0} = -0.1$.

7.4.3 Matter + Curvature + Lambda

Interesting behaviors emerge when we choose different $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ while allowing the universe to be curved. Now $\Omega_0 = \Omega_{m,0}+\Omega_{\Lambda,0}$. The Friedmann equation for a matter + curvature + Lambda universe is

$$\frac{H^2}{H_0^2} = \frac{\Omega_{m,0}}{a^3} + \frac{1-\Omega_{m,0}-\Omega_{\Lambda,0}}{a^2} + \Omega_{\Lambda,0}. \tag{7.26}$$

A matter + curvature + Lambda universe also admits a Big Crunch. Since there is no constraints on curvature, we just need $\Omega_{\Lambda,0} < 0$. Then consider the case $\Omega_{m,0} > 0$ and $\Omega_{\Lambda,0} > 0$. If $\Omega_{m,0} + \Omega_{\Lambda,0} < 1$, a negatively curved universe, then it admits a Big Chill.

If $\Omega_{m,0} + \Omega_{\Lambda,0} > 1$, then the universe is positively curved. As usual, matter will dominate at small a, and Λ will dominate at large a. However, there will be some intermediate a where the curvature term (the second term on the RHS) dominates, resulting in a negative H^2. Such a universe should start with $a \gg 1$ and $H < 0$. It contracts until it reaches a minimum value of the scale factor $a = a_{\min}$ and expand (sometimes called a *Big Bounce*). This universe will not have an initial Big Bang.

Another possible universe is a *loitering* universe (or **Lamaître's universe**). If $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ are at right values, the universe will start as matter dominated, but then enters a stage in which $a \approx$ const. for a very long time. The cosmological constant will then dominate and the universe expand exponentially.

Figure 7.5 shows the scale factor and fate of different universes.

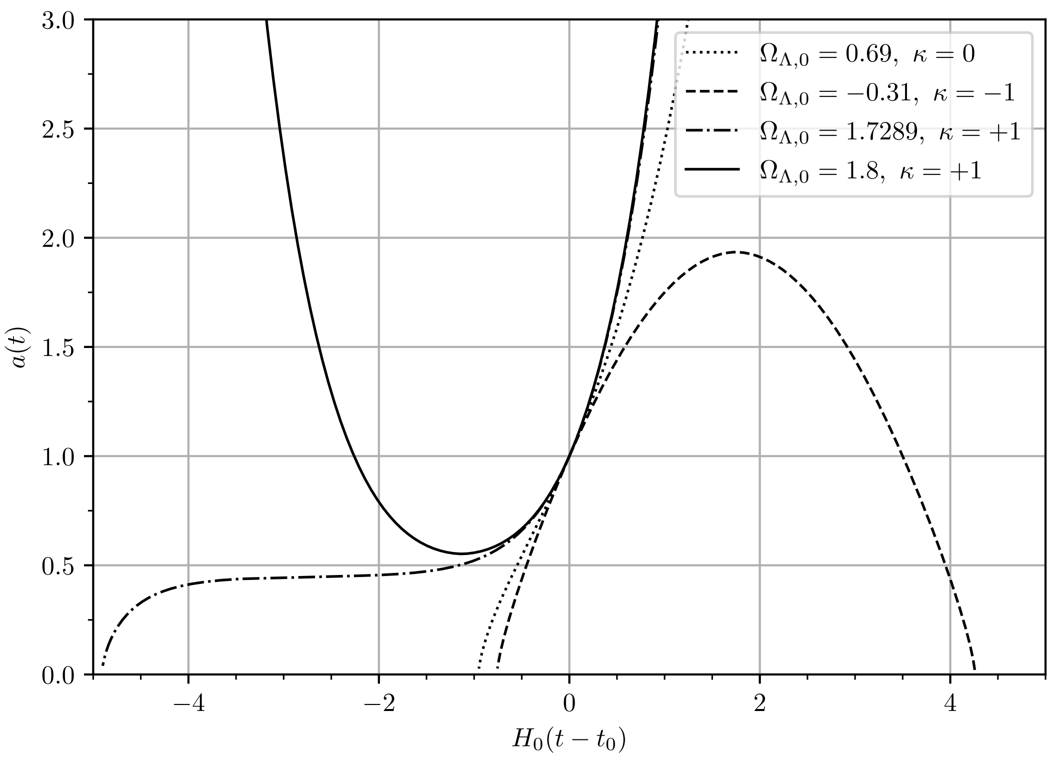


Figure 7.5: The scale factor as a function of time $a(t)$ in four universes with different fates: Big Chill (dotted line), Big Crunch (dashed line), loitering/Lamaître (dashdot line), and Big Bounce (solid line).

7.4.4 Radiation + Matter

The flat, radiation + matter model is useful when $a \sim a_{rm} = \Lambda_{r,0}/\Lambda_{m,0}$, the radiation-matter equality. The Friedmann equation is

$$\frac{H^2}{H_0^2} = \frac{\Omega_{r,0}}{a^4} + \frac{\Omega_{m,0}}{a^3}. \tag{7.27}$$

Multiplying both sides by $H_0^2 a^4$,

$$a^2\dot{a}^2 = H_0^2(\Omega_{r,0} + a\Omega_{m,0}) = H_0^2\left(\Omega_{r,0} + \frac{a\Omega_{r,0}}{a_{rm}}\right)$$

Taking the square root and rearrange some terms yields

$$H_0 dt = \frac{a\,da}{\sqrt{\Omega_{r,0}}}\left(1 + \frac{a}{a_{rm}}\right)^{-1/2}.$$

Integrate both sides,

$$H_0 t = \frac{4a_{rm}^2}{3\sqrt{\Omega_{r,0}}}\left[1 - \left(1 - \frac{a}{2a_{rm}}\right)\left(1 + \frac{a}{a_{rm}}\right)^{1/2}\right]. \tag{7.28}$$

In the limits $a \ll a_{rm}$ or $a \gg a_{rm}$, we have

$$a \approx \begin{cases} (2\sqrt{\Omega_{r,0}}H_0 t)^{1/2} \propto t^{1/2}, & a \ll a_{rm}, \\ \left(\frac{3}{2}\sqrt{\Omega_{m,0}}H_0 t\right)^{2/3} \propto t^{2/3}, & a \gg a_{rm}. \end{cases}$$

The time of radiation-matter equality is at $a = a_{rm}$,

$$t_{rm} = \frac{4}{3}\left(1 - \frac{1}{\sqrt{2}}\right)\frac{a_{rm}^2}{\sqrt{\Omega_{r,0}}}H_0^{-1} \approx 0.391\frac{\Omega_{r,0}^{3/2}}{\Omega_{m,0}^2}H_0^{-1} \approx 50000\,\mathrm{yr}$$

if we use $\Omega_{r,0} = 9.0 \times 10^{-5}$, $\Omega_{m,0} = 0.31$ and $H_0^{-1} = 14.4\,\mathrm{Gyr}$ from the Benchmark Model.

7.5 Benchmark Model (ΛCDM)

The **Lambda cold dark matter model** (ΛCDM) is the Benchmark model we are using today. It is a flat universe with radiation, matter, and dark energy. Table 7.1 lists the components of this model, summarizing previous sections in this chapter.

Components	Density parameter	
Total radiation	$\Omega_{r,0} = 9.0 \times 10^{-5}$	
Photons	$\Omega_{\gamma,0} = 5.35 \times 10^{-5}$	
Neutrinos	$\Omega_{\nu,0} = 3.65 \times 10^{-5}$	
Total matter	$\Omega_{m,0} = 0.31$	
Baryonic matter	$\Omega_{\mathrm{bary},0} = 0.048$	
Nonbaryonic dark matter	$\Omega_{\mathrm{dm},0} = 0.262$	
Cosmological Constant	$\Omega_{\Lambda,0} = 0.69$	
Important epochs		
Radiation-matter equality	$a_{rm} = 2.9 \times 10^{-4}$	$t_{rm} = 0.050\,\mathrm{Myr}$
Matter-Lambda equality	$a_{m\Lambda} = 0.77$	$t_{m\Lambda} = 10.2\,\mathrm{Gyr}$
Now	$a_0 = 1$	$t_0 = 13.7\,\mathrm{Gyr}$

Table 7.1: Properties of the Benchmark model (ΛCDM).

The scale factor $a(t)$ can be computed numerically via the Friedmann equation (7.18). Figure 7.6 shows the evolution of the scale factor.

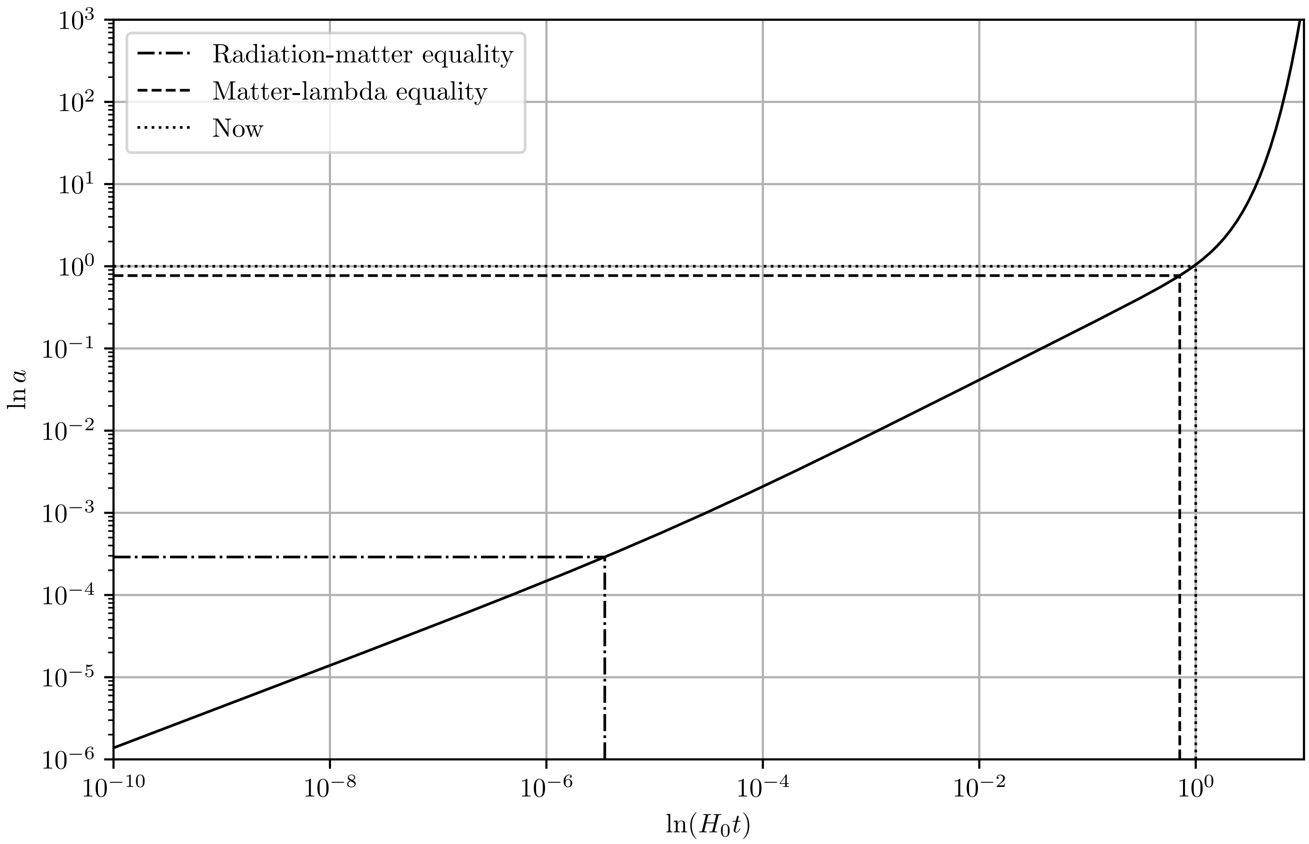


Figure 7.6: The scale factor as a function of t in units of the Hubble time.

The horizon distance of the Benchmark Model is (see Figure 7.7)

$$d_{\rm hor}(t_0) = 3.20c/H_0 = 3.35ct_0 = 14000\,\text{Mpc}.$$

The maximum $d_p(t_e)$, proper distance of the light source at the time of emission, has $z = 1.6$ and $d_p(t_e) = 0.405c/H_0$.

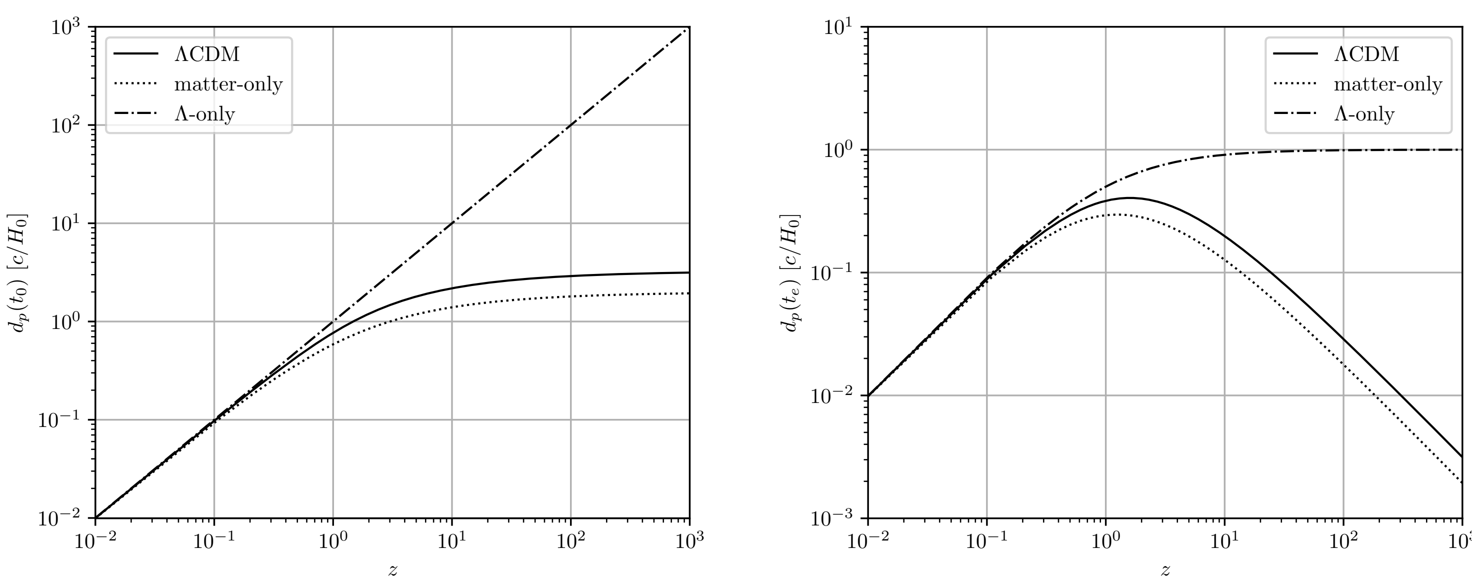


Figure 7.7: The proper distance to a light source with redshift z, in units of the Hubble distance c/H_0.

8 Cosmological Parameters and Observations

The Friedmann equation works in both ways: one can deduce the scale factor $a(t)$ from energy components of the universe, or deduce energy components from the scale factor. In reality, finding $a(t)$ is hard because we cannot observe the whole universe. In this chapter, we will discuss how to put constraints on $a(t)$ from observations of distant objects.

8.1 A Search for Two Numbers

If we cannot know the exact scale factor, we can make approximations by Taylor series around $t = t_0$,

$$a(t) = a(t_0) + \left.\frac{da}{dt}\right|_{t=t_0}(t-t_0) + \frac{1}{2}\left.\frac{d^2a}{dt^2}\right|_{t=t_0}(t-t_0)^2 + \cdots .$$

From the previous chapter, the scale factors of modeled universes do not vary rapidly with t, so the first few terms of the expansion are enough. Keeping terms up to second order and divide both sides by $a(t_0)$,

$$\frac{a(t)}{a(t_0)} \simeq 1 + \left.\frac{\dot{a}}{a}\right|_{t=t_0}(t-t_0) + \frac{1}{2}\left.\frac{\ddot{a}}{a}\right|_{t=t_0}(t-t_0)^2.$$

Invoking the normalization $a(t_0) = 1$, we can write

$$a(t) \simeq 1 + H_0(t-t_0) + \frac{1}{2}\left.\frac{\ddot{a}}{aH^2}H^2\right|_{t=t_0}(t-t_0) = 1 + H_0(t-t_0) - \frac{1}{2}q_0H_0^2(t-t_0)^2, \tag{8.1}$$

where H_0 is the Hubble constant and

$$q_0 \equiv -\left(\frac{\ddot{a}a}{\dot{a}^2}\right)_{t=t_0} = -\left(\frac{\ddot{a}}{aH^2}\right)_{t=t_0} \tag{8.2}$$

is known as the **deceleration parameter**. A positive value of q_0 corresponds to a decelerating universe, $\ddot{a} < 0$. We can use the acceleration equation (6.20) to predict q_0 for a given model of the universe. If the model universe contains N components, the acceleration says

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}\sum_{i=1}^{N}\epsilon_i(1+3w_i).$$

Dividing both sides by H^2 gives

$$-\frac{\ddot{a}}{aH^2} = \frac{1}{2}\left(\frac{8\pi G}{3c^2H^2}\right)\sum_i \epsilon_i(1+3w_i) = \frac{1}{2}\frac{1}{\epsilon_c}\sum_i \epsilon_i(1+3w_i) = \frac{1}{2}\sum_i \Omega_i(1+3w_i),$$

where we use the definition of the critical density $\epsilon_c = 3c^2H^2/8\pi G$ in (6.14). At the present $t = t_0$, the deceleration parameter is then

$$q_0 = \frac{1}{2}\sum_i \Omega_{i,0}(1+3w_i). \tag{8.3}$$

For a universe with radiation ($w = 1/3$), matter ($w = 0$), and Λ ($w = -1$),

$$q_0 = \Omega_{r,0} + \frac{1}{2}\Omega_{m,0} - \Omega_{\Lambda,0}.$$

It will be accelerating outward if $\Omega_{\Lambda,0} > \Omega_{r,0} + \Omega_{m,0}/2$. The Benchmark Model ($\Lambda$CDM) has $q_0 \approx -0.53$.

Now with the scale factor in approximated form with known H_0 and q_0, we are able to calculate other quantities. Recall that the current proper distance is given by

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

Since we now approximately know $a(t)$ in the expansion (8.1), its inverse can be calculated using Taylor expansion again,

$$\frac{1}{a(t)} \simeq 1 - H_0(t-t_0) + \left(1+\frac{q_0}{2}\right) H_0^2(t-t_0)^2. \tag{8.4}$$

Integrating $a(t)$ over time gives the proper distance

$$d_p(t_0) \simeq c(t_0-t_e) + \frac{cH_0}{2}(t_0-t_e)^2,$$

keeping terms up to second order of the lookback time $t_0 - t_e$. The first term represents the proper distance in a static universe, while the second term is a correction to the proper distance due to the expansion of the universe. It is useful to express the proper distance in terms of redshift. Using (6.9), $z+1 = 1/a(t_e)$, (8.4) can be written in terms of the lookback time

$$z \simeq H_0(t_0-t_e) + \left(1+\frac{q_0}{2}\right) H_0^2(t_0-t_e)^2.$$

This is a quadratic equation of $(t_0-t_e)^2$,

$$t_0 - t_e = \frac{-H_0 \pm \sqrt{H_0^2 + 4(1+q_0/2)H_0^2 z}}{2(1+q_0/2)H_0^2}.$$

The lookback time cannot be negative, so keep the positive solution. In the limit of small z,

$$t_0 - t_e = \frac{1}{H_0}\frac{-1+\sqrt{1+4(1+q_0/2)z}}{2(1+q_0/2)} \simeq \frac{1}{H_0}\frac{-1+1+2(1+q_0/2)z - 2(1+q_0/2)z^2}{2(1+q_0/2)^2} = \frac{1}{H_0}z\left[1-\left(1+\frac{q_0}{2}\right)z\right].$$

Substituting this lookback time into $d_p(t_0)$ and keeping to second order of z,

$$d_p(t_0) \simeq \frac{c}{H_0}\left[z - \left(1+\frac{q_0}{2}\right)z^2\right] + \frac{cH_0}{2}\frac{z^2}{H_0^2} = \frac{c}{H_0}z\left[1-\frac{1+q_0}{2}z\right]. \tag{8.5}$$

For $z \ll 2/(1+q_0)$, (8.5) reproduces the linear Hubble's law, $d_p = cz/H_0 = H_0 v$.

8.2 Measuring Distances

The current proper distance $d_p(t_0)$ is not a directly measurable quantity, because it requires one observer on Earth and another comoving observer to the galaxy to measure the distance at the same time. In this section, we will discuss how to measure other kinds of distances that can infer the current proper distance.

8.2.1 Luminosity Distance

One property of celestial object is that it emits light. We can measure the flux of light F (in W/m^2) from the galaxy. If by some means we can also know the luminosity L of the object, then the classical luminosity distance to that object is defined as

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

This is the proper distance to the object if the universe were static and Euclidean. It results from the inverse square law of light intensity, $F = L/4\pi d^2$. The objects that have relatively fixed intrinsic luminosity are called **standard candles**. However, this does not work for a universe described by a Robertson Walker metric (6.5),

$$ds^2 = -c^2dt^2 + a(t)^2[dr^2 + S_\kappa(r)^2 d\Omega^2] \quad \text{where} \quad S_\kappa(r) = \begin{cases} R_0\sin(r/R_0), & \kappa=+1, \\ r, & \kappa = 0, \\ R_0\sinh(r/R_0), & \kappa=-1. \end{cases}$$

We must account for the expansion of the universe and the curvature. Suppose we are at the origin, and we see photons from an object at comoving coordinate (r,θ,ϕ) that were emitted at time t_e. All photons emitted at t_e are now (at t_0) spread over a sphere of proper radius $d_p(t_0) = r$. The surface area of this sphere is the proper surface area $A_p(t_0)$, given by

$$A_p(t_0) = 4\pi S_\kappa(r).$$

That is, if space has no curvature, $A_p(t_0) = 4\pi r^2$. If space is positively curved, $A_p(t_0) < 4\pi r^2$. If it is negatively curved, $A_p(t_0) > 4\pi r^2$. These are the geometric effect.

The second effect is the expansion of the universe. Let the wavelength of light at emission be λ_e, and the scale factor at that time be $a(t_e)$. When it reaches us, its wavelength is stretched by

$$\lambda_0 \frac{a(t_0)}{a(t_e)} \lambda_e = (1+z)\lambda_e \quad \implies \quad E_0 = \frac{E_e}{1+z}.$$

Meanwhile, the flux is a power (energy per unit time) measurement. Because of the expansion of the universe, the time interval between photon reception will be larger. Consider two photons emitted in the same direction but separated by a time interval δt_e. Their proper distance at the time of emission will be $c\,\delta t_e$. When we detect the photons at t_0, the proper distance between them will be stretched to $c(1+z)\,\delta t_e$, resulting in a detection time interval $\delta t_0 = (1+z)\,\delta t_e$. Thus, the energy received per unit time will have two factors of $1/(1+z)$,

$$\frac{dE_0}{dt_0} = \frac{1}{1+z}\frac{dE_e}{dt_0} = \frac{1}{(1+z)^2}\frac{dE_e}{dt_e} = \frac{L}{(1+z)^2},$$

where L is the intrinsic luminosity of the object. In conclusion, taking into account both geometric and expansion effects, the relation between the observed flux F and the luminosity L of the object is

$$F = \frac{L}{4\pi S_\kappa(r)^2(1+z)^2}. \tag{8.6}$$

The **luminosity distance** is defined as

$$d_L \equiv S_\kappa(r)(1+z). \tag{8.7}$$

Since our universe is nearly flat, the radius of curvature is much larger than the current horizon distance $d_{\text{hor}}(t_0)$. We can approximate $S_\kappa(r) \approx r$. The luminosity distance is then

$$d_L = r(1+z) = d_p(t_0)(1+z) \qquad\qquad [\kappa = 0]. \tag{8.8}$$

Combining this with the current proper distance in terms of the Hubble constant and deceleration parameter using (8.5),

$$d_L \simeq \frac{c}{H_0} z\left(1 - \frac{1+q_0}{2}z\right)(1+z) \simeq \frac{c}{H_0} z\left(1 + \frac{1-q_0}{2}z\right).$$

8.2.2 Angular-Diameter Distance

If we do not know the luminosity of the object in advance, it is not possible to calculate the luminosity distance directly. Instead, if we know the proper length ℓ of the object that does not change with the expansion of the universe (such as a gravitationally bound object), then we can calculate the angular-diameter distance

$$d_A \equiv \frac{\ell}{\delta\theta},$$

where $\delta\theta$ is the measured angular distance of the object. This definition using the small-angle formula is valid when $\delta\theta \ll 1$, which is almost always true for a celestial object. Objects with fixed proper length are called **standard rulers**. Just like the classical luminosity distance, this angular-diameter distance is equal to the proper distance if the universe is static and Euclidean. Then it is of course not equal to the proper distance in general if the universe is expanding or curved.

Choose a coordinate system such that we are at the origin. Let the comoving coordinates of the two ends of the distant object at t_e be (r, θ_1, ϕ) and (r, θ_2, ϕ). When the lights at both ends travel towards us, they travel along geodesics with constant θ (one with θ_1 and the other with θ_2 to be precise) and ϕ. Thus, we will measure $\delta\theta = \theta_2 - \theta_1$. From the Robertson-Walker metric

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

The object at constant r and ϕ at t_e will span a distance

$$ds = a(t_e)S_\kappa(r)\,\delta\theta.$$

Setting this equal to the known proper length ℓ, we find that

$$\ell = a(t_e)S_\kappa(r)\,\delta\theta = \frac{S_\kappa(r)\,\delta\theta}{1+z}.$$

The **angular-diameter distance** is thus

$$d_A \equiv \frac{\ell}{\delta\theta} = \frac{S_\kappa(r)}{1+z}. \tag{8.9}$$

We can relate the angular-diameter distance with the luminosity distance using (8.7),

$$d_A = \frac{d_L}{(1+z)^2}. \tag{8.10}$$

If the universe is spatially flat, then

$$(1+z)d_A = d_p(t_0) = \frac{d_L}{1+z} \qquad [\kappa = 0]. \tag{8.11}$$

Note that the angular diameter distance in a flat universe is equal to the proper distance at the time of light emission: $d_A = d_p(t_0)/(1+z) = d_p(t_e)$. Relating the angular diameter distance to H_0 and q_0 using (8.5),

$$d_A \simeq \frac{c}{H_0}z\left(1 - \frac{3+q_0}{2}z\right).$$

In the small z limit, all the distances are more or less equal, $d_A \approx d_L \approx d_p(t_0) \approx (c/H_0)z$. However, in the limit $z \to \infty$,

$$d_p(t_0) \to d_{\text{hor}}(t_0), \quad d_L \to z d_{\text{hor}}(t_0) \to \infty, \quad d_A \to \frac{d_{\text{hor}}(t_0)}{z} \to 0.$$

Historically, the luminosity distance is much more useful than the angular-diameter distance. There are two reasons: 1. a standard ruler must have an angular size large enough that is measurable. This is not always achievable. In contrast, galaxies and galaxy clusters have large enough luminosity that can be used to obtain d_L. 2. It is difficult to assign an angular size $\delta\theta$ to galaxies/clusters that do not have well-defined edges. Moreover, many clusters are not isolated as they merge around.

8.3 Dark Matter

Matter is the dominant component before matter-Lambda equality at $a = 0.77$. Though matter is no longer the dominant component in the ΛCDM model, it still contributes $\Omega_{m,0} = 0.31$, so it is not negligible. This section discusses searches of matter density in the universe, from baryonic matter to dark matter.

8.3.1 Baryonic Matter

It is difficult to observe matter density directly through observation. A common way to infer matter density is through luminosity density in the local universe. The current luminosity density in the V band ($500\,\text{nm} < \lambda < 590\,\text{nm}$) from galaxies in the local universe is

$$\Psi_V = 1.1 \times 10^8 L_{\odot,V}/\text{Mpc}^3.$$

It tells us the mass density $\rho_\star$ of stars if we know the **mass-to-light ratio** of stars. A very crude estimate is to assume that all stars are identical to the Sun. Then $\langle M/L_V\rangle = 1\,M_\odot/L_{\odot,V}$, where $L_{\odot,V} \approx 0.12\,L_\odot = 4.6 \times 10^{25}\,\text{W}$. However, there are many stars different from the Sun. For example, an O-type star with $M = 60\,M_\odot$ has $L_V \approx 20000\,L_{\odot,V}$, giving a mass-to-light ratio of $M/L_V \approx 0.003\,M_\odot/L_{\odot,V}$. An M-type star with mass $M = 0.1\,M_\odot$ has $L_V \approx 5 \times 10^{-5}\,L_{\odot,V}$, or a mass-to-light ratio of $M/L_V \approx 2000\,M_\odot/L_{\odot,V}$. Therefore, we also need to know the distribution of stars masses. In a star-forming region, the number of stars created with masses in the range $M \to M + dM$ is given by the **initial mass function** $\chi(M)\,dM$,

$$\chi(M) \propto \begin{cases} M^{-\beta}, & M > 1\,M_\odot, \\ \dfrac{1}{M}e^{-(\log M - \log M_c)^2/2\sigma^2}, & M < 1\,M_\odot. \end{cases} \tag{8.12}$$

The index β, characteristic mass M_c, and width σ vary with locations. Typical values are $\beta = 2.3$, $M_c \approx 0.2\,M_\odot$, and $\sigma \approx 0.5$. In galaxies that ae still actively forming stars today, the mass-to-light ratio is about $M/L_V \approx 0.3\,M_\odot/L_{\odot,V}$. In galaxies that do not form new stars long ago, many luminous but short-lived O-stars run out of fuel. The mass-to-light ratio will rise to $M/L_V \approx 8\,M_\odot/L_{\odot,V}$. The overall averaged mass-to-light ratio is $\langle M/L_V \rangle \approx 4\,M_\odot/L_{\odot,V}$. Hence the mass density of stars in the universe today is

$$\rho_{\star,0} = \langle M/L_V \rangle \Psi_V \approx 4 \times 10^8\,M_\odot/\mathrm{Mpc}^3.$$

Compared to the critical density of the universe today, in the form of a mass density $\rho_{c,0} = 1.28 \times 10^{11}\,M_\odot/\mathrm{Mpc}^3$, the stars contribute only

$$\Omega_{\star,0} = \frac{\rho_{\star,0}}{\rho_{c,0}} \approx \frac{4 \times 10^8\,M_\odot/\mathrm{Mpc}^3}{1.28 \times 10^{11}\,M_\odot/\mathrm{Mpc}^3} \approx 0.003.$$

Much of the mass density are in fact not in stars. It is contributed mostly by gases between stars, between galaxies, and between galaxy clusters, by about 85%. These gases are not easy to detect, so the data comes from observations of the cosmic microwave background and theoretical predictions of nucleosynthesis in the early universe. The final result of the density parameter of baryonic matter today is

$$\Omega_{b,0} = 0.048 \pm 0.003.$$

8.3.2 Dark Matter in Galaxies

It turns out that most of the matter is not even baryonic. The non-baryonic matter are known as **dark matter**. Dark matter does not have electromagnetic interaction with any baryonic matter, but it interacts with baryonic matter gravitationally. A typical method to detect dark matter is to observe the orbital speeds of stars in spiral galaxies. Most stars in a spiral galaxy have circular orbits around the center of the galaxy. Let the orbital radius be R and orbital speed be v. A star will experience a centripetal acceleration of

$$a = \frac{v^2}{R} = \frac{GM(R)}{R^2} \quad \Longrightarrow \quad v = \sqrt{\frac{GM(R)}{R}},$$

where $M(R)$ is the mass of the disk within radius R. The surface brightness of the disk of a spiral galaxy usually falls of exponentially with distance from the center,

$$B = B_0 e^{-R/R_s},$$

where R_s is the characteristic length scale. The Milky way galaxy has $R_s \approx 4\,\mathrm{kpc}$, with surface brightness measured in the V band, while M31 (the Andromeda Galaxy) has larger, $R_s \approx 6\,\mathrm{kpc}$. This exponential law tells us that the mass of the disk inside R becomes approximately constant if R is several times R_s. If stars contribute most of the mass in a galaxy, the velocity of a circularly orbiting object outside a few R_s will fall off as $v \propto 1/\sqrt{R}$. This relation is often referred to as "Keplerian" orbit.

In 1970, Vera Rubin and Kent Ford measured emission lines from hot ionized gas located at $R = 4R_s$ in M31 and deduced the orbital speed. No Keplerian relation was found. At $R > 4R_s$, emission line at $\lambda = 21\,\mathrm{cm}$ from hydrogen in the disk of M31 were observed. The Doppler shift in this line shows that $v(R) \approx 230\,\mathrm{km/s}$, roughly a constant from $4R_s$ to $6R_s$, but not Keplerian relation $v \propto 1/\sqrt{R}$. This indicates that there must be some unseen **dark halo** hiding in the stellar disk, providing additional gravitational attraction. It turns out that almost all spiral galaxies have dark matter halos.

If we assume that $v = \text{const.}$ with radius, the mass of a spiral galaxy within a disk of radius R will be

$$M(R) = \frac{v^2 R}{G} = 1.05 \times 10^{11}\,M_\odot \left(\frac{v_{\mathrm{orbit}}}{235\,\mathrm{km/s}}\right)^2 \left(\frac{R_{\mathrm{orbit}}}{8.2\,\mathrm{kpc}}\right),$$

where we are referencing to the Sun's location and speed in our galaxy, $v = 235\,\mathrm{km/s}$ and $R = 8.2\,\mathrm{kpc}$. The luminosity in the V band of our galaxy is $L_{\mathrm{gal},V} = 2.0 \times 10^{10}\,L_{\odot,V}$, so the mass-to-light ratio is

$$\langle M/L_V \rangle_{\mathrm{gal}} \approx 64\,M_\odot/L_{\odot,V} \left(\frac{R_{\mathrm{halo}}}{100\,\mathrm{kpc}}\right),$$

where we used $v_{\mathrm{orbit}} = 235\,\mathrm{km/s}$, and R_{halo} is the radius of the dark matter halo. A rough estimate from the globular clusters and satellite galaxies orbiting our galaxy gives $R_{\mathrm{halo}} \approx 75\,\mathrm{kpc}$. The precise value of R_{halo} is still unknown. Some astronomers believe that it is 4 times larger, $R_{\mathrm{halo}} \approx 300\,\mathrm{kpc}$, reaching nearly halfway to our neighbor galaxy M31. Substituting $R_{\mathrm{halo}} \approx 75\,\mathrm{kpc}$ into $\langle M/L_V \rangle_{\mathrm{gal}}$ yields $\langle M/L_V \rangle_{\mathrm{gal}} \approx 48\,M_\odot/L_{\odot,V}$ and $M_{\mathrm{gal}} \approx 9.6 \times 10^{11}\,M_\odot$. This mass-to-light ratio is 10 times larger than that of the stars in our galaxy, which means the hark matter halo is much more massive than the stellar disk.

8.3.3 Dark Matter in Clusters

In the 1930s, when Fritz Zwicky was studying the Coma cluster of galaxies, he discovered that the dispersion in the radial velocity of the cluster's galaxies was around 1000 km/s. This velocity is too high to be bounded by gravitational attraction from only visible stars and gas within the galaxies. Zwicky believed that the cluster must contain a large amount of dark matter.

We shall use the virial theorem to follow Zwicky's reasoning. Consider a cluster consisting of N galaxies, each of which is approximated as a point mass with label $i = 1, 2, \ldots, N$. The ith galaxy has a mass m_i, a position $\mathbf{x}_i$, and a velocity $\dot{\mathbf{x}}_i$. The description of the motion of galaxies in a cluster require only Newtonian mechanics. The acceleration of the ith galaxy is given by the superposition of all gravitational fields generated by other galaxies,

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

The acceleration due to matter outside the cluster is neglected. The total gravitational potential energy of the system is

$$E_g = -\frac{G}{2} \sum_{j \neq i} \sum_i \frac{m_i m_j}{|\mathbf{x}_j - \mathbf{x}_i|}.$$

The factor of 1/2 is to prevent overcounting of galaxies: the potential energy in A and B is the same as that in B and A. The total kinetic energy of galaxies in the cluster is

$$E_k = \frac{1}{2} \sum_i m_i |\dot{\mathbf{x}}_i|^2.$$

Define the moment of inertia to be

$$I \equiv \sum_i m_i |\mathbf{x}_i|^2.$$

Assume that the cluster is in a steady state, with $\dot{I} = 0$ and $\ddot{I} = 0$. The second-derivative of the moment of inertia is

$$\ddot{I} = 2 \sum_i m_i (\mathbf{x}_i \cdot \ddot{\mathbf{x}}_i + \dot{\mathbf{x}}_i \cdot \dot{\mathbf{x}}_i) = 2 \sum_i m_i (\mathbf{x}_i \cdot \ddot{\mathbf{x}}_i) + 4E_k.$$

The first sum can be expressed using the acceleration of the ith galaxy,

$$\sum_i m_i (\mathbf{x}_i \cdot \ddot{\mathbf{x}}_i) = G \sum_i \sum_{j \neq i} m_i m_j \frac{\mathbf{x}_i \cdot (\mathbf{x}_j - \mathbf{x}_i)}{|\mathbf{x}_j - \mathbf{x}_i|^3}.$$

But it is the same sum if we exchange the dummy indices i and j,

$$\sum_j m_j (\mathbf{x}_j \cdot \ddot{\mathbf{x}}_j) = G \sum_j \sum_{i \neq j} m_j m_i \frac{\mathbf{x}_j \cdot (\mathbf{x}_i - \mathbf{x}_j)}{|\mathbf{x}_i - \mathbf{x}_j|^3}.$$

Hence

$$\begin{aligned}
\sum_i m_i (\mathbf{x}_i \cdot \ddot{\mathbf{x}}_i) &= \frac{1}{2} \left[\sum_i m_i (\mathbf{x}_i \cdot \ddot{\mathbf{x}}_i) + \sum_j m_j (\mathbf{x}_j \cdot \ddot{\mathbf{x}}_j) \right] \\
&= \frac{G}{2} \left[\sum_i \sum_{j \neq i} m_i m_j \frac{\mathbf{x}_i \cdot (\mathbf{x}_j - \mathbf{x}_i)}{|\mathbf{x}_j - \mathbf{x}_i|^3} + \sum_j \sum_{i \neq j} m_j m_i \frac{\mathbf{x}_j \cdot (\mathbf{x}_i - \mathbf{x}_j)}{|\mathbf{x}_i - \mathbf{x}_j|^3} \right] \\
&= \frac{G}{2} \left[\sum_i \sum_{j \neq i} m_i m_j \frac{\mathbf{x}_i \cdot (\mathbf{x}_j - \mathbf{x}_i)}{|\mathbf{x}_j - \mathbf{x}_i|^3} - \sum_i \sum_{j \neq i} m_i m_j \frac{\mathbf{x}_j \cdot (\mathbf{x}_j - \mathbf{x}_i)}{|\mathbf{x}_j - \mathbf{x}_i|^3} \right] \\
&= -\frac{G}{2} \sum_i \sum_{j \neq i} \frac{m_i m_j}{|\mathbf{x}_j - \mathbf{x}_i|} = E_g.
\end{aligned}$$

This means in steady state, $0 = \ddot{I} = 2E_g + 4E_k$. This is the familiar virial theorem,

$$E_g = -2E_k, \quad E_{\text{tot}} = -E_k = \frac{1}{2} E_g.$$

Alternatively, the potential energy can be written as

$$E_g = -\alpha \frac{GM^2}{r_h},$$

where $M = \sum_i m_i$ is the total mass of galaxies in the cluster, $\alpha(\approx 0.45$ is a good fit) is a numerical factor of order unity, and r_h is the **half-mass radius** of the cluster (the radius of a sphere centered at the center of the cluster and containing a mass $M/2$.) The kinetic energy can be written as

$$E_k = \frac{1}{2} M \langle v^2 \rangle \qquad \text{where} \qquad \langle v^2 \rangle \equiv \frac{1}{M} \sum_i m_i |\dot{\mathbf{x}}_i|^2.$$

Then the virial theorem says that

$$\boxed{M = \frac{\langle v^2 \rangle r_h}{\alpha G}.} \tag{8.13}$$

The application of the virial theorem to galaxy clusters is hard. In general, we do not know $\langle v^2 \rangle$ and r_h. We can only find the line-of-sight velocity of each galaxy from its redshift, while the velocity perpendicular to the line of sight is unknown. Determining r_h requires even more assumptions.

Example 8.1. Mass of the Coma cluster

Consider the Coma cluster as an example. The mean redshift is $\langle z \rangle = 0.0232$. This is not a high redshift, so we can use Hubble's law directly, $d_{\text{Coma}} = (c/H_0)\langle z \rangle = 102\,\text{Mpc}$. The velocity dispersion along the line of sight is observed to be

$$\sigma_r = \langle (v_r - \langle v_r \rangle)^2 \rangle^{1/2} = 880\,\text{km/s}.$$

This is one-dimensional. Assuming that the velocity dispersion is isotropic, then the three-dimensional mean square velocity will be

$$\langle v^2 \rangle = 3\sigma_r^2 = 2.32 \times 10^{12}\,\text{m}^2/\text{s}^2.$$

Estimating r_h needs more assumptions. First, assume that the mass-to-light ratio is constant with radius, so r_h also contains half the luminosity of the cluster. Second, assume that the cluster is spherically symmetric. The observed distribution of galaxies in the Coma cluster gives a half-mass radius of $r_h \approx 1.5\,\text{Mpc} \approx 4.6 \times 10^{22}\,\text{m}$. Then we can estimate the mass of the Coma cluster using the virial theorem (8.13),

$$M_{\text{Coma}} = \frac{\langle v^2 \rangle r_h}{\alpha G} \approx 4 \times 10^{45}\,\text{kg} \approx 2 \times 10^{15}\,M_\odot.$$

Observations show that only one percent of the mass of the Coma cluster is from stars, and ten percent from hot intracluster gas. The luminosity of the Coma cluster in the V band is $L_{\text{Coma},V} \approx 5 \times 10^{12}\,L_{\odot,V}$, which gives a mass-to-light ratio of

$$\langle M/L_V \rangle_{\text{Coma}} \sim 400\,M_\odot/L_{\odot,V}.$$

8.3.4 Gravitational Lensing

So far the detection of matter require only non-relativistic work. Another method to detect matter is using **gravitational lensing**, a purely general relativistic effect. General relativity predicts that the trajectory of photons will be bent by massive objects. According to general relativity, if a photon passes by a massive object of mass M at an impact parameter b, then the local curvature of spacetime will cause the photon to be deflected by an angle

$$\delta\phi = \frac{4GM}{c^2 b}. \tag{8.14}$$

For example, if a photon just graze through a massive object of radius R, (so $b = R$), here are some angles of gravitational lensing from common objects:

Object	Radius (m)	Mass (kg)	Angle of deflection
The Earth	6.4×10^6	6.0×10^{24}	$5.7'' \times 10^{-4}$
The Sun	7.0×10^8	2.0×10^{30}	$1.7''$
A white dwarf	1.5×10^7	2.0×10^{30}	$1.4'$
A neutron star	1.2×10^4	3.0×10^{30}	$43°$

Gravitational lensing was confirmed in 1919 after Einstein's prediction using general relativity. A solar eclipse allowed observation of a star field near the Sun. Then a photograph of this star field was compared to the same star field takne 6 months earlier. The stars were deflected by the amount predicted by Einstein, caused by gravitational lensing from the Sun.

Gravitational lensing is a method to detect MACHOs, or **MAssive Compact Halo Objects**. The MACHOs can be cold white dwarfs, black holes, brown dwarfs that are too dim to be detected, but they can act as gravitational lenses. They exist in dark halos within our galaxies. Suppose there is a MACHO of mass M between our galaxy and a star in the Large Megellanic Cloud (LMC). We will then observe the star to be distorted and amplified. If the MACHO is exactly along the line between the observer and the lensed star, then the image of the star is a perfect *ring* with angular radius

$$\theta_{\mathrm{E}} = \sqrt{\frac{4GM}{c^2}\frac{D_{LS}}{D_S D_L}}, \tag{8.15}$$

where D_S is the distance from the observer to the lensed star, D_L is the distance from the observer to the lensing MACHO, and $D_{LS} = D_S - D_L$ (see Figure 8.1) This angular radius is known as the **Einstein radius**.

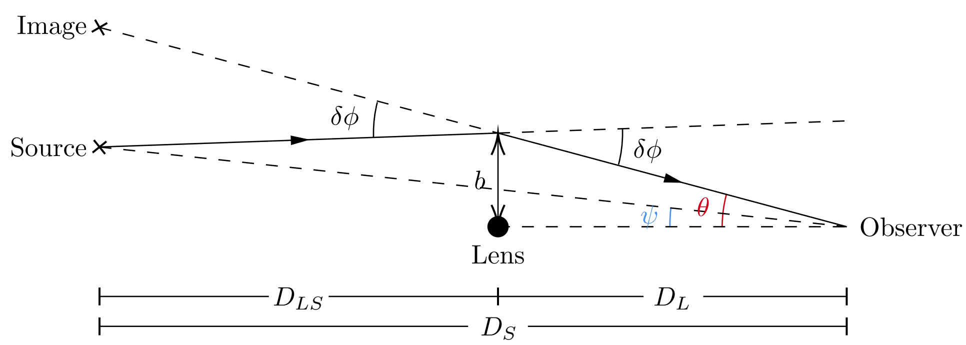


Figure 8.1: An illustration of gravitational lensing. The trajectory of photon can be approximated as straight lines because the size of the lens is small compared to D_S, just like in geometric optics. Note that θ is not necessarily equal to $\delta\phi$.

In a strong lensing situation, the lensed star may appear in multiple locations due to the large angle of deflection. If $x = 0.5$ (the MACHO is halfway between us and the star), then

$$\theta_{\mathrm{E}} \approx 4'' \times 10^{-4} \left(\frac{M}{M_\odot}\right)^{1/2} \left(\frac{D_S}{50\,\mathrm{kpc}}\right)^{-1/2}.$$

If the MACHO is not exactly along the line between the observer and the lensed star, then image of the star will be separated into several arcs. The probability of detecting a MACHO lensing a star in the LMC is low. The MACHOs and the stars in the LMC are in relative motion. A typical signature of lensing event is a star that becomes brighter/angular distance decreases, and then becomes dimmer/angular distance increases. The typical timescale for a lensing event is about

$$\Delta t = \frac{2\theta_{\mathrm{E}} D_L}{v}.$$

where v is the relative transverse velocity of the MACHO relative to background stars as seen on Earth. The search of MACHOs concluded that at most 8 percent of the halo mass is in the form of MACHOs. The rest should be in the form of dark matter.

Gravitational lensing can also occur in large scale structures. For example, a galaxy cluster with $M \sim 10^{14}\,M_\odot$ at a distance

$\sim 500\,\mathrm{Mpc}$ from us is able to lens a background galaxy at $d \sim 1000\,\mathrm{Mpc}$. The Einstein radius is

$$\theta_{\mathrm{E}} \approx 0.5' \left(\frac{M}{10^{14}\,M_{\odot}}\right)^{1/2} \left(\frac{d}{1000\,\mathrm{Mpc}}\right)^{-1/2} .$$

9 The Early Universe

In this chapter, we will visit all the way to the first second of the universe. using the cosmic microwave background and nucleosynthesis, the products of photon decoupling and neutrino decoupling. Then we introduce some flaws of the standard Hot Big Bang model in the theory of cosmic inflation.

*Note: this chapter involves multiple Saha-like equations and concepts of radiation. It is useful to consult Section, 2.3.4, 2.3.5, in Stellar Structure and Evolution and Appendix A.7 .

9.1 Cosmic Microwave Background

The cosmic microwave background (CMB) is a uniform photon gas and a blackbody at $T_0 = 2.7255\,\mathrm{K}$ with nearly perfect fit. The corresponding energy density (2.12) and number density (2.10) of CMB photons are

$$\epsilon_{\gamma,0} = a_{\mathrm{rad}} T_0^4 = 0.2606\,\mathrm{MeV/m^3}, \quad n_{\gamma,0} = bT_0^3 = 4.107 \times 10^8\,\mathrm{m^{-3}}.$$

The energy density of baryonic matter is $\epsilon_{\mathrm{bary},0} = \Omega_{\mathrm{bary},0}\epsilon_{c,0} \approx 234\,\mathrm{MeV/m^3}$, about 900 times that of the CMB photons. Most of the energy of a baryon (like a proton or neutron) is the rest mass energy $E_{\mathrm{bary}} \approx 939\,\mathrm{MeV}$. Thus, the number density of baryons is

$$n_{\mathrm{bary},0} = \frac{\epsilon_{\mathrm{bary},0}}{E_{\mathrm{bary}}} \approx \frac{234\,\mathrm{MeV/m^3}}{939\,\mathrm{MeV}} \approx 0.25\,\mathrm{m^{-3}},$$

much lower than that of photons. The ratio of baryons to photons in the universe, η, is

$$\eta = \frac{n_{\mathrm{bary},0}}{n_{\gamma,0}} \approx \frac{0.25\,\mathrm{m^{-3}}}{4.107 \times 10^8\,\mathrm{m^{-3}}} \approx 6.1 \times 10^{-10}, \tag{9.1}$$

or 1.6 billion photons per baryon.

9.1.1 Recombination and Decoupling

To trace back to te origin of the CMB, we first define three moments in the history of the universe.

1. The epoch of **recombination** is the time at which the baryonic component of the universe transition from an ionized plasma to a gas of neutral atoms. We define it to be the instant when the number density of ions is equal to that of neutral atoms.
2. The epoch of **photon decoupling** is the time when the rate at which photons scatter from electrons becomes smaller than the Hubble parameter.
3. The epoch of **last scattering** is the time at which a typical CMB photon underwent its last scattering from an electron.

For simplicity, assume that the only baryonic component in the universe are hydrogen during the recombination epoch. The hydrogen can be in two forms: a neutral hydrogen atom (H), or a nucleus known as a proton (p). If the whole universe is neutral, then the number density of free electrons should be the same as that of free protons, $n_e = n_p$. The ionization fraction x, is then

$$x \equiv \frac{n_p}{n_p + n_{\mathrm{H}}} = \frac{n_p}{n_{\mathrm{bary}}} = \frac{n_e}{n_{\mathrm{bary}}}.$$

The dominant reaction that occured during the recombination epoch is the photoionization of a hydrogen and its reverse process:

$$\mathrm{H} + \gamma \longleftrightarrow p + e^-.$$

Choose a time period before the recombination, for example, at $a = 10^{-5}$ or $z = 10^5$. In the ΛCDM model, the universe was about 70 years old. By (5.3), the background radiation had a temperature of $T \approx 3 \times 10^5\,\mathrm{K}$. The mean photon energy is $\langle \epsilon_\gamma \rangle = 2.7kT \approx 60\,\mathrm{eV}$, in the ultraviolet range. Recall that there are 1.6 billion photons per baryon, so any hydrogen atom formed will be ionized by dense and energetic photons. Thus, the ionization fraction at that time is $x \approx 1$.

When the universe is fully ionized, photons scattered off electrons mainly through Thomson scattering, with a Thomson scattering cross-section of $\sigma_{\mathrm{T}} = 6.65 \times 10^{-29}\,\mathrm{m}^2$. The mean free path of the photon is (2.16),

$$l = \frac{1}{\sigma_{\mathrm{T}} n_e},$$

and the rate of Thomson scattering is $\Gamma = c/l = \sigma_{\mathrm{T}} n_e c$. For a fully ionized universe, $n_e = n_p = n_{\mathrm{bary}}$. Since baryon number is conserved, the number of density of baryons, and hence the number density of electrons, goes like

$$n_e = n_{\mathrm{bary}} = \frac{n_{\mathrm{bary},0}}{a^3} \quad \Longrightarrow \quad \Gamma = \frac{\sigma_{\mathrm{T}} n_{\mathrm{bary},0} c}{a^3} = \frac{5.0 \times 10^{-21}\,\mathrm{s}^{-1}}{a^3}.$$

At $a = 10^{-5}$, the scattering rate is about $\Gamma = 5.0 \times 10^{-6}\,\mathrm{s}^{-1}$, or three times a week. When $a < a_{rm} \approx 2.9 \times 10^{-4}$, the universe was in radiation era. The Friedmann equation is

$$\frac{H^2}{H_0^2} = \frac{\Omega_{r,0}}{a^4} \quad \Longrightarrow \quad H = \frac{H_0 \Omega_{r,0}^{1/2}}{a^2} = \frac{2.1 \times 10^{-20}\,\mathrm{s}^{-1}}{a^2}.$$

For example, at $a = 10^{-5}$, the Hubble parameter is about $H \approx 2.1 \times 10^{-10}\,\mathrm{s}^{-1} < \Gamma$. As long as $\Gamma > H$, the Hubble parameter, or equivalently, their mean free path $l < c/H$, we say that photons remain coupled to electrons. In this case, the photons, electrons, and protons are in thermal equilibrium because they are constantly interacting with each other. Once photons decouple with electrons, the universe becomes transparent, and baryonic matter will have a different temperature with the CMB.

If hydrogen was always fully ionized, we may set $\Gamma = H$ and calculate that the photon decoupling should occur at $a \approx 0.0254$, or $z \approx 38$ and at a CMB temperature of $T \approx 110\,\mathrm{K}$. (Note at this scale factor, one should take both radiation and matter into account and compute the Hubble parameter.) However, at such a low temperature, the mean energy of photons is $\langle \epsilon_\gamma \rangle = 2.7kT \approx 0.026\,\mathrm{eV} \ll 13.6\,\mathrm{eV}$. In other words, the photons could not keep the hydrogen atoms ionized, in contradiction to our assumption. Hence we need to be more careful about ionization. The best way to study ionization is to invoke the Saha equation.

Recall that the ground state Saha equation for the reaction $H + \gamma \longleftrightarrow p + e^-$ is given by (2.13):

$$\frac{n_p n_e}{n_{\mathrm{H}}} = \left(\frac{2\pi m kT}{h^2}\right)^{3/2} e^{-\chi/kT} \quad \text{where} \quad \chi = 13.6\,\mathrm{eV}.$$

It is derived from the equilibrium condition $\mu_{\mathrm{H}} + \mu_\gamma = \mu_p + \mu_e$, where the chemical potential and quantum concentration are

$$\mu = mc^2 + kT \ln\left(\frac{n}{g n_Q}\right), \quad n_Q = \left(\frac{2\pi m kT}{h}\right)^{3/2}.$$

Here g is the degeneracy states of a particle. In the ground state, $g_p = g_e = 2$ and so $g_{\mathrm{H}} = 4$. We also make the approximation that $m_p = m_{\mathrm{H}}$. Relating n_{H} with n_p from ionization fraction,

$$x \equiv \frac{n_p}{n_p + n_{\mathrm{H}}} \quad \Longrightarrow \quad n_{\mathrm{H}} = \frac{1-x}{x} n_p.$$

With $n_e = n_p$, the Saha equation becomes

$$\frac{x}{1-x} n_p = \left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} e^{-\chi/kT}.$$

To eliminate n_p, we use $\eta \equiv n_{\mathrm{bary}}/n_\gamma = n_p/x n_\gamma$. Then by (2.10), we write the number density of photons in terms of the temperature, $n_\gamma = bT^3$,

$$n_p = \eta x b T^3 = 16\pi \eta x \left(\frac{k}{hc}\right)^3 \zeta(3) T^3.$$

Substituting this back into the Saha equation,

$$\frac{\eta x^2 b T^3}{1-x} = \left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} e^{-\chi/kT} \quad \Longrightarrow \quad \frac{\eta x^2}{1-x} = \frac{\sqrt{\pi}}{2^{5/2}\zeta(3)} \left(\frac{m_e c^2}{kT}\right)^{3/2} e^{-\chi/kT}.$$

For calculation convenience, we plug in all the numbers ($\zeta(3) \approx 1.202$),

$$\frac{x^2}{1-x} = \frac{0.26}{\eta} \left(\frac{m_e c^2}{kT}\right)^{3/2} e^{-\chi/kT}. \tag{9.2}$$

It is a quadratic equation, with parameter T and η. The solution is

$$x = \frac{-1 + \sqrt{1 + 4y}}{2y} \quad \text{where} \quad y(T, \eta) = 3.84\eta \left(\frac{kT}{m_e c^2} \right)^{3/2} e^{\chi/kT}. \tag{9.3}$$

If we define the instant of recombination to be $x = 1/2$, then (assuming $\eta = 6.1 \times 10^{-10}$) the recombination temperature is

$$kT_{\text{rec}} = 0.324\,\text{eV} = \frac{\chi}{42}.$$

This corresponds to a temperature of $T_{\text{rec}} = 3760\,\text{K}$ on the Kelvin scale and a redshift $z_{\text{rec}} = 1380$. According to the ΛCDM model, the age of the universe at the time of recombination is $t_{\text{rec}} = 250000\,\text{yr}$. Figure 9.1 shows the ionization fraction as a function of redshift. We see that the recombination epoch happens gradually in the thermodynamic standard so that chemical equilibrium and thermal equilibrium is a good approximation. But in the cosmological standard, the recombination is rapid.

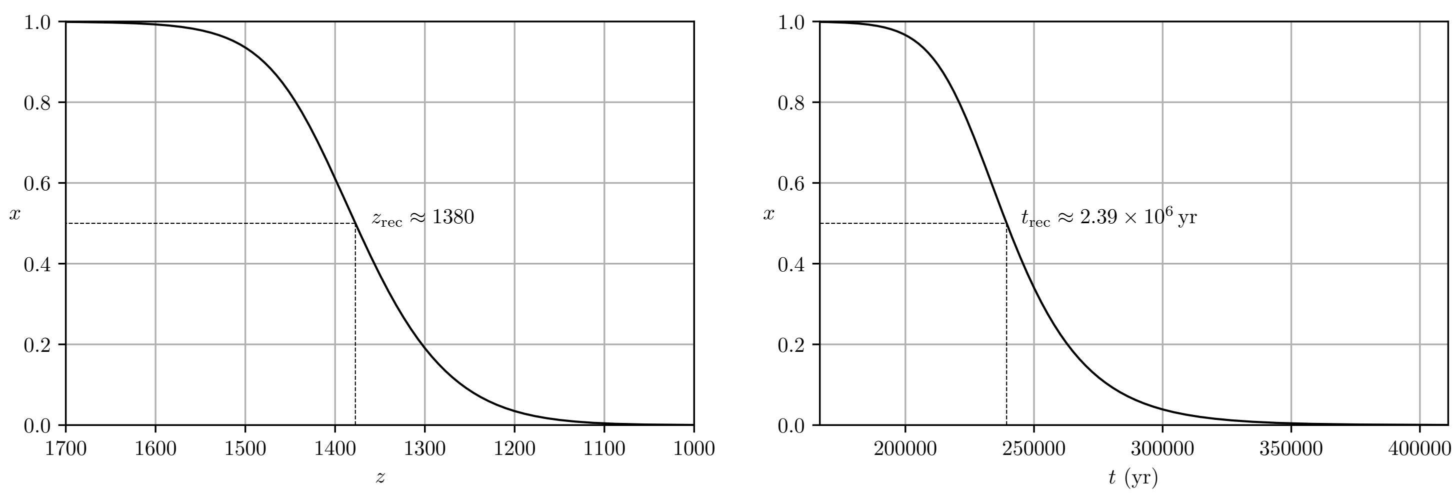


Figure 9.1: Ionization fraction x as a function of redshift z and as a function of time t during the epoch of recombination. A baryon-to-photon ratio $\eta = 6.1 \times 10^{-10}$ is assumed. The redshift or recombination is $z_{\text{rec}} \approx 1380$ and the corresponding age of the universe is $t_{\text{rec}} \approx 239000$ years.

Also, the baryon-to-photon ratio does not affect much on the recombination temperature (see Figure 9.2).

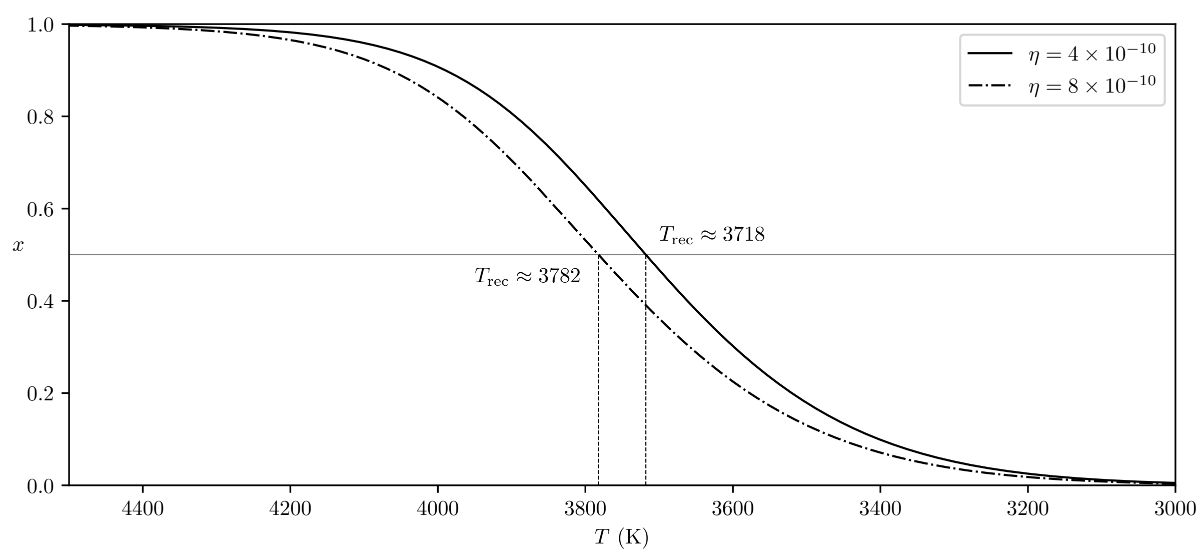


Figure 9.2: The ionization fraction x as a function of temperature T near the recombination epoch. We see that doubling the baryon-to-photon ratio η only increase the recombination by about $64\,\text{K}$.

We can work out the rate of photon scattering when the hydrogen is partially ionized. The scattering rate depends on the number density of free electrons, and hence the redshift,

$$\Gamma(z) = n_e(z)\sigma_{\mathrm{T}} c = x(z)(1+z)^3 n_{\mathrm{bary},0}\sigma_{\mathrm{T}} c \approx (5.0 \times 10^{-21}\,\mathrm{s}^{-1})x(z)(1+z)^3,$$

where we used $n_{\mathrm{bary},0} \approx 0.25\,\mathrm{m}^{-3}$. During the recombination, the universe is already matter dominated. The Friedmann equation says

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

Using $\Omega_{m,0}$ in the ΛCDM model, the Hubble parameter during the recombination is

$$H(z) = (1.23 \times 10^{-18}\,\mathrm{s}^{-1})(1+z)^{3/2}. \tag{9.4}$$

Setting $\Gamma = H$, we find that the redshift of photon decoupling is obeys

$$1 + z_{\mathrm{dec}} = \frac{39.3}{x(z_{\mathrm{dec}})^{2/3}} \qquad \Longrightarrow \qquad z_{\mathrm{dec}} = 1120.$$

This can be calculated numerically, or read from Figure 9.1 roughly. However, the exact redshift of photon decoupling is a bit smaller, $z_{\mathrm{dec}} = 1090$. This is because near $\Gamma \sim H$, the Saha equation is not a good approximation because the photon and the hydrogen are not in true thermal equilibrium. To get z_{dec}, one needs non-equilibrium physics. We will stick with $z_{\mathrm{dec}} = 1090$, with a corresponding temperature $T_{\mathrm{dec}} = 2970\,\mathrm{K}$, and $t_{\mathrm{dec}} = 371000\,\mathrm{yr}$.

Now we want to find the time of last scattering. The probability of a photon undergoes a scattering is $dP = \Gamma(t)\,dt$. For a photon detected at time t_0, its expected number of scattering from time t to t_0 is

$$\tau(t) = \int_t^{t_0} \Gamma(t')\,dt',$$

where τ is the **optical depth**. The time of last scattering is the time at which $\tau = 1$. We may compute τ by a change of variables

$$\tau(a) = \int_a^1 \Gamma(a')\frac{da'}{\dot{a}'} = \int_a^1 \frac{\Gamma(a')}{H(a')}\frac{da'}{a'} \quad \text{or} \quad \tau(z) = \int_0^z \frac{\Gamma(z')}{H(z')}\frac{dz'}{1+z'} = 0.0041\int_0^z x(z')\sqrt{1+z'}\,dz'.$$

Again, the Saha equation does not work very well here because the photons and matter at the time of last scattering is certainly not in equilibrium. We will take the redshift of the time of last scattering to be $z_{\mathrm{ls}} \approx z_{\mathrm{dec}} \approx 1090$.

In conclusion, the CMB photons we observe today have the time of last scattering at around 370000 year after the Hot Big Bang. We are looking far back in time compared to the age of the universe now. Table 9.1 summaries the events discussed in this section, with a reference of radiation-matter inequality.

Event	Redshift	Scale factor	Temperature (K)	Time (Myr)
Radiation-matter equality	3440	2.9×10^{-4}	9390	0.050
Recombination	1380	7.2×10^{-4}	3760	0.25
Photon decoupling and last scattering	1090	9.2×10^{-4}	2970	0.37

Table 9.1: Events in the early universe.

9.1.2 Temperature Fluctuations

The cosmic microwave background is not a perfect homogeneous photon gas. There are temperature fluctuations within. The physical size of the temperature fluctuations can be derived from its angular size. Let the observed angular size be $\delta\theta$. The physical size ℓ is related to $\delta\theta$ by $\ell = d_A\delta\theta$, where d_A is the angular diameter distance to the surface of last scattering. Since $z_{\mathrm{ls}} = 1090 \gg 1$, the angular diameter distance is approximately

$$d_A \approx \frac{d_{\mathrm{hor}}(t_0)}{z_{\mathrm{ls}}} \approx \frac{14000\,\mathrm{Mpc}}{1090} \approx 12.8\,\mathrm{Mpc},$$

where $d_{\rm hor}(t_0)$ is the current horizon distance in the ΛCDM model. The physical size of the temperature fluctuation is then

$$\ell = d_A \delta\theta = 12.8\,{\rm Mpc}\left(\frac{\delta\theta}{1\,{\rm rad}}\right) = 3.7\,{\rm kpc}\left(\frac{\delta\theta}{1'}\right).$$

Because the angular diameter in a flat universe is also $d_p(t_e)$, $\ell \approx 18\,{\rm kpc}$ is the physical size at the time of last scattering. The physical size today is $\ell(1+z_{\rm ls}) \approx 20\,{\rm Mpc}$.

The surface of last scattering is a sphere to us, so the temperature fluctuation $\delta T/T$ of the cosmic microwave background is defined on the surface of the sphere. It is useful to to expand $\delta T/T$ in spherical harmonics $Y_\ell^m(\theta,\phi)$,

$$\frac{\delta T}{T}(\theta,\phi) = \sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\theta,\phi),$$

The statistical property of $\delta T/T$ are encoded in the correlation function $C(\theta)$, defined as

$$C(\theta) = \left\langle \frac{\delta T}{T}(\hat{\mathbf{n}})\frac{\delta T}{T}(\hat{\mathbf{n}}')\right\rangle, \tag{9.5}$$

where $\cos\theta = \hat{\mathbf{n}}\cdot\hat{\mathbf{n}}'$, and $\hat{\mathbf{n}}$ and $\hat{\mathbf{n}}'$ are directions to a point on the sphere for an observer. When $\delta T/T$ is expressed in spherical harmonics, the correlation function can be written as

$$C(\theta) = \frac{1}{4\pi}\sum_{\ell=0}^{\infty}(2\ell+1)C_\ell P_\ell(\cos\theta),$$

where P_ℓ are Legendre polynomials. The coefficients C_ℓ are the multipole moments of the correlation function. As usual, $\ell=0$ is the monopole, $\ell=1$ the dipole, and so on. The $\ell \geq 2$ moments tells the temperature fluctuations at the time of last scattering.

9.1.3 Causes of Temperature Fluctuations

An important distance at the time of last scattering is the horizon distance. According to (7.12),

$$d_{\rm hor}(t_{\rm ls}) = a(t_{\rm ls})c\int_0^{t_{\rm ls}}\frac{dt}{a(t)} \approx 2.24ct_{\rm ls} \approx 0.251\,{\rm Mpc}. \tag{9.6}$$

The value is calculated from the radiation + matter model, as Λ does not have much effect at early times. The angular size of a hot spot with physical size $d_{\rm hor}(t_{\rm ls})$ as observed on Earth will be

$$\theta_{\rm hor} = \frac{d_{\rm hor}(t_{\rm ls})}{d_A} = \frac{0.251\,{\rm Mpc}}{12.8\,{\rm Mpc}} \approx 1.1^\circ.$$

There are both large scale fluctuations $\theta > \theta_{\rm hor}$ and small scale fluctuations $\theta < \theta_{\rm hor}$ in temperature, but their causes are different.

Large Scale Fluctuations. These temperature fluctuations results from density fluctuations and their gravitational effect. We can first calculate the energy density of dark matter, baryonic matter, and photons at $t_{\rm ls}$:

$$\begin{aligned}
\epsilon_{\rm dm}(z_{\rm ls}) &= \Omega_{\rm dm,0}\epsilon_{c,0}(1+z_{\rm ls})^3 \approx 1.7\times10^{12}\,{\rm MeV/m^3},\\
\epsilon_{\rm bary}(z_{\rm ls}) &= \Omega_{\rm bary,0}\epsilon_{c,0}(1+z_{\rm ls})^3 \approx 3.1\times10^{11}\,{\rm MeV/m^3},\\
\epsilon_\gamma(z_{\rm ls}) &= \Omega_{\gamma,0}\epsilon_{c,0}(1+z_{\rm ls})^4 \approx 3.9\times10^{11}\,{\rm MeV/m^3},
\end{aligned}$$

where we used $\epsilon_{\rm dm}, \epsilon_{\rm bary} \propto a^{-3}$ and $\epsilon_\gamma \propto a^{-4}$. The density ratio of the three is $\epsilon_{\rm dm} : \epsilon_\gamma : \epsilon_{\rm bary} = 5.5 : 1.24 : 1$. Thus, dark matter dominated the gravitational potential at $t_{\rm ls}$.

Now consider the energy density of dark matter that is not perfectly homogeneous, but with some perturbation in space,

$$\epsilon(\mathbf{r}) = \bar\epsilon + \delta\epsilon(\mathbf{r}),$$

where $\bar\epsilon$ is the homogeneous piece (the mean energy density), and $\delta\epsilon$ is the local deviation from the mean energy density. In Newtonian mecahnics, the density fluctuations give rise to the potential fluctuation $\delta\Phi$ through Poisson's equation,

$$\nabla^2(\delta\Phi) = \frac{4\pi G}{c^2}\delta\epsilon.$$

The fluctuation in potential will then cause photons to gain or lose energy, depending on whether they are at local maximum or minimum of $\delta\Phi$. The temperature fluctuation resulting from the potential fluctuation needs to be calculated from general relativity. The answer is

$$\boxed{\frac{\delta T}{T} = \frac{1}{3}\frac{\delta\Phi}{c^2}.} \tag{9.7}$$

This type of creation of temperature fluctuations is called the **Sachs-Wolfe effect**.

Small Scale Fluctuations. Small scale fluctuations are more complicated. Immediately before photon decoupling, all the photons and matter is in a photon-baryon fluid. This fluid moves according to the gravitational potential from dark matter. For example, when the photon-baryon fluid fall toward the center of the potential well, both its density and pressure increases. If the fluid is compressed enough, the pressure will cause it to expand and the pressure itself will drop. Then the fluid will fall inward again. This type of oscillation is called **acoustic oscillations**, like standing sound waves.

Now consider the instance of photon decoupling. Roughly speaking, if the photon-baryon fluid is at maximum compression, the temperature of the liberated photon will be higher because $T \propto \epsilon_\gamma^{1/4}$ for photons. Similarly, it will be colder if the fluid is at maximum expansion. If the fluid is in the process of expansion or contraction, the temperature depends on the direction of photons relative to us. The potential wells generated by dark matter have a typical size of the **sound horizon distance**,

$$d_s(t_{\rm ls}) = a(t_{\rm ls})\int_0^{t_{\rm ls}} \frac{c_s(t)\,dt}{a(t)}. \tag{9.8}$$

This formula is similar to proper distance, except the speed of sound $c_s(t)$ is time-dependent. The sound speed of the photon-baryon fluid is approximately the same as that of a pure photon gas, $c_s \approx c/\sqrt{3}$. Hence the sound horizon distance at the last scattering is estimated to be

$$d_s(t_{\rm ts}) \approx \frac{1}{\sqrt{3}} d_{\rm hor}(t_{\rm ls}) \approx 0.145\,\text{Mpc}.$$

The angular size of a hot spots with physical size $d_s(t_{\rm ts})$ is $\theta_s \approx d_s(t_{\rm ls})/d_A \approx 0.7^\circ$.

9.2 Nucleosynthesis

The cosmic microwave background tells us the physics of the universe near the epoch of recombination. The energy scale during this time period is of the order 10 eV. Before the recombination epoch, the temperature of the universe, and hence the energy scale, are even higher. We will encounter the nuclear energy scale in MeV all the way to Planck energy, 10^{19} GeV.

The nuclear counterpart of the epoch of recombination is called the epoch of **Big Bang nucleosynthesis**. During nucleosynthesis, heavy nuclei form from fusion reactions of protons and neutrons. Nucleosynthesis occur before the radiation-matter equality. In radiation dominated era, the temperature scales like $T \propto a^{-1} \propto t^{-1/2}$, just like the cosmic microwave background.

9.2.1 Protons and Neutrons

Before nucleosynthesis, most of the energy density of baryonic matter is in the form of protons and neutrons. The rest mass energy difference between a neutron and a proton is

$$\chi_n = m_n c^2 - m_p c^2 = 1.29\,\text{MeV}.$$

Free neutrons are not stable particles. In fact, it is this energy difference that make them unstable. Neutrons decay within a mean lifetime of about $\tau_n = 880\,$s, or 15 minutes, via beta decay

$$n \longrightarrow p + e^- + \bar{\nu}_e.$$

A free neutron will spontaneously decay into a proton, an electron, and an electron antineutrino. An electron has a rest mass of $0.511\,\text{MeV} < \chi_n$, so the remaining energy are carried away as kinetic energy by daughter particles (neutrinos are often regarded as massless). Note that we emphasize on *free* neutrons. If a neutron binds to a nucleus, then it is in general stable, because it requires an additional amount of energy to overcome the nuclear binding energy to decay. Otherwise, since the decay is an exponential $N(t) = N_0 e^{-t/\tau}$, there will be no neutrons left today.

Consider the state of the universe when its age was $t = 0.1\,\text{s}$, at a temperature of $T \approx 3 \times 10^{10}\,\text{K}$. Such a temperature corresponds to a mean photon energy of $\langle \epsilon_\gamma \rangle = 2.7kT \sim 10\,\text{MeV}$. This energy is much larger than the rest mass of an electron. Thus, the energetic photons were able to produce an electron-positron pair: $\gamma + \gamma \longleftrightarrow e^+ + e^-$. Moreover, at $t = 0.1\,\text{s}$, neutrinos were not decoupled to baryonic matter, so protons and neutrons can convert to each other via

$$n + \nu_e \;\longleftrightarrow\; p + e^- \quad \text{and} \quad n + e^+ \;\longleftrightarrow\; p + \bar{\nu}_e.$$

If all particles are in a thermal equilibrium and a diffusive equilibrium, we can derive a Saha-like equation for proton-neutron equilibrium using the condition $\mu_n = \mu_p$. (We do not include μ_e and μ_ν because they can be created or annihilated easily at $T \approx 3 \times 10^{10}\,\text{K}$. Their chemical potentials are effectively zero, just like photons.) Using

$$\mu = mc^2 + kT \ln\left(\frac{n}{gn_Q}\right) \quad \text{and} \quad n_Q = \left(\frac{2\pi mkT}{h^2}\right)^{3/2},$$

the equilibrium condition reads

$$m_n c^2 + kT \ln\left(\frac{n_n}{g_n n_{Q,n}}\right) = m_p c^2 + kT \ln\left(\frac{n_p}{g_p n_{Q,p}}\right).$$

The proton and neutron are both spin-1/2 particles—they admit two spin states, so $g_n = g_p = 2$. Moreover, we can approximate $n_{Q,n} \approx n_{Q,p}$ because their mass are very close to each other. Eventually, we will get

$$\frac{n_n}{n_p} = e^{-\chi_n/kT}, \quad \text{where} \quad \chi_n \equiv (m_n - m_p)c^2. \tag{9.9}$$

This simple equation states that as time goes on, and as temperature drops, proton is more favored because it is the lightest baryon. If protons and neutrons were always in equilibrium, then there will 1 neutron in a million protons after several minutes. Hence this equilibrium should not last very long.

Every time there is a neutrino present in a process, it is mediated by a weak interaction. Weak interactions have very small cross-sections compared to electromagnetic or strong interactions. At temperatures $kT \sim 1\,\text{MeV}$, the cross-section for $n + \nu_e \to p + e^-$ is

$$\sigma_w \sim 10^{-47}\,\text{m}^2 \left(\frac{kT}{1\,\text{MeV}}\right)^2 \propto t^{-1}.$$

(Recall that the Thomson scattering cross-section is much larger, $\sigma_\text{T} = 6.65 \times 10^{-29}\,\text{m}^2$.) Also, the number density of neutrinos falls off like $n_\nu \propto a^{-3} \propto t^{-3/2}$. Altogether, the interaction rate of the process $n + \nu_e \to p + e^-$ scales like

$$\Gamma = n_\nu c \sigma_w \propto t^{-5/2}.$$

The Hubble parameter only scales like $H = \dot{a}/a \propto t^{-1}$. Therefore, there exist a time/temperature where $\Gamma = H$. This particular time is called neutron-proton **freeze-out**. Essentially, we can also call it the time of neutrino decoupling. The freeze-out temperature is about $kT_\text{freeze} \approx 0.8\,\text{MeV}$, or $T_\text{freeze} \approx 9 \times 10^9\,\text{K}$. The age of the universe at this point is $t_\text{freeze} \sim 1\,\text{s}$. Using T_freeze, we can find the neutron-to-proton ratio to be $n_n/n_p = e^{-\chi_n/kT_\text{freeze}} \approx 0.2$, one neutron for every five protons. This explains why most of the mass today are in the form of free hydrogen (i.e. proton). There are not many neutrons to fuse with excessive protons, so heavier elements are scarce. In fact, most of the heavier elements are fused by stars and other astronomical events like merging neutron stars.

We don't need to worry about proton-proton fusion in Big Bang nucleosynthesis. It is nearly impossible for two protons to fuse into a ^{2}He and proceed further fusion to ^{4}He because of the Coulomb barrier between them. This is because the temperature decreased rapidly during nucleosynthesis and it would not be enough time for substantial proton-proton fusion reaction to occur. (By the same reason, stars can fuse protons into helium because they are stable and in hydrostatic equilibrium.) Therefore, we focus on neutron-proton fusion as there is no Coulomb barrier. The strong nuclear force will do the job to fuse a neutron and a proton into ^{2}D, a deuterium nucleus, or a deuteron.

9.2.2 Deuterium Synthesis

Neutrinos are now decoupled from baryons and electrons, but photons are not. The first major reaction in nucleosynthesis is

$$p + n \longleftrightarrow \mathrm{D} + \gamma, \quad \chi_{\mathrm{D}} = (m_n + m_p - m_{\mathrm{D}})c^2 = 2.22\,\mathrm{MeV},$$

where χ_{D} is the binding energy of a deuteron. The reverse process of the above reaction is the photodissociation, similar to ionization, but with a more energetic photon. This again suggests us to derive a Saha-like equation to study the number density of deuteron formed. The equilibrium condition is $\mu_n + \mu_p = \mu_{\mathrm{D}}$. Following the procedure to derive the Saha equation, it is straightforward to show that

$$\frac{n_p n_n}{n_{\mathrm{D}}} = \frac{g_p g_n}{g_{\mathrm{D}}} \left(\frac{m_p m_n}{m_{\mathrm{D}}}\right)^{3/2} \left(\frac{2\pi kT}{h^2}\right)^{3/2} e^{-\chi_{\mathrm{D}}/kT}.$$

The difference is that a deuteron has a degeneracy factor $g_{\mathrm{D}} = 3$, while $g_p = g_n = 2$. The masses are approximately related as $m_p \approx m_n \approx m_{\mathrm{D}}/2$, and we can get an equation for deuteron number density:

$$\frac{n_p n_n}{n_{\mathrm{D}}} = \frac{4}{3} \left(\frac{\pi m_p kT}{h^2}\right)^{3/2} e^{-\chi_{\mathrm{D}}/kT}. \tag{9.10}$$

Now we shall define a temperature at which the nucleosynthesis happens. Because neutrons are the limiting reactants, it is convenient to define T_{nuc} at which $n_n/n_{\mathrm{D}} = 1$. Half the free neutrons are fused into deuteron at T_{nuc}. To a rough estimate, we assume 80% of baryons to be protons,

$$n_p \approx 0.8 n_{\mathrm{bary}} = 0.8\eta n_\gamma = 0.8\eta b T^3 = 0.8\eta \left[1.949 \left(\frac{\pi kT}{hc}\right)^3\right].$$

Substituting n_p into (9.10),

$$\frac{n_n}{n_{\mathrm{D}}} = \frac{4}{3n_p} \left(\frac{\pi m_p kT}{h^2}\right)^{3/2} e^{-\chi_{\mathrm{D}}/kT} = \frac{5}{3\pi^{3/2} \times 1.949\eta} \left(\frac{m_p c^2}{kT}\right)^{3/2} e^{-\chi_{\mathrm{D}}/kT} \approx \frac{0.154}{\eta} \left(\frac{m_p c^2}{kT}\right)^{3/2} e^{-\chi_{\mathrm{D}}/kT}.$$

Plugging in all the numbers, we find that $kT_{\mathrm{nuc}} \approx 66\,\mathrm{keV}$, or $T_{\mathrm{nuc}} \approx 7.6 \times 10^8\,\mathrm{K}$. The age of the universe at nucleosynthesis is $t_{\mathrm{nuc}} \approx 200\,\mathrm{s}$. This timescale is not negligible compared to the mean lifetime of a neutron, $\tau_n = 880\,\mathrm{s}$. By t_{nuc}, the neutron-to-proton ratio drops from $n_n/n_p = 1/5$ to

$$\frac{n_n}{n_p} \approx \frac{e^{-200/880}}{5 + (1 - e^{-200/880})} \approx 0.15.$$

9.2.3 Beyond Deuterium

After the formation of large amount of deuteron, they can fuse into ^{3}H, ^{3}He, and ^{4}He in various ways:

$$\begin{aligned}
\mathrm{D} + p &\longleftrightarrow {}^3\mathrm{He} + \gamma, \\
\mathrm{D} + n &\longleftrightarrow {}^3\mathrm{H} + \gamma, \\
\mathrm{D} + \mathrm{D} &\longleftrightarrow {}^4\mathrm{He} + \gamma, \\
\mathrm{D} + \mathrm{D} &\longleftrightarrow {}^3\mathrm{H} + p, \\
\mathrm{D} + \mathrm{D} &\longleftrightarrow {}^3\mathrm{He} + n,
\end{aligned}$$

$$\begin{aligned}
{}^3\mathrm{H} + p &\longleftrightarrow {}^4\mathrm{He} + \gamma, \\
{}^3\mathrm{He} + n &\longleftrightarrow {}^4\mathrm{He} + \gamma, \\
{}^3\mathrm{H} + \mathrm{D} &\longleftrightarrow {}^4\mathrm{He} + n, \\
{}^3\mathrm{He} + \mathrm{D} &\longleftrightarrow {}^4\mathrm{He} + p.
\end{aligned}$$

Tritium is unstable, as it decays via the process ${}^3\mathrm{H} \to {}^3\mathrm{He} + e^- + \bar{\nu}_e$. However, its mean lifetime is relatively long (18 years) compared to the timescale of nucleosynthesis. All other processes above involve only strong interactions, so their cross-sections are much larger than that of weak interactions. The production of ^{4}He was very efficient during nucleosynthesis.

If one look at a plot of binding energy per nucleon as a function of mass number A, ^{4}He is a highly stable nucleus compared to its neighboring nuclei such as ^{5}He, ^{5}Li. Thus, fusing additional proton or neutron to ^{4}He has a small chance. Instead, ^{6}Li, ^{7}Li, and ^{7}Be (mean lifetime 53 days) are formed through

$$\begin{aligned}
{}^4\mathrm{He} + \mathrm{D} &\longleftrightarrow {}^6\mathrm{Li} + \gamma, \\
{}^4\mathrm{He} + {}^3\mathrm{H} &\longleftrightarrow {}^7\mathrm{Li} + \gamma, \\
{}^4\mathrm{He} + {}^3\mathrm{He} &\longleftrightarrow {}^7\mathrm{Be} + \gamma.
\end{aligned}$$

Again, there are no stable nuclei with mass number $A = 8$. Most heavier elements are not formed during nucleosynthesis. When the age of the universe is around $t \sim 1000\,\mathrm{s}$, nucleosynthesis is essentially over. Most nuclei are in the form of either protons or ^{4}He.

To determine the fraction of primordial elements formed, one need to rely on computers. The primordial helium mass fraction Y_p is determined to be 0.24. Ratios such as D/H, ^{3}He/H, and ^{7}Li/H varies with baryon-to-ratio η. If we know some data about these ratios, it is possible to deduce η. For example, distant quasars that illuminate intergalactic clouds reveals D/H to be $(2.53 \pm 0.04) \times 10^{-5}$. Results from computers show that this gives a baryon-to-photon ratio $\eta = (6.0 \pm 0.1) \times 10^{-10}$, which is consistent to what we find from the CMB.

9.2.4 Baryon-Antibaryon Asymmetry

The energy density at nucleosynthesis $T_{\mathrm{nuc}} \approx 7.6 \times 10^8\,\mathrm{K}$ is

$$\epsilon_{\mathrm{nuc}} = a_{\mathrm{rad}} T^4 \approx 1.6 \times 10^{33}\,\mathrm{MeV/m^3},$$

which corresponds to a mass density $\rho_{\mathrm{nuc}} \approx 2800\,\mathrm{kg/m^3}$. However, most energy density were in the form of radiation. The mass density of baryons were only

$$\rho_{\mathrm{bary}}(t_{\mathrm{nuc}}) = \Omega_{\mathrm{bary},0}\rho_{c,0}\left(\frac{T_{\mathrm{nuc}}}{T_0}\right)^3 \approx 0.009\,\mathrm{kg/m^3}.$$

The low baryon-to-photon is puzzling, but there is an even more puzzling asymmetry called **baryon-antibaryon asymmetry**. This is obvious: almost all baryonic matter are in the form of baryons today. Antibaryons, such as antiproton or antineutron are only produced in colliders. To find the reason behind $n_{\mathrm{antibary}} \ll n_{\mathrm{bary}} \ll n_\gamma$, we need to go to the *very* early universe.

When the age of the universe were only a few microseconds, the temperature was $kT \gtrsim 150\,\mathrm{MeV}$. Quarks and antiquarks at this time were not confined in hadrons (baryons and mesons). Pair production and annihilation of quarks were possible,

$$\gamma + \gamma \longleftrightarrow q_i + \bar{q}_i,$$

where q_i represents the quark flavor: up, down, strange, charm, bottom, top (u, d, s, c, b, t). The significant baryon-antibaryon asymmetry can be explained by the following—it was not significant at the beginning. Suppose the asymmetry at the beginning of the universe were

$$\delta_q \equiv \frac{n_q - n_{\bar{q}}}{n_q + n_{\bar{q}}} \ll 1.$$

As the universe cooled, quark-antiquark pairs could not be produced, but they can still annihilate to photons. Also, quark confinement forced quarks to form bound states into protons and neutrons. The very few leftover baryons that do not have a partner to annihilate are what make us today. It seems like the number of baryons and antibaryons are asymmetric, but there were just more of them in the past. This also explains the low baryon-to-photon ratio. The majority of quarks and antiquarks were annihilate into photons, causing $n_\gamma \gg n_{\mathrm{bary}}$. However, the exact mechanism that produce that δ_q is still unknown.

9.3 Inflation

The classical Hot Big Bang model is a success of cosmology, supported by the cosmic microwave background and other evidence. Nonetheless, it has several problems. The most famous ones are known as the **flatness problem**, the **horizon problem**, and the **monopole problem**.

9.3.1 The Flatness Problem

According to the ΛCDM model, the universe today is nearly flat, with a constraint on the current density parameter being $|1-\Omega_0| \leq 0.005$. What about the curvature in the past? The Friedmann equation in terms of the density parameter is given by (6.16):

$$1-\Omega(t) = -\frac{\kappa c^2}{R_0^2 a^2 H^2}, \quad \text{with} \quad 1-\Omega_0 = -\frac{\kappa c^2}{R_0^2 H_0^2}.$$

Combining the two equations gives the density parameter as a function of time,

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

Let's focus on the era where radiation and matter dominate. The Friedmann equation of a radiation + matter universe is (7.27)

$$\frac{H(t)^2}{H_0^2} = \frac{\Omega_{r,0}}{a^4} + \frac{\Omega_{m,0}}{a^3}.$$

Then we can write the density parameter as a function of scale factor only:

$$1-\Omega(t) = \frac{(1-\Omega_0)a^2}{\Omega_{r,0} + a\Omega_{m,0}}. \tag{9.11}$$

In radiation-dominated era and matter-dominated era, we have

$$|1-\Omega|_r \propto a^2 \propto t, \quad |1-\Omega|_m \propto a \propto t^{2/3}.$$

Suppose the limit $|1-\Omega_0| \leq 0.005$ is accurate. Then at radiation-matter equality $a_{rm} \approx 2.9 \times 10^{-4}$, $|1-\Omega_{rm}| \leq 2 \times 10^{-6}$. At Big Bang nucleosynthesis $a_{\text{nuc}} \approx 3.6 \times 10^{-9}$, $|1-\Omega_{\text{nuc}}| \leq 7 \times 10^{-16}$. If we continue to smaller scale factor, or the earlier universe, the density parameter is closer and closer to 1. At the Planck time $t_{\text{P}} \approx 5 \times 10^{-44}\,\text{s}$, the universe has nearly no curvature,

$$|1-\Omega_{\text{P}}| \leq 2 \times 10^{-62}.$$

This is the flatness problem. It is to accept it as a *coincidence* that the universe is so close to flat at the start. *If* the universe were not this flat, it will result in a Big Crunch ($\Omega > 1$) or a Big Chill ($\Omega < 1$) very soon after it was born. The universe today would be very different.

9.3.2 The Horizon Problem

So far we always assume the cosmological principle, which states that the universe is homogeneous and isotropic. It is convenient to work with in cosmology, but meanwhile it is also true on large scales based on observations. However, it does not violate any physical law for the universe to be not homogeneous or isotropic.

The actual problem is even more serious. If we look at the cosmic microwave background, we are looking at the surface of last scattering, which is almost perfectly homogeneous. This means these CMB photons were in equilibrium before the last scattering. Now consider the one patch at $(\theta, \phi) = (\pi/2, \phi)$ on the last scattering surface. Its current proper distance to us is

$$d_p(t_0) = c\int_{t_{\text{ls}}}^{t_0} \frac{dt}{a(t)} \approx 0.98 d_{\text{hor}}.$$

We we look exactly in the opposite direction $(\pi/2, \phi+\pi)$, this patch looks almost identical (in temperature, for example) to the previous one. Also, it is located $0.98d_{\text{hor}}$ away from us. This means the two patches, separated by a distance $1.96d_{\text{hor}} > d_{\text{hor}}$, look almost identical. In principle, two points separated by a distance $d_p > d_{\text{hor}}$ are not causally related—they do not and *did* not have chance to communicate with each other unless something is traveling faster than the speed of light. In fact, the two patches need not be opposite. They just need to have an angular separation of

$$\theta > \theta_{\text{hor}} = \frac{d_{\text{hor}}(t_{\text{ls}})}{d_A} \approx \frac{0.251\,\text{Mpc}}{12.8\,\text{Mpc}} \approx 1.1^\circ.$$

Therefore, there exist thousands of such patches on the last scattering surface that look the same but not causally related. It is hard to explain it by a coincidence.

9.3.3 The Monopole Problem

To be written.

9.3.4 The Inflation Solution

The **inflation** theory hypothesizes that there was a period in the very early universe where the expansion was accelerating outward. According to the acceleration equation (6.20),

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(\epsilon + 3P),$$

there existed a component where $w < -1/3$ that was dominating during inflation. The simplest component with $w < -1/3$ would be a cosmological constant Λ_i, where the subscript i stands for inflation. The Friedmann equation is of the form

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{\Lambda_i c^2}{3},$$

with an exponential solution

$$a(t) \propto e^{H_i t}, \quad \text{where} \quad H_i = \left(\frac{\Lambda_i c^2}{3}\right)^{1/2}.$$

The inflation theory solves the flatness problem and the horizon problem. For simplicity, suppose at some time t_i in the radiation dominated phase, Λ_i suddenly dominated over all other components, and ended in t_f. Hence the growth of the scale factor switched to $e^{H_i t}$ from t_i to t_f. In this case, the scale factor can be written as

$$a(t) = \begin{cases} a_i (t/t_i)^{1/2}, & t < t_i, \\ a_i e^{H_i (t - t_i)}, & t_i \le t < t_f, \\ a_i e^{H_i (t_f - t_i)} (t/t_f)^{1/2}, & t \ge t_f. \end{cases}$$

The growth in the scale factor between t_i and t_f, is then e^N, were $N \equiv H_i(t_f - t_i)$. During inflation, we want $t_f - t_i \ll H_i^{-1}$ so the exponential growth is large. Consider the following example. Let $t_i \approx t_{\text{GUT}} \approx 10^{-36}\,\text{s}$ and $H_i \approx t_{\text{GUT}}^{-1} \approx 10^{36}\,\text{s}^{-1}$. The energy density of the cosmological constant (6.22) would be

$$\epsilon_{\Lambda_i} = \frac{c^4}{8\pi G}\Lambda_i = \frac{3c^2}{8\pi G} H_i^2 \sim 10^{105}\,\text{TeV/m}^3,$$

which is enormous compared to $\epsilon_{\Lambda,0} = 0.69\epsilon_{c,0} \approx 0.0034\,\text{TeV/m}^3$.

Solution to the Flatness Problem. Recall that the Friedmann equation in terms of the density parameter is

$$1 - \Omega(t) = -\frac{\kappa c^2}{R_0^2 a^2 H^2} \quad \Longleftrightarrow \quad |1 - \Omega(t)| = \frac{c^2}{R_0^2 a^2 H^2}$$

for any non-flat universe. During inflation, $H(t) = H_i = \text{const.}$ and $a(t) \propto e^{H_i t}$, which means

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

The deviation of the density parameter from 1, or equivalently the curvature, was suppressed exponentially. If $t_f = (N+1)t_i \approx (N+1) \times 10^{-36}\,\text{s}$, then

$$|1 - \Omega(t_f)| = e^{-2N} |1 - \Omega(t_i)|.$$

In the ΛCDM model, we can calculate the scale factor all the way to t_f, where $a(t_f) \approx 2 \times 10^{-28}\sqrt{N+1}$. With $|1 - \Omega_0| \le 0.005$ and (9.11), the deviation of Ω from 1 at t_f will be

$$|1 - \Omega(t_f)| \le 2 \times 10^{-54}(N+1).$$

Relating to the start of inflation t_i, we have

$$e^{-2N}|1-\Omega(t_i)| \leq 2\times 10^{-54}(N+1). \tag{9.12}$$

If $|1-\Omega(t_i)| \sim 1$, the number of e-folding is $N \approx 60$, or a growth in scale factor of $a(t_f)/a(t_i) \sim e^{60} \sim 10^{26}$. In the future we may discover that the current $|1-\Omega_0|$ is more restricted than 0.005. Then it requires N to be greater than 60.

Solution to the Horizon Problem. Recall that the horizon distance at any time t is given by (7.12)

$$d_{\rm hor}(t) = a(t)c\int_0^t \frac{dt}{a(t)}.$$

Before inflation, $a = a_i(t/t_i)^{1/2}$, so the horizon distance at t_i was

$$d_{\rm hor}(t_i) = a_i c\int_0^{t_i} \frac{dt}{a_i(t/t_i)^{1/2}} = 2ct_i,$$

where $a_i = a(t_i)$. At the end of inflation, the horizon distance is

$$d_{\rm hor}(t_f) = a(t_f)c\int_0^{t_f}\frac{dt}{a(t)} = a_i e^N c\left[\int_0^{t_i}\frac{dt}{a_i(t/t_i)^{1/2}} + \int_{t_i}^{t_f}\frac{dt}{a_i e^{H_i(t-t_i)}}\right].$$

For the second integral,

$$\int_{t_i}^{t_f}\frac{dt}{a_i e^{H_i(t-t_i)}} = \frac{e^{H_i t_i}}{a_i}\int_{t_i}^{t_i(N+1)} e^{-H_i t}\,dt = -\frac{e^{H_i t_i}}{a_i H_i}e^{-H_i t}\bigg|_{t_i}^{t_i(N+1)} = \frac{1}{a_i H_i}(1-e^{-H_i t_i N}).$$

If N is large, we can neglect $e^{-H_i t_i N}$, and the horizon distance at the end of inflation was

$$d_{\rm hor}(t_f) = e^N c(2t_i + H_i^{-1}).$$

Let's plug in some numbers. If $t_i \approx 10^{-36}\,\text{s}$, the horizon distance is $d_{\rm hor}(t_i) = 2ct_i \approx 6\times 10^{-28}\,\text{m}$. Assume that $N = 65$, a bit more than minimum $N_{\rm min} = 60$, and take $H_i \approx t_i^{-1}$. The horizon distance immediately after inflation was $d_{\rm hor}(t_f) \approx e^N(3ct_i) \sim 15\,\text{m}$. The scale factor was $a(t_f) \approx 2\times 10^{-28}\sqrt{N+1} \sim 2\times 10^{-27}$. If we continue integrating the horizon distance to the time of last scattering where $a(t_{\rm ls}) \approx 9.1\times 10^{-4}$, we will find that $d_{\rm hor}(t_{\rm ls}) \approx 200\,\text{Mpc}$. Compared to $d_{\rm hor}(t_{\rm ls}) \approx 0.25\,\text{Mpc}$ calculated without inflation (9.6), the one with inflation is 800 times larger.

To show that inflation enables causality, we will calculate the proper distance related to the surface of last scattering. The current proper distance to the surface of last scattering is $d_p(t_0) \approx 14000\,\text{Mpc}$. When $a(t_f) \sim 2\times 10^{-27}$, this surface is in a radius of $d_p(t_f) = a(t_f)d_p(t_0) \sim 0.9\,\text{m}$. Just before the inflation, $a(t_i) = e^{-N}a(t_f) \sim 10^{-55}$, this surface has a radius of $d_p(t_i) \sim 5\times 10^{-29}\,\text{m}$. This distance is smaller than $d_{\rm hor}(t_i) \sim 2ct_i \sim 6\times 10^{-28}\,\text{m}$. Thus, they definitely had time to exchange information and reach an equilibrium before inflation. This explains homogeneity and isotropy of the current universe.

10 Structure Formation

The universe on a scale smaller than $\sim 100\,\mathrm{Mpc}$ is certainly not homogeneous. There are structures formed by self gravitation, and they can be seen as density fluctuations in the universe. On scales $\sim 50\,\mathrm{Mpc}$, the high density regions are known as **superclusters**, and the low density regions are known as **voids**. Superclusters are typically planar or linear structures, while the underdense voids are typically spherical. Superclusters are made of galaxy clusters, and galaxy clusters contain galaxies, which then contains stars, planets, and so on.

10.1 Gravitational Instability

We already know that density is not homogeneous at the time of last scattering, inferred from the cosmic microwave background. A slight inhomogeneity of density will cause regions to have stronger gravity than others. If the density is sufficiently large—large enough to overcome the expansion of the universe—then they will collapse and become gravitationally bound objects. Such a basic mechanism for growing large structures is **gravitational instability**. In addition, overdense regions will attract more objects from underdense regions, making them even more denser. This is sometimes called the **Matthew effect** (the rich get richer and the poor get poorer).

10.1.1 Gravitational Collapse

Quantitatively, let the energy density of the universe be a function of position and time, $\epsilon(\mathbf{r}, t)$. At any time t, there is a spatially averaged density

$$\bar{\epsilon}(t) = \frac{1}{V}\int_V \epsilon(\mathbf{r}, t)\, d^3\mathbf{r},$$

where the volume V over which we are averaging should be as large as possible. Define the **density contrast** $\delta(\mathbf{r}, t)$ to be

$$\delta(\mathbf{r}, t) \equiv \frac{\epsilon(\mathbf{r}, t) - \bar{\epsilon}(t)}{\bar{\epsilon}(t)} = \frac{\epsilon}{\bar{\epsilon}} - 1. \tag{10.1}$$

It is negative in underdense regions and positive in overdense regions. Its minimum value is -1, corresponding to $\epsilon = 0$. There is no upper limit for $\delta(\mathbf{r}, t)$, though in most cases we also want it to be small. Consider a region of the universe that is approximately static, homogeneous, pressureless, with density $\bar{\rho}$. Now add a sphere of radius R with some mass ΔM such that within the sphere the density is $\rho = \bar{\rho}(1 + \delta)$, where $\delta \ll 1$. If δ is uniform inside the sphere, the gravitational acceleration at the surface of the sphere is

$$\ddot{R} = -\frac{G\Delta M}{R^2} = -\frac{G}{R^2}\left(\frac{4\pi R^3 \bar{\rho}\delta}{3}\right) \quad \Longrightarrow \quad \frac{\ddot{R}}{R} = -\frac{4\pi G\bar{\rho}}{3}\delta(t). \tag{10.2}$$

By conservation of mass, the mass of the sphere is held constant at any time

$$M = \frac{4\pi R^3}{3}\bar{\rho}(1 + \delta).$$

Then we get the second equation involving $R(t)$ and $\delta(t)$,

$$R(t) = R_0(1 + \delta)^{-1/3} \quad \text{where} \quad R_0 \equiv \left(\frac{3M}{4\pi\bar{\rho}}\right)^{1/3}.$$

If $|\delta| \ll 1$, we can approximate $R(t)$ and its second derivative as

$$R(t) \simeq R_0\left(1 - \frac{1}{3}\delta\right), \quad \ddot{R} \simeq -\frac{1}{3}R_0\ddot{\delta} = -\frac{1}{3}R(1+\delta)^{1/3}\ddot{\delta} \simeq -\frac{1}{3}R\left(1 + \frac{1}{3}\delta\right)\ddot{\delta} \simeq -\frac{1}{3}R\ddot{\delta}.$$

Hence the mass conservation says

$$\frac{\ddot{R}}{R} \simeq -\frac{1}{3}\ddot{\delta}, \qquad [\delta \ll 1]. \tag{10.3}$$

Combining this with (10.2), we get a simple equation for the density contrast,

$$\ddot{\delta} = 4\pi G\bar{\rho}\delta.$$

The general solution of this equation are exponentials:

$$\delta(t) = A_1 e^{t/t_{\rm dyn}} + A_2 e^{-t/t_{\rm dyn}}, \quad \text{where} \quad t_{\rm dyn} \equiv \frac{1}{\sqrt{4\pi G\bar{\rho}}} \approx 9.6\,\text{hr}\left(\frac{\bar{\rho}}{1\,\text{kg/m}^3}\right)^{-1/2}.$$

is the dynamical timescale, depending only on $\bar{\rho}$. After a sufficient amount of time, $\delta(t)$ is dominated by the growing exponential. The constants A_1 and A_2 are fixed by initial conditions like $\delta(0)$ and $\dot{\delta}(0)$.

10.1.2 The Jeans Length

The dynamical timescale is a short timescale, cosmologically speaking. For example, if $\bar{\rho} = 1\,\text{kg/m}^3$, the $t_{\rm dyn} \sim 10$ hours. The thing that prevents gravitational collapse is pressure. For a sphere of gas in a background mean density $\bar{\rho} = \bar{\epsilon}/c^2$, the dynamical timescale is

$$t_{\rm dyn} \sim \frac{1}{\sqrt{G\bar{\rho}}} = \left(\frac{c^2}{G\bar{\epsilon}}\right)^{1/2}.$$

The pressure gradient does not respond immediately after a density fluctuation. Instead, they travel at sound speeds c_s. For a gas with equation-of-state parameter $w > 0$, the sound speed is given by

$$c_s = c\left(\frac{dP}{d\epsilon}\right)^{1/2} = \sqrt{w}c. \tag{10.4}$$

The timescale for the pressure gradient to build up in a sphere of radius R is then

$$t_P \sim \frac{R}{c_s} = \frac{R}{\sqrt{w}c}.$$

To establish hydrostatic equilibrium, the pressure timescale should be smaller than the dynamical timescale, $t_P < t_{\rm dyn}$, or

$$\frac{R}{c_s} \lesssim \left(\frac{c^2}{G\bar{\epsilon}}\right)^{1/2} \quad \Longrightarrow \quad R \lesssim c_s\left(\frac{c^2}{G\bar{\epsilon}}\right)^{1/2}.$$

In other words, the radius of the sphere needs to be smaller than some length scale to establish hydrostatic equilibrium. This length scale is the **Jeans length**, defined by

$$\lambda_{\rm J} = c_s\left(\frac{\pi c^2}{G\bar{\epsilon}}\right) = 2\pi c_s t_{\rm dyn}. \tag{10.5}$$

(The factor of π requires a more precise derivation.) If hydrostatic equilibrium is established, then the density fluctuations just oscillate. If the radius of the density fluctuation is much larger than the Jeans length, then there is not enough time for the pressure gradient to build up, so gravitational collapse happens. On cosmological scales, consider a flat universe with mean energy density $\bar{\epsilon}$ and $|\delta| \ll 1$. The characteristic time for the expansion of the universe is the Hubble time

$$H^{-1} = \left(\frac{3c^2}{8\pi G\bar{\epsilon}}\right)^{1/2} = \left(\frac{3}{2}\right)^{1/2} t_{\rm dyn},$$

from the Friedmann equation (6.13) with $\kappa = 0$. Note that the Hubble time is of the same order as the dynamical timescale. The Jeans length, expressed in terms of the Hubble parameter, is

$$\lambda_{\rm J} = 2\pi c_s t_{\rm dyn} = 2\pi\left(\frac{2}{3}\right)^{1/2}\frac{c_s}{H} = 2\pi\left(\frac{2w}{3}\right)^{1/2}\frac{c}{H}. \tag{10.6}$$

The Jeans length of a universe with only radiation ($w = 1/3$) is $\lambda_{\rm J,rad} \approx 3c/H$, three times the Hubble distance. This Jeans length is too large to have gravitationally collapse in a scale smaller than the Hubble distance. To have gravitationally collapsed structures much smaller than the Hubble distance, like galaxy clusters, the universe must have a non-relativistic component with $\sqrt{w} \ll 1$.

10.1.3 Structure Formation

An important time for gravitational collapse is the time of photon decoupling/last scattering. Before photon decoupling ($z_{\rm dec} \approx z_{\rm ls} \approx 1090$, $t \approx 0.37\,\text{Myr}$), all photons, electrons, and baryons are in a photon-baryon fluid. More specifically, the energy density of photons still dominates over the energy density of baryons at that time. The Jeans length is approximately

$$\lambda_{\rm J}(\text{before}) \approx \frac{3c}{H(z_{\rm dec})} \approx 0.66\,\text{Mpc} \approx 2.0\times10^{22}\,\text{m}.$$

The baryonic Jeans mass, the mass of baryons in a sphere of radius $\lambda_{\rm J}$, is

$$M_{\rm J} \equiv \left(\frac{4\pi\lambda_{\rm J}^3}{3}\right)\rho_{\rm bary}. \tag{10.7}$$

Just before the decoupling, $\rho_{\rm bary} \approx 5.6\times10^{-19}\,\text{kg/m}^3$, the baryonic Jeans mass was

$$M_{\rm J} = \frac{4\pi\lambda_{\rm J}(\text{before})^3}{3}\rho_{\rm bary} \approx 2\times10^{49}\,\text{kg} \approx 10^{19}\,M_\odot,$$

much larger than a typical supercluster ($M_{\rm bary}\sim10^{16}\,M_\odot$). Immediately after the decoupling, photons and baryons no longer exchange energies and become two separate gases. The sound speed of the baryonic gas is

$$c_{s,\text{bary}} = \sqrt{w}c = \left(\frac{kT}{\bar{m}c^2}\right)^{1/2}c,$$

where $\bar{m}$ is the average mass of baryonic particles. Immediately after the decoupling, $kT_{\rm dec}\approx 0.26\,\text{eV}$ and the average rest mass energy of atoms is

$$\bar{m}c^2 = (1-Y_p)m_pc^2 + Y_p(4m_p)c^2 \approx 1.22m_pc^2 \approx 1140\,\text{MeV},$$

where $Y_p = 0.24$ is the primordial helium mass fraction. The sound speed of the baryonic gas is then $c_{s,\text{bary}}\approx 1.5\times10^{-5}c$. Compared to the sound speed of the photon gas, $c_{s,\gamma} = c/\sqrt{3}\approx 0.58c$, they differ by a factor of 2.6×10^{-5}. Since the Jeans length is linearly proportional to the sound speed, $\lambda_{\rm J}(\text{after})/\lambda_{\rm J}(\text{before})\approx 2.6\times10^{-5}$. Therefore, the baryonic Jeans mass changes drastically,

$$\frac{M_{\rm J}(\text{after})}{M_{\rm J}(\text{before})} = \left[\frac{\lambda_{\rm J}(\text{after})}{\lambda_{\rm J}(\text{before})}\right]^3 \quad\Longrightarrow\quad M_{\rm J}(\text{after}) = (2.6\times10^{-5})^3 M_{\rm J}(\text{before}) \approx 2\times10^5\,M_\odot.$$

It is even smaller than the Small Megellanic cloud ($M_{\rm bary}\sim10^9\,M_\odot$). In conclusion, after the photon decoupling, density perturbations within regions with $M > M_{\rm J}\approx 2\times10^5\,M_\odot$ will undergo gravitational collapse. Structure formation is triggered after the photon decoupling.

10.2 Instability in an Expanding Universe

In the last section, we concluded that in such a universe $\delta\sim e^{t/t_{\rm dyn}}$, but assuming that the universe is static. We also find that the dynamical timescale $t_{\rm dyn}$ and the expansion timescale H^{-1} are of the same order. This tells us that expansion also plays a role in the density perturbation.

10.2.1 A Cheat by Newtonian Gravity II

Consider a universe that contains only matter. The average density decreases as the universe expands by $\bar{\rho}(t)\propto a(t)^{-3}$. A sphere of radius R and uniform density contrast has matter density

$$\rho(t) = \bar{\rho}(t)[1+\delta(t)].$$

Assume that $|\delta|\ll 1$ and all its derivatives $\dot{\delta}\ll 1$, $\ddot{\delta}\ll 1$, etc. Mass conservation inside the sphere says that

$$M = \frac{4\pi R^3\bar{\rho}}{3}(1+\delta).$$

Solving for R,

$$R(t)\propto \bar{\rho}^{-1/3}(1+\delta)^{-1/3}\propto a(1+\delta)^{-1/3}.$$

The constant of proportionality is unimportant; call it A so that $R(t) = Aa(1+\delta)^{-1/3}$. Taking the time derivatives of this equation,

$$\dot{R} = A\dot{a}(1+\delta)^{-1/3} - \frac{1}{3}Aa(1+\delta)^{-4/3}\dot{\delta},$$
$$\ddot{R} = A\ddot{a}(1+\delta)^{-1/3} - \frac{1}{3}A\dot{a}(1+\delta)^{-4/3}\dot{\delta} - \frac{1}{3}A\dot{a}(1+\delta)^{-4/3}\dot{\delta} + \frac{4}{9}Aa(1+\delta)^{-7/3}\dot{\delta}^2 - \frac{1}{3}Aa(1+\delta)^{-4/3}\ddot{\delta}.$$

Taking the ratio of $\ddot{R}$ and R,

$$\frac{\ddot{R}}{R} = \frac{\ddot{a}}{a} - \frac{2}{3}\frac{\dot{a}}{a}(1+\delta)^{-1}\dot{\delta} + \frac{4}{9}(1+\delta)^{-2}\dot{\delta}^2 - \frac{1}{3}(1+\delta)^{-1}\ddot{\delta}.$$

Since we assume that $\delta, \dot{\delta}, \ddot{\delta} \ll 1$, we can expand this result to leading order in δ and all its derivatives,

$$\frac{\ddot{R}}{R} \simeq \frac{\ddot{a}}{a} - \frac{2\dot{a}}{3a}\dot{\delta} - \frac{1}{3}\ddot{\delta}. \tag{10.8}$$

This is what mass conservation tells us. The gravitational acceleration at the surface of the sphere is

$$\ddot{R} = -\frac{GM}{R^2} = -\frac{G}{R^2}\left(\frac{4\pi R^3\bar{\rho}}{3}\right) = -\frac{4\pi G R\bar{\rho}}{3}(1+\delta).$$

This gives the equation of motion for a point at the surface of the sphere,

$$\frac{\ddot{R}}{R} = -\frac{4\pi G\bar{\rho}}{3}(1+\delta). \tag{10.9}$$

Combining (10.8) and (10.9), we have an equation relating the scale factor and the matter density,

$$\frac{\ddot{a}}{a} - \frac{2\dot{a}}{3a}\dot{\delta} - \frac{1}{3}\ddot{\delta} = -\frac{4\pi G\bar{\rho}}{3}(1+\delta).$$

The first term on the LHS cancels with the first term on the RHS because the background density ($\delta = 0$) and scale factor satisfy

$$\frac{\ddot{a}}{a} = -\frac{4\pi G\bar{\rho}}{3}.$$

Thus, we have a linear differential equation for the perturbation,

$$-\frac{1}{3}\ddot{\delta} - \frac{2\dot{a}}{3a}\dot{\delta} + \frac{4\pi G\bar{\rho}}{3}\delta = 0 \quad \Longrightarrow \quad \ddot{\delta} + 2H\dot{\delta} - 4\pi G\bar{\rho}\delta = 0. \tag{10.10}$$

In a static universe ($H = 0$), this equation reduces to $\ddot{\delta} = 4\pi G\bar{\rho}\delta$ and $\delta \propto e^{t/t_{\text{dyn}}}$. Because the above equation is like a damped harmonic oscillator equation, the term $2H\dot{\delta}$ is sometimes called the *Hubble friction*. It is this turn that slows down the gravitational collapse.

10.2.2 General Relativistic Perturbation

Equation (10.10) is not far from the density perturbation equation derived from general relativity,

$$\ddot{\delta} + 2H\dot{\delta} - \frac{4\pi G\bar{\epsilon}_m}{c^2}\delta = 0. \tag{10.11}$$

Here ϵ_m is the matter energy density. The relativistic version (10.11) also generalize to a multiple-component universe, where ϵ_m may be only a small part of the total energy density ϵ. But still, δ is the density contrast of matter alone. It is more convenient to write (10.11) in terms of the density parameter,

$$\ddot{\delta} + 2H\dot{\delta} - \frac{3}{2}\Omega_m H^2\delta = 0, \quad \text{where} \quad \Omega_m = \frac{\bar{\epsilon}_m}{\epsilon_c} = \frac{8\pi G\bar{\epsilon}_m}{3c^2H^2}. \tag{10.12}$$

We will now analyze when the matter perturbation will grow rapidly. Consider a flat universe with $\Omega_0 = 1$. In radiation era $a \ll 1$, $\Omega_m \ll \Omega_r \approx \Omega_0$, the scale factor $a(t)$ and the Hubble parameter $H(t)$ grows like

$$a(t) = \left(\frac{t}{t_0}\right)^{1/2} \quad \text{and} \quad H(t) = \frac{\dot{a}}{a} = \frac{1}{2t}.$$

Then (10.12) becomes

$$\ddot{\delta} + \frac{1}{t}\dot{\delta} \simeq 0 \quad \Longrightarrow \quad \delta(t) \simeq A_1 + A_2 \ln t,$$

where A_1 and A_2 are integration constants fixed by initial conditions. The density contrast grows only logarithmically. In Lambda-dominated era far in the future, $a \gg 1$, $\Omega_m \ll \Omega_\Lambda \approx \Omega_0$, the scale factor and the Hubble parameter grows like

$$a(t) = e^{H_\Lambda(t-t_0)}, \quad H(t) = \frac{\dot{a}}{a} = H_\Lambda \quad \text{where} \quad H_\Lambda = \left(\frac{8\pi G\epsilon_\Lambda}{3c^2}\right)^{1/2}.$$

(See Equation 7.16.) Then (10.12) becomes

$$\ddot{\delta} + 2H_\Lambda\dot{\delta} \simeq 0 \quad \Longrightarrow \quad \delta(t) \simeq A_1 + A_2 e^{-2H_\Lambda t}.$$

It is basically a constant as time goes on, but meanwhile, the matter density decays like $a^{-3} = e^{-3H_\Lambda t}$. Therefore, the density fluctuations that make up all the structures must form during the matter-dominated era. In a matter dominated universe, $\Omega_m \approx \Omega_0 = 1$, and

$$a(t) = \left(\frac{t}{t_0}\right)^{2/3} \quad \text{and} \quad H(t) = \frac{2}{3t}.$$

(10.12) becomes

$$\ddot{\delta} + \frac{4}{3t}\dot{\delta} - \frac{2}{3t^2}\delta \simeq 0.$$

Guess the solution to be in the form At^ξ, and

$$\xi(\xi-1)At^{\xi-2} + \frac{3}{4t}\xi At^{\xi-1} - \frac{2}{3t^2}At^\xi = 0.$$

Canceling out all the $At^{\xi-2}$, we have an equation for ξ,

$$\xi(\xi-1) + \frac{4}{3}\xi - \frac{2}{3} = 0 \quad \Longrightarrow \quad \xi_1 = -1, \quad \xi_2 = \frac{2}{3}.$$

Hence the solution of the density contrast is

$$\delta \simeq A_1 t^{2/3} + A_2 t^{-1}. \tag{10.13}$$

The dominant term as time goes on is $t^{2/3}$, so

$$\delta \propto t^{2/3} \propto a(t) \propto \frac{1}{1+z}.$$

But remember that the perturbation is only valid when $|\delta| \ll 1$.

The matter perturbation here includes both baryonic matter and dark matter, and is in fact dominated by dark matter. In the last section we concluded that baryonic matter collapses after photon decoupling z_{dec} because before that moment baryonic matter and radiation are in one fluid. If baryonic matter were the dominant part of matter density, then they will follow the $\delta \propto t^{2/3} \propto a$ trend above. However, dark matter is the dominant one. Dark matter do not interact with photons or any other particles we know of, so they are not in a fluid with radiation. As a consequence, they follow $t^{2/3}$ right after the moment when matter dominates over radiation, at $z \lesssim z_{rm}$. But radiation-matter equality $z_{rm} = 3440$ occurs before photon decoupling z_{dec}, which means dark matter gets a head start in structure formation. When photons decouple with baryonic matter at z_{dec}, structures of dark matter already form. Baryons then fall into the gravity well of dark matter structure, so the density contrast of baryonic matter increase rapidly (see Figure 10.1).

10.3 The Power Spectrum

In previous analysis, we assume that the density perturbation is spherically symmetric. This is usually not the case, but (10.11) and (10.12) generalize to low-amplitude perturbations of any shape. At some time t_i in the universe, as long as the density contrast $|\delta| \ll 1$, the expansion of the universe is nearly isotropic and homogeneous:

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

It is convenient to label the density contrast using the comoving coordinate $\mathbf{r} = (r, \theta, \phi)$, $\delta(\mathbf{r})$. We are interested in the statistical properties of $\delta(\mathbf{r})$, just as the temperature fluctuation $\delta T/T(\theta, \phi)$. But instead of using spherical harmonics Y_ℓ^m, the density contrast

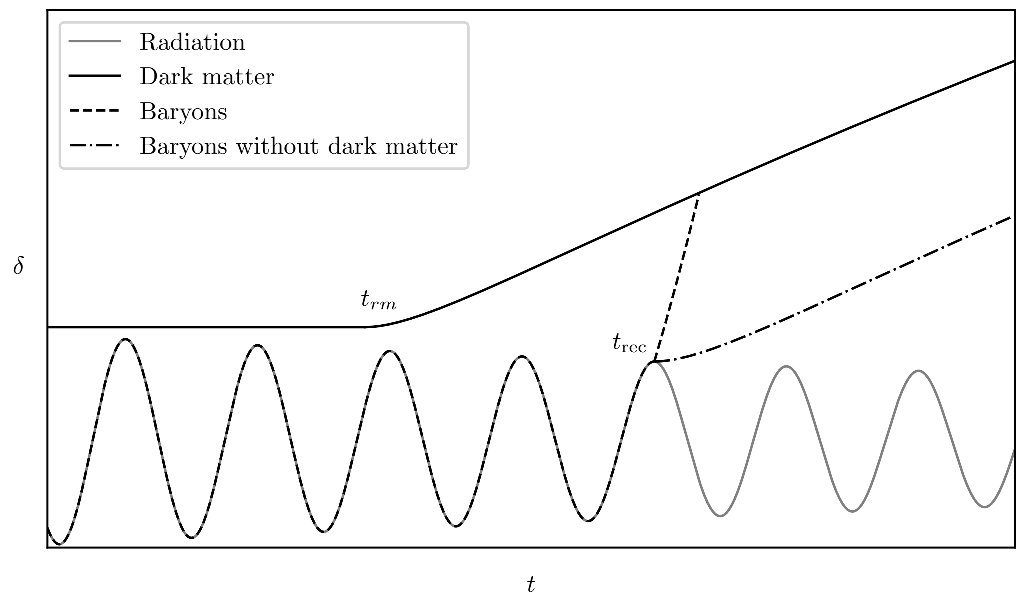


Figure 10.1: A schematic figure. Baryonic matter density perturbation increase rapidly with dark matter getting a head start. Photon density perturbation is oscillating because its Jeans length is too large.

δ, as defined in a three-dimensional space, is expanded in terms of Fourier components:

$$\delta(\mathbf{r}) = \frac{V}{(2\pi)^3}\int \delta(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{r}}d^3\mathbf{k}, \tag{10.14}$$

where V is the volume of a comoving large box in which $\delta(\mathbf{r})$ is specified. The Fourier components can be found by the inverse Fourier transform

$$\delta(\mathbf{k}) = \frac{1}{V}\int \delta(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}}d^3\mathbf{r}. \tag{10.15}$$

The quantity $\mathbf{k}$ is the comoving wavenumber. For simplicity, we will write $\delta(\mathbf{k})$ as $\delta_{\mathbf{k}}$. Each Fourier component is a complex number which can be written as $\delta_{\mathbf{k}} = |\delta_{\mathbf{k}}|e^{i\phi_{\mathbf{k}}}$. If $|\delta_{\mathbf{k}}| \ll 1$, it obeys (10.12),

$$\ddot{\delta}_{\mathbf{k}} + 2H\dot{\delta}_{\mathbf{k}} - \frac{3}{2}\Omega_m H^2\delta_{\mathbf{k}} = 0. \tag{10.16}$$

This equation is valid as long as the proper wavelength $\lambda_k = a(t)2\pi/k$ is larger than the Jeans length $\lambda_{\rm J}$ and smaller than the Hubble distance c/H. The Fourier decomposition is useful because when perturbations with smaller wavelength reach $|\delta_{\mathbf{k}}| \sim 1$, the ones with longer wavelength can still be described by (10.16) separately. Physically, it means that we can study large structure perturbations even when there are smaller structures formed within them.

The **power spectrum** is defined as the mean square amplitude of the Fourier components:

$$P(k) = \langle|\delta_{\mathbf{k}}|^2\rangle. \tag{10.17}$$

The average is over all possible orientation of the wavenumber $\mathbf{k}$ with magnitude k. If the phases $\phi_{\mathbf{k}}$ of different Fourier components are uncorrelated with each other, then $\delta(\mathbf{r})$ is called a **Gaussian field**. If $\delta(\mathbf{r})$ is a Gaussian field, the probability $\mathcal{P}$ of selecting (randomly) a value of $\delta(\mathbf{r})$ is the Gaussian distribution

$$\mathcal{P}(\delta) = \frac{1}{\sqrt{2\pi}\sigma_\delta}e^{-\delta^2/2\sigma_\delta^2}.$$

The standard deviation is given by

$$\sigma_\delta^2 = \frac{V}{(2\pi)^3}\int P(k)\,d^3\mathbf{k} = \frac{V}{(2\pi)^3}\int_0^{2\pi}d\phi\int_0^{\pi}d\theta\int_0^{\infty}dk\,P(k)k^2\sin\theta = \frac{V}{2\pi^2}\int_0^{\infty}P(k)k^2\,dk,$$

where we used spherical coordinates in the k-space for integration. Most inflationary scenarios predict that the density fluctuations created by inflation is an isotropic, homogeneous Gaussian field. Also, the power spectrum of inflationary density fluctuations can be well described by a power law,

$$P(k) \propto k^n.$$

The inflation theory favors $n \approx 1$. The power spectrum with $n = 1$ is known as the **Harrison-Zel'dovich spectrum**. From observations of the temperature fluctuations of the CMB, $n = 0.97 \pm 0.01$.

For a universe with $P(k) \propto k^n$, consider a randomly picked sphere of radius r. The average mass of non-relativistic matter of each sphere is

$$\langle M \rangle = \frac{4\pi r^3 \rho_{m,0}}{3} = 1.67 \times 10^{11} M_\odot \left(\frac{r}{1\,\mathrm{Mpc}}\right)^3.$$

The actual mass of each sphere will vary. The mean square density fluctuation is given by

$$\left\langle \left(\frac{M - \langle M \rangle}{\langle M \rangle}\right)^2 \right\rangle = \frac{V}{2\pi^2} \int P(k) \left[\frac{3j_1(kr)}{kr}\right]^2 k^2\, dk = \frac{9V}{2\pi^2 r^2} \int P(k) j_1(kr)^2\, dk, \tag{10.18}$$

where $j_n(x)$ is the spherical Bessel function of the first kind, and $j_1(x) = (\sin x - x\cos x)/x^2$. Suppose the power spectrum has the form $P(k) \propto k^n$. By making a change of variable $u = kr$, $dk = du/r$,

$$\left\langle \left(\frac{M - \langle M \rangle}{\langle M \rangle}\right)^2 \right\rangle = \frac{9V}{2\pi^2 r^2} \int_0^\infty \left(\frac{u}{r}\right)^n j_1(u)^2 \frac{du}{r} = \frac{9V}{2\pi^2} r^{-3-n} \int_0^\infty u^n j_1(u)^2\, du.$$

The root-mean-square mass fluctuation in these spheres will have an r dependence

$$\frac{\delta M}{M} \equiv \left\langle \left(\frac{M - \langle M \rangle}{\langle M \rangle}\right)^2 \right\rangle^{1/2} \propto r^{-(3+n)/2} \propto M^{-(3-n)/6}. \tag{10.19}$$

The potential fluctuations satisfy

$$\delta\Phi \sim \frac{G\delta M}{r} \propto \frac{\delta M}{M^{1/3}} \propto M^{-(3+n)/6} M^{2/3} \propto M^{(1-n)/6}.$$

Note that only when $n = 1$ the potential fluctuations do not diverge on both large and small scales. Hence the Harrison-Zel'dovich spectrum is often referred to as a *scale invariant* power spectrum of density perturbations.

10.3.1 Hot versus Cold

During inflation, the power spectrum is $P(k) \propto k^n$, with $n \approx 1$. Between the end of inflation t_f and radiation-matter equality $t_{rm} \approx 4.7 \times 10^4$ yr, the shape of the power spectrum depends on whether the dark matter is cold or hot. Consider a particular time when the (matter) mass within a Hubble volume (a sphere of proper radius c/H) is equal to $M_{\mathrm{gal}} \approx 10^{12}\, M_\odot$. According to (7.18), the Hubble parameter in a radiation-dominated universe is

$$H(a) = H_0 \frac{\sqrt{\Omega_{r,0}}}{a^2}.$$

The mass within a Hubble volume in radiation era is

$$M_{\mathrm{H}} = \frac{4\pi}{3} \left(\frac{c}{H_0} \frac{a^2}{\sqrt{\Omega_{r,0}}}\right)^3 \frac{\rho_{m,0}}{a^3} = 1.6 \times 10^{28}\, M_\odot a^3.$$

Setting this equal to $M_{\mathrm{gal}} \approx 10^{12}\, M_\odot$ gives a scale factor $a \approx 4 \times 10^{-6}$. It corresponds t a temperature $kT \approx 60\,\mathrm{eV}$ and a cosmic time $t \approx 12\,\mathrm{yr}$. The mean energy of a relativistic particle is $3kT$ (instead of $\frac{3}{2}kT$ for non-relativistic particle according to the equipartition theorem, because the energy of a relativistic particle is linear in momentum p, not quadratic.) If a particle has a rest mass energy $mc^2 \ll 3kT$, then we say it is relativistic. If $mc^2 \gg 3kT$, then it is non-relativistic. **Hot dark matter** is relativistic at $t \approx 12\,\mathrm{yr}$. For example, neutrinos have mass $< 1\,\mathrm{eV}$, so it is definitely relativistic at $t \approx 12\,\mathrm{yr}$ when the mean energy is $3kT \approx 180\,\mathrm{eV}$. **Cold dark matter** is non-relativistic at $t \approx 12\,\mathrm{yr}$. For example, Weakly Interacting Massive Particles (WIMPs) also decouple at $t \sim 1\,\mathrm{s}$, but they are too massive ($mc^2 \sim 100\,\mathrm{GeV}$) to be relativistic at $kT \approx 180\,\mathrm{eV}$.

Consider a universe filled with hot dark matter consisting of particles with mass m_h. They become non-relativistic at a temperature

$3kT \approx m_h c^2$, or

$$T_h \approx \frac{m_h c^2}{3k} \approx 12000\,\mathrm{K}\left(\frac{m_h c^2}{3\,\mathrm{eV}}\right). \tag{10.20}$$

Since the temperature at radiation-matter equality is 9390 K, a particle with mass $m_h c^2 > 2.4\,\mathrm{eV}$ will become non-relativistic when the universe is still radiation-dominated. The temperature (10.20) corresponds to a cosmic time

$$t_h \approx 42000\,\mathrm{yr}\left(\frac{m_h c^2}{3\,\mathrm{eV}}\right)^{-2}.$$

Before time t_h, hot dark matter particles move near the speed of light in random motion called **free streaming**. It will smooth out any density fluctuations with wavelength smaller than ct_h in the hot dark matter smaller. In other words, no density fluctuations exist on scales smaller than

$$d_{\min} \approx ct_h \approx 13\,\mathrm{kpc}\left(\frac{m_h c^2}{3\,\mathrm{eV}}\right)^{-2} \quad\Longleftrightarrow\quad r_{\min} = \frac{d_{\min}}{a(t_h)} \approx \left(\frac{T_h}{2.7255\,\mathrm{K}}\right) d_{\min} \approx 55\,\mathrm{Mpc}\left(\frac{m_h c^2}{3\,\mathrm{eV}}\right)^{-1}.$$

The matter within a sphere of radius $r_{\min}$ has mass

$$M_{\min} = \frac{4\pi r_{\min}^3 \rho_{m,0}}{3} \approx 2.7 \times 10^{16}\,M_\odot \left(\frac{m_h c^3}{3\,\mathrm{eV}}\right)^{-3}.$$

We see that the smaller mass of hot dark matter particle, the larger density fluctuations would be smoothed out. For example, if $m_h c^2 \sim 50\,\mathrm{eV}$, there is no density fluctuations smaller than the Local Group. If $m_h c^2 \sim 3\,\mathrm{eV}$, there is no density fluctuations smaller than a supercluster. One can find that for wavenumbers $k \ll 2\pi/r_{\min}$, the power spectrum of hot dark matter is close to $P(k) \propto k$, the original power spectrum. Hot dark matter affects the power spectrum when the wavenumbers $k \gg 2\pi/r_{\min}$: $P(k)$ decays rapidly as a function of k. Because $k \propto r^{-1}$, this means only scales larger than $r_{\min}$ have growing density fluctuations. Smaller scale density fluctuations must wait for hot dark matter particles to become non-relativistic. In other words, superclusters form before smaller structures when the universe is filled with hot dark matter. This scenario is called the **top-down** scenario.

However, observations show that smaller structures form before larger structures. Galaxies in our universe are relatively older compared to superclusters, which are still undergoing collapse. This suggests that most of the dark matter in the universe is cold dark matter, with no free streaming. Recall that density fluctuations before t_{rm} do not grow significantly if their proper wavelength $\lambda_k = a(t)2\pi/k$ is small compared to the Hubble distance c/H. Only fluctuations with λ_k larger than the Hubble distance grow.

The Hubble distance at WIMPs decoupling $t_d \sim 1\,\mathrm{s}$ is $a_d \sim 3 \times 10^{-10}$, which corresponds to

$$\frac{c}{H} = 2ct_d \approx 6 \times 10^8\,\mathrm{m}, \quad r_d = \frac{2ct_d}{a_d} \approx 60\,\mathrm{pc}, \quad M_d = \frac{4\pi r_d^3 \rho_{m,0}}{3} \approx 0.05\,M_\odot, \quad \text{and} \quad k_d = \frac{2\pi}{r_d} \approx 10^5\,\mathrm{Mpc}^{-1}.$$

Hence density fluctuations with $k < k_d$ and $M > M_d$ can grow in amplitude, as long as their comoving wavelength remains larger than r_d. At $t_{rm} = 0.047\,\mathrm{Myr}$, $a_{rm} = 2.9 \times 10^{-4}$,

$$\frac{c}{H} \approx 1.8ct_{rm} \approx 0.027\,\mathrm{Mpc}, \quad r_{rm} = \frac{1.8ct_{rm}}{a_{rm}} \approx 90\,\mathrm{Mpc}, \quad M_{rm} \approx 1.3 \times 10^{17} M_\odot, \quad k_{rm} = \frac{2\pi}{r_{rm}} \approx 0.07\,\mathrm{Mpc}^{-1}.$$

Density fluctuations with $k < k_{rm}$ and $M > M_{rm}$ will grow steadily throughout the entire radiation era. Their power spectrum remains $P(k) \propto k$. Density fluctuations with $k_{rm} < k < k_d$ will grow in amplitude until their physical wavelength $a(t)r \propto t^{1/2}$ is smaller than the Hubble distance $c/H \propto t$. They will grow again after t_{rm} (matter dominated) according to the conclusion in the last section. Density fluctuations with $k > k_{rm}$ are suppressed throughout the entire radiation era. With this shape of power spectrum, we will find that $\delta M/M$ is largest in amplitude for smallest mass scales, consistent with observations. This scenario is called the **bottom-up** scenario.

In conclusion, the majority of dark matter in the universe is in the form of cold dark matter. This is why the current standard model of cosmology is called ΛCDM (Lambda Cold Dark Matter) model. It means the dominant energy density today is in the form of a cosmological constant Λ and cold dark matter.

Reionization optical depth:

$$\tau_* = \frac{2}{3\Omega_{m,0}} \frac{\Gamma_0}{H_0}\left[\sqrt{\Omega_{m,0}(1+z_*)^3 + \Omega_{\Lambda,0}} - 1\right], \quad \Gamma_0 = c\sigma_{\mathrm{T}} n_{\mathrm{bary},0}.$$

10.3.2 Baryon Acoustic Oscillations

To be written.

10.4 Baryons and Photons

The matter content in the universe is not solely made up of dark matter. There are also baryonic matter that can interact with photons. Baryonic structures form after photon decoupling $z_{\text{dec}} = 1090$. Some parts of the universe today ($t_0 \approx 13.7\,\text{Gyr}$) are far above average baryonic density. The average baryonic density today is

$$\rho_{\text{bary},0} = 4.2 \times 10^{-28}\,\text{kg/m}^3 = 6.2 \times 10^9\,M_\odot/\text{Mpc}^3.$$

In the region within a few hundred parsecs of the Sun, for example, the density of stars and interstellar gas is

$$\rho = 6.4 \times 10^{-21}\,\text{kg/m}^3 = 0.095\,M_\odot/\text{pc}^3, \quad \delta = \frac{\rho - \rho_{\text{bary},0}}{\rho_{\text{bary},0}} \sim 2 \times 10^7.$$

For individual stars like the Sun, $\bar{\rho}_\odot \approx 1400\,\text{kg/m}^3$, which corresponds to a density contrast $\delta_\odot \sim 3 \times 10^{30}$. Even though there are very dense objects today, most of the baryonic matter is still in the form of intergalactic gas.

10.4.1 Baryonic Matter Today

The current baryonic mass of the universe can be divided into several parts:

- Stars, stellar remnants, brown dwarfs, and planets: 7%.
- Interstellar gas: 1%.
- Circumgalactic gas that is gravitationally bound within the dark halo of a galaxy, but lying outside the main distribution of the galaxy's stars: 3%.
- Intracluster gas that is gravitationally bound within a cluster of galaxies, but is not bound to any particular galaxy within the cluster: 4%.
- Diffuse intergalactic gas at a temperature $T < 10^5\,\text{K}$ outside galaxies and clusters: 40%. It is called "diffuse" because it is less dense than the average baryonic density, $\delta \le 0$.
- Warm-hot intergalactic gas at a temperture $10^5 < T < 10^7\,\text{K}$ found in long filamnets between clusters: 45%. Typically it is denser than the average baryonic density, $3 < \delta < 300$.

It is found that most of the intergalactic gas in hydrogen are ionized, with ionization fraction x very close to 1. But we know that after the epoch of recombindation, the ionization fraction dropped close to zero. Something must have reionized hydrogen after the recombination.

10.4.2 Reionization of Hydrogen

Consider the scattering process between photons and free electrons. The cross-section is dominated by the Thomson scattering cross-section $\sigma_\text{T} = 6.65 \times 10^{-29}\,\text{m}^2$. The rate of photon scattering from a free electron is

$$\Gamma = \sigma_\text{T} n_e c,$$

where n_e is the number density of free electrons. Suppose the instant of reionization is at time t_*. The optical depth (expected number of scattering) from time t_* to t_0 is given by

$$\tau_* = \int_{t_*}^{t_0} \Gamma(t)\,dt = c\sigma_\text{T} \int_{t_*}^{t_0} n_e(t)\,dt.$$

The CMB photons can scatter from free electrons. The fact that we can see the CMB and deduce (right) physics at the time of last scattering means that the CMB photons should not scatter many times after t_*. Otherwise, the information from the CMB is smeared out from these scattering. Observations from the CMB temperature fluctuations shows that $\tau_* = 0.066 \pm 0.016$ for the reionized gas at low redshift. This means 1/15 of the CMB photons scatters from free electrons before reaching us.

Now we can make some estimate on the time t_*. Here are some assumptions:

1. Assume that the baryonic matter is a uniformly distributed.
2. Let baryonic matter of the universe be pure hydrogen, either in neutral atoms with number density n_{H}, or in free protons with number density n_p. Hence we have
$$n_{\mathrm{H}} + n_p = n_{\mathrm{bary}} = \frac{n_{\mathrm{bary},0}}{a^3}.$$
3. Assume that all hydrogen atoms is ionized at time t_* instantaneously. Then the number density of free electrons before t_* is $n_e = 0$, and that after t_* is $n_e = n_p = n_{\mathrm{bary},0}/a^3$.

With these assumptions, after the instant of reionization, the optical depth is

$$\tau_* = \Gamma_0 \int_{t_*}^{t_0} \frac{dt}{a(t)^3}, \quad \text{where} \quad \Gamma_0 = c\sigma_{\mathrm{T}} n_{\mathrm{bary},0} = 1.58 \times 10^{-4}\,\mathrm{Gyr}^{-1} \approx 0.0023 H_0$$

is the scattering rate of photons and free electrons today. Making a change of variable from t to a, and then a to z (by $1+z = 1/a$),

$$\tau_* = \Gamma_0 \int_{a(t_*)}^{1} \frac{da}{\dot{a}a^3} = \Gamma_0 \int_{a(t_*)}^{1} \frac{da}{H(a)a^4} = \Gamma_0 \int_0^{z_*} \frac{(1+z)^2\,dz}{H(z)}.$$

In recent times, the universe is matter-Λ dominated, with Hubble parameter (7.18)

$$H(z) = H_0[\Omega_{m,0}(1+z)^3 + \Omega_{\Lambda,0}]^{1/2}.$$

The integral has an analytic solution

$$\tau_* = \frac{2}{3\Omega_{m,0}} \frac{\Gamma_0}{H_0} \left[\sqrt{\Omega_{m,0}(1+z_*)^3 + \Omega_{\Lambda,0}} - 1 \right].$$

Plugging in $\Omega_{m,0} = 0.31$, $\Omega_{\Lambda,0} = 0.69$, $\Gamma_0 \approx 0.0023 H_0$,

$$\tau_* = 0.00485 \left[\sqrt{0.31(1+z_*)^3 + 0.69} - 1 \right].$$

Setting this equal to the observed value from the CMB $\tau_* = 0.066 \pm 0.016$ gives the redshift of reionization $z_* = 7.8 \pm 1.3$. This corresponds to the age of the universe $t_* \sim 650\,\mathrm{Myr}$. The time elapse between recombination and reionization is $t_* - t_{\mathrm{rec}} \sim 0.05 t_0$.

We have the time of reionization. Let's discuss how hydrogen can be ionized. Basically, we are looking for sources of photons with energy $h\nu > 13.6\,\mathrm{eV}$. One source of these photons is an O-type star with $M \geq 30\,M_\odot$. The production of ionizing photons is $\dot{N}_* \approx 5 \times 10^{48}\,\mathrm{s}^{-1}$. In its entire life $t \approx 6\,\mathrm{Myr} \approx 2 \times 10^{14}\,\mathrm{s}$, the total ionizing photons produced is $N_* \approx 10^{63}$. Another source is an **active galactic nucleus** (AGN). An AGN is a supermassive black hole that is capable of radiating ionizing photons from its accretion disk. It is located at the center of a galxy. A luminous AGN has a rate of ionizing photon production of

$$\dot{N}_* \approx 3 \times 10^{56}\,s^{-1} \left(\frac{L_{\mathrm{AGN}}}{10^{13}\,L_\odot} \right).$$

The most luminous AGNs are known as **quasars**. It can have a luminosity $L \sim 10^{13}\,L_\odot$, much more luminous than an O-type star.

10.5 Galaxies and Stars

10.5.1 The First Stars and Quasars

To be written.

10.5.2 Making Galaxies

Stars are not uniformly distributed in the universe. They are clustered in galaxies with interstellar gas and a dark matter halo. The number density of galaxies in the luminosity range $L \to L + dL$ is the **Schechter luminosity function**

$$\Phi(L)\, dL = \Phi^* \left(\frac{L}{L^*}\right)^\alpha e^{-L/L^*} \frac{dL}{L^*}. \tag{10.21}$$

The observed luminosity function has a power-law $\alpha \approx -1$ and $\Phi^* \approx 0.005\,\mathrm{Mpc}^{-3}$, and $L_V^* \approx 2 \times 10^{10}\, L_{\odot,V}$ in the V band. The Milky Way galaxy has $L_V \approx L_V^* \approx 2 \times 10^{10}\, L_{\odot,V}$ and a baryonic mass $M_{\text{bary}} \approx 1.2 \times 10^{11}\, M_\odot$. The total mass with the dark halo is uncertain, $M_{\text{tot}} \approx 1 - 2 \times 10^{12}\, M_\odot$. From (10.21), it is clear that galaxies with $L > L^*$ are rare, so making galaxies with total mass $M_{\text{tot}} > 10^{12}\, M_\odot$ is difficult. Basically, it means there is an upper limit on the mass of a galaxy.

Consider an overdense sphere of radius R at radiation-matter equality $t_{rm} \approx 0.047\,\mathrm{Myr}$. The relation between the mass M inside the sphere, density contrast δ, and radius R is

$$M = \frac{4\pi R(t)^3 \rho_m(t)}{3}(1+\delta) \quad \text{where} \quad \rho_m(t) = \rho_{m,0} a^{-3} = \rho_{m,0}(1+z)^3.$$

The perturbation requires $\delta(t) \ll 1$, while a collapse under its own gravity starts when $\delta \approx 1$. Let t_{coll} be the time of gravitational collapse of the sphere. From (10.13), we know that

$$\delta \propto t^{2/3} \propto a$$

in the matter-dominated era. Then the collapse time t_{coll} can be approximated as

$$t_{\text{coll}} \approx \delta_{rm}^{-3/2} t_{rm} \quad \text{and} \quad 1 + z_{\text{coll}} \approx \delta_{rm}(1 + z_{rm}).$$

This means a region with larger δ_{rm} will collapse earlier and have a higher redshift. At t_{coll}, the density of the sphere is

$$\rho \approx 2\rho_m(t_{\text{coll}}) = 2\rho_{m,0}(1 + z_{\text{coll}}).$$

The galaxy forms when the sphere collapse to a gravitationally bound halo with radius R_{halo}. It is in virial equilibrium with $E_k = \frac{1}{2}E_p$. Suppose the initial energy before the collapse is entirely in the form of potential energy $E_{p,rm}$ Then we have

$$E_{k,\text{coll}} + E_{p,\text{coll}} = \frac{1}{2}E_{p,\text{coll}} = E_{p,rm}.$$

The total potential energy within a uniform sphere can be calculated as $-3GM^2/5R$. The key here is that $E_p \propto 1/R$, so $E_{p,\text{coll}}/2 = E_{p,rm}$ means that the galaxy has $R_{\text{halo}} = R_{rm}/2$. Its density will be 8 times larger,

$$\rho_{\text{halo}} = 8\rho(t_{\text{coll}}) = 16\rho_{m,0}(1 + z_{\text{coll}})^3.$$

If we know ρ_{halo}, we can find out when a sphere starts to collapse to form a galaxy.

Apart from the collapse time, we can also find the temperature of particles in a galaxy. Suppose the galaxy is also in hydrostatic equilibrium (2.1) and the gas within the galaxy is ideal,

$$\frac{dP_{\text{gas}}}{dr} = -\frac{GM(r)\rho_{\text{gas}}(r)}{r^2}, \quad P_{\text{gas}} = \frac{\rho_{\text{gas}} kT_{\text{gas}}}{\bar{m}}.$$

Combining the two equations above, the mass function $M(r)$ can be written as

$$M(r) = -\frac{dP}{dr}\frac{r^2}{G\rho_{\text{gas}}} = -\frac{d\rho_{\text{gas}}}{dr}\frac{kTr^2}{G\bar{m}} - \frac{dT}{dr}\frac{kr^2}{G\bar{m}} = -\frac{kT_{\text{gas}}r}{G\bar{m}}\left[\frac{d\ln\rho_{\text{gas}}}{d\ln r} + \frac{d\ln T_{\text{gas}}}{d\ln r}\right]. \tag{10.22}$$

The mean particle mass incorporates all baryons and electrons. With helium mass fraction $Y = 0.24$, we can find that $\bar{m} = 0.6m_p$. Assume that T_{gas} is uniform throughout the whole galaxy and the density obeys a power law $\rho_{\text{gas}} \propto r^{-\beta}$ ($\beta > 0$). Evaluating the mass at $r = R_{\text{halo}}$ so $M(r) = M_{\text{tot}}$, Equation 10.22 says

$$kT_{\text{gas}} = \frac{GM_{\text{tot}}\bar{m}}{\beta R_{\text{halo}}}. \tag{10.23}$$

This temperature is called the **virial temperature** of the virialized halo. Expressing the virial temperature in terms of $\rho_{m,0}$,

$$R_{\text{halo}} = \left(\frac{3M_{\text{tot}}}{4\pi\rho_{\text{halo}}}\right)^{1/3} = \left[\frac{3M_{\text{tot}}}{64\pi\rho_{m,0}(1+z_{\text{coll}})^3}\right]^{1/3}, \qquad kT_{\text{gas}} = \frac{4}{\beta}\left(\frac{\pi M_{\text{tot}}^2\rho_{m,0}}{3}\right)^{1/3} G\bar{m}(1+z_{\text{coll}}).$$

Hence massive galaxies have high virial temperature. We can make further assumptions: let the gas have similar profile to the dark halo (observed to have $\beta \approx 2$) and $\bar{m} = 0.6m_p$. The virial temperature of the gas in the galaxy is

$$T_{\text{gas}} \approx 10^6\,\text{K}\left(\frac{M_{\text{tot}}}{10^{12}\,M_\odot}\right)^{2/3}\left(\frac{1+z_{\text{coll}}}{5}\right).$$

(The scaling of the redshift is because $z_{\text{coll}} = 4$ is the intense star-forming era.)

The reason why galaxies with $M_{\text{tot}} > 10^{12}\,M_\odot$ again comes from the virial theorem. The hot baryonic gas emits photons. As the gas loses energy, it cool down and contracts. But a more massive dark halo will have a gas that is less efficient in radiating away energy. In a low-mass halo, the virial temperature is lower: e.g. a $10^{10}\,M_{\text{halo}}$ has $T_{\text{gas}} \approx 70000\,\text{K}$. At such a temperature, hydrogen is ionized, but helium is not completely ionized—it can still have one bound electrons. The bound electrons can radiate efficiently by spontaneous emission. At even lower temperatures, spontaneous emission occur in both neutral helium and hydrogen. On the contrary, a massive halo with $M_{\text{tot}} \sim 10^{12}\,M_\odot$ and $T_{\text{gas}} > 10^6\,\text{K}$ has a fully ionized gas. Electrons radiate via bremsstrahlung (Free-free emission), essentially by acceleration when they are near protons.

The luminosity density for bremsstrahlung emission is found to be

$$\Psi = 5.3\times 10^{-32}\,\text{W/m}^3\left(\frac{\rho_{\text{gas}}}{10^{-24}\,\text{kg/m}^3}\right)^2\left(\frac{T}{10^6\,\text{K}}\right)^{1/2}.$$

The low-density gas is highly transparent, so all the luminosity escapes from the halo. The thermal energy density needed to radiate away is

$$\epsilon = \frac{3}{2}nkT = 2.1\times 10^{-14}\,\text{J/m}^3\left(\frac{\rho_{\text{gas}}}{10^{-24}\,\text{kg/m}^3}\right)\left(\frac{T}{10^6\,\text{K}}\right).$$

Thus, the time for the ionized gas to cool by bremstrahlung emission is

$$t_{\text{cool}} = \frac{\epsilon}{\Psi} = 13\,\text{Gyr}\left(\frac{\rho_{\text{gas}}}{10^{-24}}\right)^{-1}\left(\frac{T}{10^6\,\text{K}}\right).$$

If the gas has a temperature $T > 10^6\,\text{K}$, it needs to have a density $\rho_{\text{gas}} > 10^{-24}\,\text{kg/m}^3$ to cool within the age of the universe. We can show that it is unlikely. Suppose the ratio of the baryonic gas to the total mass in the halo is the same as that of the universe, $f = 0.048/0.31 = 0.15$. The average baryonic mass density is

$$\rho_{\text{bary}} = f\rho_{\text{halo}} = 16f\rho_{m,0}(1+z_{\text{coll}})^3 \approx 0.8\times 10^{-24}\,\text{kg/m}^3\left(\frac{f}{0.15}\right)\left(\frac{1+z_{\text{coll}}}{5}\right)^3.$$

In conclusion, a virialized halo with $M_{\text{tot}} > 10^{12}\,M_\odot$ that starts to collapse at $z_{\text{coll}} > 4$ will not have enough time to cool to form galaxies.

10.5.3 Making Stars

To be written.

Part IV

Appendix

A Statistical Mechanics

A.1 Definitions and Theorems

A.1.1 The Equipartition Theorem

The **equipartition theorem** says that every quadratic degrees of freedom of particles contributes to the average internal energy $\frac{1}{2}kT$ in a system. For example, the translational kinetic energy is $\frac{p^2}{2m}$, quadratic in p_x, p_y, and p_z, so the average translational kinetic energy of particles is $\frac{3}{2}kT$. Other examples could be rotational kinetic energy $\frac{L^2}{2I}$, vibrational kinetic energy $\frac{1}{2}mv^2$, elastic potential energy $\frac{1}{2}kx^2$, etc.

A.1.2 Chemical Potential

The **chemical potential** is a measure of how much energy is needed to add a particle to a system. It is defined as

$$\mu_i = \left(\frac{\partial U}{\partial N_i}\right)_{S,V} = \left(\frac{\partial G}{\partial N_i}\right)_{T,P},$$

where U is the total internal energy, G is the Gibbs free energy, N_i is the number of particles of ith species, V is the volume, S is the entropy, T is the temperature, and P is the pressure. If two systems are in diffusive equilibrium, we say that they are at the same chemical potential. (Think of thermal equilibrium, where the two system have the same temperature). The chemical potential is divided into an internal part and an external part. The external part is the usual potential energy per particle.

A.2 Density of States of an Ideal Gas

The density of states of an ideal gas can be derived from infinite square well, or particle-in-a-box case. The 1D time-independent Schrödinger equation is

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + V(x)\psi(x) = E\psi(x).$$

For the infinite square well, the potential V looks like

$$V(x) = \begin{cases} 0, & 0 < x < L, \\ \infty, & \text{otherwise.} \end{cases}$$

Outside the well, the probability of finding the particle is zero so $\psi(x) = 0$. Inside the well, we have $V(x) = 0$ and the time-independent Schrödinger equation becomes

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} = k^n\psi, \quad \text{where} \quad k = \frac{\sqrt{2mE}}{\hbar}.$$

The solution of this equation with boundary condition $\psi(0) = \psi(L) = 0$ is

$$\psi(x) = \sqrt{\frac{2}{L}}\sin\left(\frac{n\pi x}{L}\right), \quad n = 1, 2, 3, \ldots$$

The possible values of energy are

$$E_n = \frac{\hbar^2 k_n^2}{2m} = \frac{n^2\pi^2\hbar^2}{2ma^2}.$$

Generalizing to 3D infinite square well of length L, or volume $V = L^3$, the wavenumber of the particle is

$$\mathbf{k} = (k_x, k_y, k_z) = \frac{\pi}{L}(n_x, n_y, n_z),$$

where n_x, n_y, n_z are positive integers. The momentum of the particle is

$$\mathbf{p} = \hbar\mathbf{k} = \frac{\hbar\pi}{L}(n_x, n_y, n_z) = \frac{h}{2L}(n_x, n_y, n_z).$$

This is saying that each state can occupy a momentum volume of $(h/2L)^3$, each assigned with different quantum numbers. We want to know, for a given momentum interval $p \to p + dp$, how many states there are. In momentum space, the momentum vector can point anywhere on the surface of constant p, but in the first octant because $n_x, n_y, n_z > 0$. The interval will have a volume $4\pi p^2 dp$, so the number of states within will be

$$\frac{1}{8}\frac{4\pi p^2 dp}{(h/2L)^3} = \frac{L^3}{h^3}4\pi p^2 dp.$$

Since L^3 is the spatial volume of the box, the density of states in the box is

$$g(p)\,dp = \frac{4\pi p^2 dp}{h^3}.$$

A.3 Occupation Number/Occupancy

For quantum statistics, the idea is to consider a system of one single particle state instead of a particle. In other words, the system is a particular spatial wavefunction with a particular spin orientation. The reservoir consists of all the other possible single-particle states. Let the energy of a single-particle state occupied by one particle to be ϵ. Then when the state is unoccupied, the energy is 0, and when the state is occupied by n particles, the energy is $n\epsilon$. The probability of having n particles in this state is

$$\mathcal{P}(n) = \frac{1}{\mathcal{Z}}e^{-(n\epsilon-\mu n)/kT} = \frac{1}{\mathcal{Z}}e^{-n(\epsilon-\mu)/kT}.$$

If the particles are fermions, then n can only be 0 or 1 because no more than one fermions can occupy the same state. The grand partition function for fermions is thus

$$\mathcal{Z} = 1 + e^{-(\epsilon-\mu)/kT}.$$

The average number of particles in that state is called **occupancy** of the state:

$$\bar{n} = \sum_n n\mathcal{P}(n) = 0 \cdot \mathcal{P}(0) + 1 \cdot \mathcal{P}_1 = \frac{e^{-(\epsilon-\mu)/kT}}{1 + e^{-(\epsilon-\mu)/kT}} = \frac{1}{e^{(\epsilon-\mu)/kT} + 1}.$$

For fermions, the occupancy of a state is just the probability of the state being occupied. This formula is called the **Fermi-Dirac distribution**:

$$\bar{n}_{\mathrm{FD}} = \frac{1}{e^{(\epsilon-\mu)/kT} + 1}. \tag{A.1}$$

It is a function of ϵ, μ, and T. Note that $\bar{n}_{\mathrm{FD}}$ goes to zero when $\epsilon \gg \mu$ and goes to 1 when $\epsilon \ll \mu$. Hence a state with energy much smaller than μ tends to be occupied.

If the particles are bosons, n can be any non-negative integer. The grand partition function is

$$\mathcal{Z} = 1 + e^{-(\epsilon-\mu)/kT} + e^{-2(\epsilon-\mu)/kT} + \cdots = \sum_n^{\infty}\left[e^{-(\epsilon-\mu)/kT}\right]^n = \frac{1}{1 - e^{-(\epsilon-\mu)/kT}}.$$

The Gibbs factor cannot grows to infinity, so we require that $\mu < \epsilon$ and the series will converge. The average number of particles in a state is

$$\bar{n} = \sum_n n\mathcal{P}(n) = \sum_n n\frac{e^{-nx}}{\mathcal{Z}} = -\frac{1}{\mathcal{Z}}\sum_n \frac{\partial}{\partial x}e^{-nx} = -\frac{1}{\mathcal{Z}}\frac{\partial \mathcal{Z}}{\partial x},$$

where $x \equiv (\epsilon - \mu)/kT$ (this formula also works for fermions). So the bosons have occupancy

$$\bar{n} = -(1 - e^{-x})\frac{\partial}{\partial x}\frac{1}{1 - e^{-x}} = \frac{1}{e^{(\epsilon-\mu)/kT} - 1}.$$

This is called the **Bose-Einstein distribution**:

$$\bar{n}_{\mathrm{BE}} = \frac{1}{e^{(\epsilon-\mu)/kT} - 1}. \tag{A.2}$$

A.4 Maxwell-Boltzmann Distribution

A.4.1 Speed Distribution

From the equipartition theorem, we already know that the root-mean-square speed of particles in an ideal gas is

$$v_{\text{rms}} = \sqrt{\frac{3kT}{m}}.$$

As its name suggests, this is just some average speed. The distribution of speeds is unknown from this expression. We will use Boltzmann statistics to derive the speed distribution, the probability density of some particle moving at a given speed, of an ideal gas.

The distribution function $\mathcal{D}(v)$, like any other probability density, is normalized. The probability of finding the particle's speed between v_1 and v_2 is

$$\mathcal{P}(\text{between } v_1 \text{ and } v_2) = \int_{v_1}^{v_2} \mathcal{D}(v)\, dv.$$

The space is three dimensional, so for a given speed, there can be many possible velocity vectors. This means

$$\mathcal{D}(v) \propto \begin{pmatrix}\text{probability of a particle} \\ \text{having velocity } \mathbf{v}\end{pmatrix} \times \begin{pmatrix}\text{number of vectors } \mathbf{v} \\ \text{corresponding to speed } v\end{pmatrix}.$$

We know that the velocity is related to kinetic energy by $E = \frac{1}{2}mv^2$. Hence the probability of having $\mathbf{v}$ is proportional to the Boltzmann factor $e^{-mv^2/2kT}$. For the second factor, consider a three-dimensional velocity space. Any velocity vector $\mathbf{v}$ corresponding to speed v lies on the surface of a sphere with radius v in this space. We see that a large v has more possible velocity vectors because it corresponds to a larger sphere. Thus, the number of $\mathbf{v}$ corresponding to speed v is proportional to $4\pi v^2$. Combining the two factors, we have

$$\mathcal{D}(v) = C \cdot 4\pi v^2 e^{-mv^2/2kT}.$$

To determine C, use the normalization condition for a distribution,

$$1 = \int_0^\infty \mathcal{D}(v)\, dv = 4\pi C \int_0^\infty v^2 e^{-mv^2/2kT}\, dv.$$

Making a change of variables $x = v\sqrt{m/2kT}$, this equation becomes

$$1 = 4\pi C \left(\frac{2kT}{m}\right)^{3/2} \int_0^\infty x^2 e^{-x^2}\, dx = 4\pi C \left(\frac{2kT}{m}\right)^{3/2} \frac{\sqrt{\pi}}{4}.$$

This gives $C = (m/2\pi kT)^{3/2}$. The final result for the distribution function is

$$\mathcal{D}(v) = \left(\frac{m}{2\pi kT}\right)^{3/2} 4\pi v^2 e^{-mv^2/2kT}. \tag{A.3}$$

This is known as the **Maxwell speed distribution** for an ideal gas.

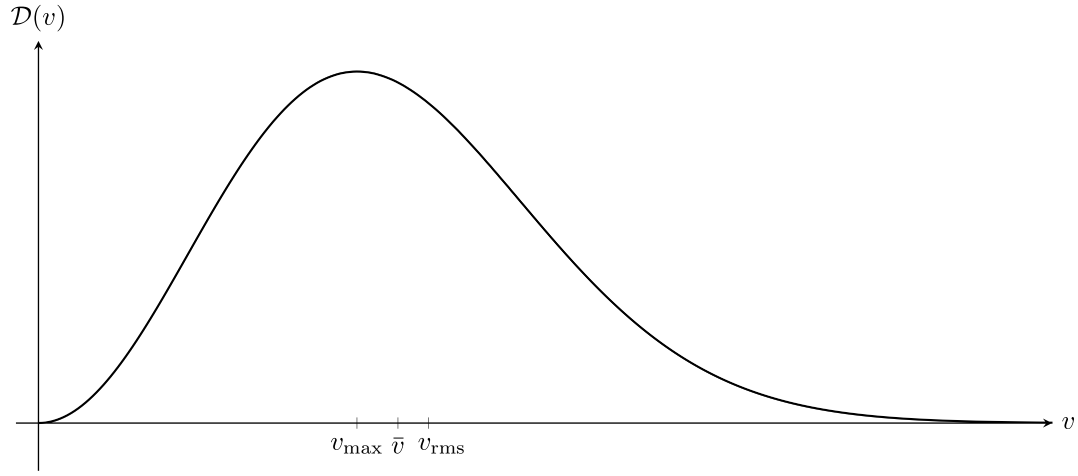


Figure A.1: The Maxwell speed distribution falls off as $v \to 0$ and as $v \to \infty$. The average speed is slightly larger than the most likely speed, while the rms speed is a even more larger.

At small v, the probability looks like a parabola. At large v, the probability is exponentially decaying. There are two more speeds we can drawn from this graph. One is the most probable speed $v_{\max}$. This is obtained by taking the derivative of the distribution:

$$\frac{d\mathcal{D}}{dv} \propto \frac{d}{dv}\left(v^2 e^{-mv^2/2kT}\right) = 2ve^{-mv^2/2kT} - v^2\left(\frac{mv}{kT}\right)e^{-mv^2/2kT} = 2v\left(1 - \frac{mv^2}{2kT}\right)e^{-mv^2/2kT}.$$

Setting this equal to zero gives $v = 0$ or $v^2 = 2kT/m$. At $v = 0$, the distribution is at minimum, so the most probable speed is at $v_{\max} = \sqrt{2kT/m}$.

The average speed $\bar{v}$ is determined by the following expression:

$$\bar{v} = \int_0^\infty v\mathcal{D}(v)\,dv = \left(\frac{m}{2\pi kT}\right)^{3/2} 4\pi \int_0^\infty v^3 e^{-mv^2/2kT}\,dv.$$

To do this integral analytically, you can first evaluate the integral of $ve^{-\lambda v^2}$, where $\lambda = -m/2kT$,

$$\int_0^\infty ve^{-\lambda v^2}\,dv = \frac{1}{2}\int_0^\infty e^{-\lambda v^2}\,d(v^2) = -\frac{1}{2\lambda}e^{-\lambda v^2}\bigg|_0^\infty = -\frac{1}{2\lambda}.$$

Taking the derivative with respect to λ gives the desired integral

$$\int_0^\infty v^3 e^{-\lambda v^2}\,dv = -\frac{\partial}{\partial\lambda}\int_0^\infty ve^{-\lambda v^2}\,dv = -\frac{\partial}{\partial\lambda}\left(\frac{1}{2\lambda}\right) = \frac{1}{2\lambda^2}.$$

Plugging in $\lambda = -m/2kT$, average speed is

$$\bar{v} = \left(\frac{m}{2\pi kT}\right)^{3/2} 4\pi \int_0^\infty v^3 e^{-mv^2/2kT}\,dv = \left(\frac{m}{2\pi kT}\right)^{3/2} 4\pi \cdot \frac{1}{2}\left(\frac{2kT}{m}\right)^2 = \sqrt{\frac{8kT}{\pi m}}.$$

Note that $v_{\max} < \bar{v} < v_{\rm rms}$. You may wonder why the average speed is not $v_{\rm rms}$. This is because the root-mean-square speed is defined (or related) to the average of speed squared. To see this, let's calculate the average of v^2 and see what we get:

$$\begin{aligned}\langle v^2\rangle &= \int_0^\infty v^2\mathcal{D}(v)\,dv = \frac{4}{\sqrt{\pi}}\left(\frac{m}{2kT}^{3/2}\right)\int_0^\infty v^4 e^{-mv^2/2kT}\,dv \\ &= \frac{4}{\sqrt{\pi}}\left(\frac{m}{2kT}\right)^{3/2}\left(\frac{2kT}{m}\right)^{5/2}\int_0^\infty x^4 e^{-x^2}\,dx = \frac{8kT}{\sqrt{\pi}m}\int_0^\infty x^4 e^{-x^2}\,dx \\ &= \frac{8kT}{\sqrt{\pi}m}\cdot\frac{3\sqrt{\pi}}{8} = \frac{3kT}{m}.\end{aligned}$$

This is consistent with the equipartition theorem.

A.4.2 Energy Distribution

It is useful to convert the speed distribution to energy distribution, in case we want to know the most probable energy. The energy and speed are related by $E = \frac{1}{2}mv^2$, or $v = \sqrt{2E/m}$. In deriving speed distribution, we have

$$\mathcal{D}(v) = C \cdot 4\pi v^2 e^{-mv^2/2kT}.$$

Substituting $v = \sqrt{2E/m}$ and $C = (m/2\pi kT)^{3/2}$ into the normalization condition,

$$1 = 4\pi C \int_0^\infty v^2 e^{-mv^2/2kT}\, dV = 4\pi C \int_0^\infty \left(\frac{2E}{m}\right)\sqrt{\frac{1}{2mE}} e^{-E/kT}\, dE = 2\left(\frac{1}{kT}\right)^{3/2} \int_0^\infty \sqrt{\frac{E}{\pi}} e^{-E/kT}\, dE.$$

Essentially, this is saying that the energy distribution is

$$\mathcal{D}(E) = 2\sqrt{\frac{E}{\pi}}\left(\frac{1}{kT}\right)^{3/2} e^{-E/kT} = \frac{2\pi\sqrt{E}}{(\pi kT)^{3/2}} e^{-E/kT}. \tag{A.4}$$

A.5 Ideal Gas Law for Both Relativistic and Non-Relativistic Gas

Classical ideal gases obey Maxwell-Boltzmann distribution with the occupation number given by

$$f(p) = e^{-[\epsilon(p)-\mu]/kT}.$$

The number density and the pressure of the gas are the integrals

$$\begin{aligned} n &= \int g_s g(p) f(p)\, dp = \frac{4\pi g_s}{h^3} e^{\mu/kT} \int_0^\infty p^2 e^{-\epsilon(p)/kT} dp \\ P &= \frac{1}{3}\int p v g_s g(p) f(p)\, dp = \frac{4\pi g_s}{3h^3} e^{\mu/kT} \int_0^\infty p^3 v(p) e^{-\epsilon(p)/kT} dp. \end{aligned}$$

The ratio P/n is then

$$P/n = \frac{\int p^3 v(p) e^{-\epsilon(p)/kT} dp}{3\int p^2 e^{-\epsilon(p)/kT} dp}.$$

The non-relativistic limit has energy per particle $p^2/2m + mc^2$, and velocity distribution $v(p) = p/m$. We will use the Gaussian integral (for reference): $\int e^{-\lambda x^2} dx = \sqrt{\pi/\lambda}$. Defining $\lambda = 1/2mkT$, we have (of by a factor of $\frac{4\pi g_s}{h^3} e^{(\mu - mc^2)/kT}$ for both number density and pressure):

$$\begin{aligned} n &\sim \int_0^\infty p^2 e^{-\lambda p^2} dp = -\frac{1}{2}\frac{d}{d\lambda}\int_{-\infty}^{+\infty} e^{-\lambda p^2} dp = -\frac{1}{2}\frac{d}{d\lambda}\sqrt{\frac{\pi}{\lambda}} = \frac{1}{4}\sqrt{\frac{\pi}{\lambda}}. \\ P &\sim \frac{1}{3}\int_0^\infty \frac{p^4}{m} e^{-\lambda p^2} dp = -\frac{1}{3m}\frac{dn}{d\lambda} = -\frac{1}{3m}\frac{d}{d\lambda}\left(\frac{1}{4}\sqrt{\frac{\pi}{\lambda^3}}\right) = \frac{1}{8}\sqrt{\frac{\pi}{\lambda^5}}. \end{aligned}$$

This gives

$$P/n = \frac{\sqrt{\pi/\lambda^5}/8m}{\sqrt{\pi/\lambda^3}/4} = \frac{1}{2m\lambda} = \frac{2mkT}{2m} = kT \implies \boxed{P = nkT.}$$

In the relativistic limit, $\epsilon(p) = pc$, and $v \approx c$. Let $\beta = c/kT$, the number density and pressure are

$$n \sim \int_0^\infty p^2 e^{-\beta p} dp = \frac{d^2}{d\beta^2}\int_0^\infty e^{-\beta p} dp = \frac{d^2}{d\beta^2}\frac{1}{\beta} = \frac{2}{\beta^3}.$$

$$P \sim \frac{1}{3}\int_0^\infty p^3 c e^{-\beta p} dp = -\frac{c}{3}\frac{d^3}{d\beta^3}\int_0^\infty e^{-\beta p} dp = -\frac{c}{3}\left(-\frac{6}{\beta^4}\right) = \frac{2c}{\beta^4}.$$

$$P/n = \frac{2c/\beta^4}{2/\beta^3} = \frac{c}{\beta} = \frac{c}{c/kT} = kT \implies \boxed{P = nkT.}$$

We see that the usual ideal gas law works for both classical relativistic and non-relativistic gases.

A.6 Degenerate Fermi Gas and Degeneracy Pressure

Pretend the temperature is at $T = 0$. At $T = 0$, the Fermi-Dirac distribution becomes a step function:

$$\bar{n} = \frac{1}{e^{(\epsilon-\mu)/kT} + 1} \qquad \longrightarrow \qquad \bar{n} = \begin{cases} 1, & \epsilon < \mu \\ 0, & \epsilon > \mu. \end{cases}$$

In this case we define μ as the **Fermi energy**, ϵ_{F}:

$$\epsilon_{\mathrm{F}} \equiv \mu(T = 0).$$

When a Fermi gas is so cold that nearly all states with $\epsilon < \epsilon_{\mathrm{F}}$ are occupied and nearly all states with $\epsilon > \epsilon_{\mathrm{F}}$ are unoccupied, it is said to be **degenerate**. (This name is totally unrelated to the one in quantum mechanics.) The value of ϵ_{F} is determined by the number of electrons in the gas. Suppose you are adding electrons to a box one at a time. The electrons will fill up the energy level due to Pauli exclusion principle, and the last electron will have energy just below ϵ_{F}. To add one more electron, you need to put in the energy $\epsilon_{\mathrm{F}} = \mu$. This is perfectly consistent with the definition $\mu = (\partial U/\partial N)_{S,V}$.

Now consider free electrons (electrostatic force ignored) inside a box of $V = L^3$. According to quantum mechanics in 3D, electrons can have momentum

$$p_x = \frac{hn_x}{2L}, \quad p_y = \frac{hn_y}{2L}, \quad p_z = \frac{hn_z}{2L}, \quad \text{where} \quad n_x, n_y, n_z = 1, 2, 3, \ldots$$

and energy levels

$$\epsilon = \frac{p^2}{2m} = \frac{h^2}{2mL^2}(n_x^2 + n_y^2 + n_z^2).$$

We can specify the allowed energies in the "n-space", where the three axes are labeled n_x, n_y, and n_z, and each state is represented by the vector $\mathbf{n} = (n_x, n_y, n_z)$. Note that in this space, the energy is proportional to the square of the distance from the origin, $n_x^2 + n_y^2 + n_z^2$. When we add electrons into the box, we are filling up the n-space starting from the center, until all energy levels are filled. All states live within the first octant of a sphere with radius $n_{\max}$, which specifies highest energy states. The Fermi energy is thus

$$\epsilon_{\mathrm{F}} = \frac{h^2 n_{\max}^2}{8mL^2}.$$

The total number of states within the box is the volume of 1/8 (first octant) of the sphere, times 2 (this accounts for two spin orientations of electrons), so

$$N = 2 \times (\text{volume of eighth-sphere}) = 2 \cdot \frac{1}{8} \cdot \frac{4}{3}\pi n_{\max}^3 = \frac{\pi n_{\max}^3}{3}.$$

Note that the Fermi energy depends only on the number density of electrons N/V, so it is an intensive quantity. These two equations give the Fermi energy of the gas as a function of total electrons N and the volume $V = L^3$,

$$\epsilon_{\mathrm{F}} = \frac{h^2}{8m}\left(\frac{3N}{\pi V}\right)^{2/3}. \tag{A.5}$$

The total energy of the system is just the triple sum, which can be approximated as an integral,

$$\begin{aligned} U &= 2\sum_{n_x,n_y,n_z} \epsilon(\mathbf{n}) \approx 2\int \epsilon(\mathbf{n})\, dn_x\, dn_y\, dn_z \\ &= 2\int_0^{n_{\max}} dn \int_0^{\pi/2} d\theta \int_0^{\pi/2} d\phi\, n^2 \epsilon(n) \sin\theta \\ &= \pi \int_0^{n_{\max}} \epsilon(n) n^2\, dn = \frac{\pi h^2}{8mL^2}\int_0^{n_{\max}} n^4\, dn = \frac{\pi h^2 n_{\max}^5}{40mL^2}. \end{aligned}$$

We use spherical coordinates in n-space and multiplying by 2 accounts for the spin orientations of electrons. The final answer can be written in terms of the Fermi energy and total number of electrons:

$$U = \frac{3}{5}N\epsilon_{\mathrm{F}} \tag{A.6}$$

This means the average energy of electrons is 3/5 the Fermi energy. Though the above analysis is based on electrons in a box, the result works for any Fermi gas in any volume at $T = 0$. Plugging in some numbers, the Fermi energy is $\epsilon_{\mathrm{F}} \sim \mathrm{eV}$, while the average thermal energy of a particle at room temperature is about $kT \approx 1/40\,\text{eV}$. Thus there are two equivalent statements:

$$\frac{V}{N} \ll v_Q \quad \iff \quad kT \ll \epsilon_\text{F}.$$

The approximation $T \approx 0$ is quite accurate in many situations. The temperature that a Fermi gas would have such that $kT = \epsilon_\text{F}$ is called the **Fermi temperature**: $T_\text{F} \equiv \epsilon_\text{F}/k$. We can calculate the pressure of a degenerate Fermi gas using $P = -(\partial U/\partial V)_{S,V}$:

$$P = -\frac{\partial}{\partial V}\left[\frac{3}{5}\frac{Nh^2}{8m}\left(\frac{3N}{\pi V}\right)^{2/3}\right] = \frac{2N\epsilon_\text{F}}{5V} = \frac{2}{3}\frac{U}{V}.$$

This is called the **degeneracy pressure** of a Fermi gas. Such pressure has nothing to do with electrostatic or other kinds of fundamental forces. It comes entirely from Pauli exclusion principle, and it is what keeps electrons from falling into protons by electrostatic force.

A.6.1 Small Nonzero Temperatures

The Fermi gas with zero temperature assumption cannot provide any information about the heat capacity. We shall see how nonzero but small temperatures would affect the Fermi gas.

At temperature T, any particle may acquire a thermal energy of roughly kT. However, most of the electrons cannot acquire such amount of energy because the higher energy states are already occupied. Only those with energy about $\epsilon_\text{F} - kT$ can just to the next level. (This obviouly leave *some* space for lower energy electrons.) Because of the energy tolerance kT, the number of electrons that are affected by an increase in T is proportional to T. Meanwhile, this number should be proportional to N as well because it is an extensive quantity. These arguments give the increase in energy of a Fermi gas when temperature goes from zero to T,

$$\begin{aligned}\Delta E &\propto (\text{number of affected electrons}) \times (\text{energy acquired by each})\\ &\propto (NkT) \times (kT)\\ &\propto N(kT)^2.\end{aligned}$$

By dimensional analysis, the proportionality constant should have units of $(\text{energy})^{-1}$. The only constant available in this context is ϵ_F, so the increase in energy must be some dimensionless constant times $N(kT)^2/\epsilon_\text{F}$. This dimensionless constant is determined to be $\pi^2/4$. In conclusion, the total energy of a degenerate Fermi gas for $kT \ll \epsilon_\text{F}$ and the heat capacity are

$$U = \frac{3}{5}N\epsilon_\text{F} + \frac{\pi^2}{4}\frac{N(kT)^2}{\epsilon_\text{F}}, \quad C_V = \left(\frac{\partial U}{\partial T}\right)_V = \frac{\pi^2 Nk^2T}{2\epsilon_\text{F}}.$$

A.6.2 The Density of States

To quantitatively determine the energy of a degenerate Fermi gas, consider the change of variable from n to the electron energy ϵ:

$$\epsilon = \frac{h^2}{8mL^2}n^2, \quad n = \sqrt{\frac{8mL^2}{h^2}}\sqrt{\epsilon}, \quad dn = \sqrt{\frac{8mL^2}{h^2}}\frac{1}{2\sqrt{\epsilon}}\,d\epsilon.$$

The energy integral at $T = 0$ becomes

$$U = \int_0^{\epsilon_\text{F}} \epsilon\left[\frac{\pi}{2}\left(\frac{8mL^2}{h^2}\right)^{3/2}\sqrt{\epsilon}\right]d\epsilon.$$

The term in the square bracket is the **density of states**, denotes as $g(\epsilon)$,

$$g(\epsilon) = \frac{\pi(8m)^{3/2}}{2h^3}V\sqrt{\epsilon} = \frac{3N}{2\epsilon_\text{F}^{3/2}}\sqrt{\epsilon}. \tag{A.7}$$

It indicates the number of states within the energy interval $\epsilon, \epsilon + d\epsilon$. Note that the density of states does not depend on N, where in the second equality N is canceled by $\epsilon_\text{F}^{3/2}$. It is a useful quantity to be integrated. If you want the number of states within ϵ_1 and ϵ_2, just integrate $g(\epsilon)$ in this range. Moreover, the density of states is a pure quantum mechanical quantity, and of course different models would have different forms of $g(\epsilon)$. Once the density of states of some system is known, it can be applied to thermal physics.

For an electron gas at zero temperauture, the total number of electrons is just

$$N = \int_0^{\epsilon_F} g(\epsilon)\, d\epsilon.$$

But when $T \neq 0$, not all states are occupied according to the previous section. In fact, we already know the probability of a state with energy ϵ being occupied—the Fermi-Dirac distribution. Therefore, we will modify the integrals of total number of electrons and total energy to

$$N = \int_0^{\infty} g(\epsilon)\bar{n}_{\mathrm{FD}}(\epsilon)\, d\epsilon = \int_0^{\infty} \frac{g(\epsilon)}{e^{(\epsilon-\mu)/kT}+1}\, d\epsilon.$$
$$U = \int_0^{\infty} \epsilon g(\epsilon)\bar{n}_{\mathrm{FD}}(\epsilon)\, d\epsilon = \int_0^{\infty} \frac{\epsilon g(\epsilon)}{e^{(\epsilon-\mu)/kT}+1}\, d\epsilon.$$

Note that the chemical potential is no longer equal to the Fermi energy when $T \neq 0$, but is slightly smaller instead. Figure A.2 shows the graph of the integrand $g(\epsilon)\bar{n}_{\mathrm{FD}}$ of the N-integral.

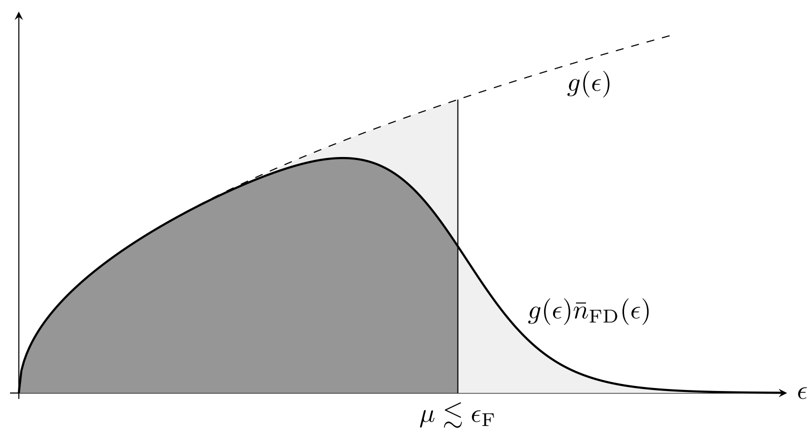


Figure A.2: At nonzero T, the number of fermions per unit energy is given by the density of states times the Fermi-Dirac distribution. Because increasing the temperature does not change the total number of fermions, the two lightly shaded areas must be equal. Since $g(\epsilon)$ is greater above ϵ_F than below, this means that the chemical potential decreases as T increases. This graph is drawn for $T/T_F = 0.1$; at this temperature μ is about 1% less than ϵ_F.

The chemical potential now plays a role of fixing the total number of particles. Once we get N in a closed form from the integral, it is possible to solve for $\mu(T)$. After obtaining $\mu(T)$, plug it into the U-integral gives the energy as a function of T, and hence the heat capacity. Unfortunately, these integrals do not have a closed form. They can only be evaluated by approximation in the limit $kT \ll \epsilon_F$.

A.6.3 The Sommerfeld Expansion

The **Sommerfeld expansion** can tell us the answer of $\mu(T)$ and $U(T)$ in the limit $kT \ll \epsilon_F$:

$$\frac{\mu}{\epsilon_F} = 1 - \frac{\pi^2}{12}\left(\frac{kT}{\epsilon_F}\right)^2 + \ldots, \tag{A.8}$$

$$U = \frac{3}{5}N\epsilon_F + \frac{\pi^2}{4}\frac{N(kT)^2}{\epsilon_F} + \ldots \tag{A.9}$$

There emerges the $\pi^2/4$ coefficient for the energy correction, and also the chemical potential decreases as T increases. The following is the proof:

Chemical Potential The integral for N is

$$N = \int_0^{\infty} g(\epsilon)\bar{n}_{\mathrm{FD}}(\epsilon)\, d\epsilon = g_0 \int_0^{\infty} \epsilon^{1/2}\bar{n}_{\mathrm{FD}}\, d\epsilon,$$

where g_0 is some constants before $\sqrt{\epsilon}$ in $g(\epsilon)$. The interesting region is near $\epsilon \approx \mu$ (see Figure A.2), so we should isolate this region by integration by parts:

$$N = \cancel{\frac{2}{3} g_0 \epsilon^{3/2} \bar{n}_{\mathrm{FD}} \Big|_0^\infty} + \frac{2}{3} g_0 \int_0^\infty \epsilon^{3/2} \left(-\frac{d\bar{n}_{\mathrm{FD}}}{d\epsilon} \right) d\epsilon.$$

By making a change of variable $x = (\epsilon - \mu)/kT$, the derivative $d\bar{n}_{\mathrm{FD}}/d\epsilon$ is

$$-\frac{d\bar{n}_{\mathrm{FD}}}{d\epsilon} = -\frac{d}{d\epsilon} \left[e^{(\epsilon-\mu)/kT} + 1 \right]^{-1} = \frac{1}{kT} \frac{e^x}{(e^x+1)^2}.$$

The integral becomes

$$N = \frac{2}{3} g_0 \int_0^\infty \frac{1}{kT} \frac{e^x}{(e^x+1)^2} \epsilon^{3/2} \, d\epsilon = \frac{2}{3} g_0 \int_{-\mu/kT}^\infty \frac{e^x}{(e^x+1)^2} \epsilon^{3/2} \, dx.$$

The integrand decays exponentially whn $|\epsilon - \mu| \gg kT$. Now make two approximations

1. Extend the lower limit of the integrand to $-\infty$ to make the integral symmetric.
2. Take the first few terms of the Taylor series about $\epsilon = \mu$.

$$\epsilon^{3/2} = \mu^{3/2} + \frac{3}{2}(\epsilon - \mu)\mu^{1/2} + \frac{3}{8}(\epsilon-\mu)^2 \mu^{-1/2} + \dots.$$

The integral becomes integrable with just polynomials and exponential,

$$N = \frac{2}{3} g_0 \int_{-\infty}^{+\infty} \frac{e^x}{(e^x+1)^2} \left[\mu^{3/2} + \frac{3}{2} xkT\mu^{1/2} + \frac{3}{8}(xkT)^2 \mu^{-1/2} + \dots \right] dx.$$

The first term:

$$\int_{-\infty}^{+\infty} \frac{e^x}{(e^x+1)^2} \, dx = \int_{-\infty}^{+\infty} -\frac{d\bar{n}_{\mathrm{FD}}}{d\epsilon} \, d\epsilon = \bar{n}_{\mathrm{FD}}(-\infty) - \bar{n}_{\mathrm{FD}}(\infty) = 1.$$

The second term is an odd function of x, so

$$\int_{-\infty}^{+\infty} \frac{xe^x}{(e^x+1)^2} \, dx = 0.$$

The third integral can be evaluated analytically. The result is

$$\int_{-\infty}^{+\infty} \frac{x^2 e^x}{(e^x+1)^2} \, dx = \frac{\pi^2}{3}.$$

Summing up these results give the total number of electrons:

$$N = \frac{2}{3} g_0 \mu^{3/2} + \frac{1}{4} g_0 (kT)^2 \mu^{-1/2} \cdot \frac{\pi^2}{3} + \dots = N \left[N \left(\frac{\mu}{\epsilon_{\mathrm{F}}} \right)^{3/2} + \frac{\pi^2}{8} \frac{(kT)^2}{\epsilon_{\mathrm{F}}^{3/2} \mu^{1/2}} + \dots \right]$$

Canceling the N, we see that $\mu/\epsilon_{\mathrm{F}} \approx 1$ because of the low temperature limit ($kT/\epsilon_F \ll 1$). The correction term is small, so we can approximate $\mu \approx \epsilon_{\mathrm{F}}$. The equation for $\mu/\epsilon_{\mathrm{F}}$ becomes

$$1 = \left(\frac{\mu}{\epsilon_{\mathrm{F}}} \right)^{3/2} + \frac{\pi^2}{8} \frac{(kT)^2}{\epsilon_{\mathrm{F}}^2} + \dots$$

Solve for $\mu/\epsilon_{\mathrm{F}}$:

$$\frac{\mu}{\epsilon_{\mathrm{F}}} = \left[1 - \frac{\pi^2}{8} \left(\frac{kT}{\epsilon_{\mathrm{F}}} \right)^2 + \dots \right]^{2/3} = 1 - \frac{\pi^2}{12} \left(\frac{kT}{\epsilon_{\mathrm{F}}} \right)^2 + \dots$$

Energy The U-integral is

$$U = \int_0^\infty \epsilon g(\epsilon) \bar{n}_{\mathrm{FD}}(\epsilon) \, d\epsilon = g_0 \int_0^\infty \epsilon^{3/2} \bar{n}_{\mathrm{FD}}(\epsilon) \, d\epsilon.$$

Integrating by parts,

$$U = \cancel{\frac{2}{5} g_0 \epsilon^{5/2} \bar{n}_{\mathrm{FD}}(\epsilon) \Big|_0^\infty} + \frac{2}{5} g_0 \int_0^\infty \epsilon^{5/2} \left(-\frac{d\bar{n}_{\mathrm{FD}}}{d\epsilon} \right) d\epsilon.$$

Let $x = (\epsilon - \mu)/kT$,

$$U = \frac{2}{5} g_0 \int_0^\infty \frac{1}{kT} \frac{e^x}{(e^x+1)^2} \epsilon^{5/2} \, d\epsilon = \frac{2}{5} g_0 \int_{-\mu/kT}^\infty \frac{e^x}{(e^x+1)^2} \epsilon^{5/2} \, dx.$$

Do the same approximation as in the chemical potential case: extend the lower limit of the integral to $-\infty$ and Taylor expand $\epsilon^{5/2}$,

$$\epsilon^{5/2} = \mu^{5/2} + \frac{5}{2}(\epsilon - \mu)\mu^{3/2} + \frac{15}{8}(\epsilon - \mu)^2 \mu^{1/2} + \cdots = \mu^{5/2} + \frac{5}{2}(xkT)\mu^{3/2} + \frac{15}{8}(xkT)^2 \mu^{1/2} + \ldots$$

The three integrals are the same: the first one is 1; the second one is zero; the third one is $\pi/3$, so

$$\begin{aligned} U &\approx \frac{2}{5} g_0 \mu^{5/2} + \frac{3}{4} g_0 (kT)^2 \mu^{1/2} \cdot \frac{\pi^2}{3} + \ldots \\ &= \frac{2}{5} g_0 \mu^{5/2} + \frac{\pi^2}{4} g_0 (kT)^2 \mu^{1/2} + \ldots \\ &= \frac{3}{5} \frac{N\mu^{5/2}}{\epsilon_{\mathrm{F}}^{3/2}} + \frac{3\pi^2}{8} \frac{N(kT)^2}{\epsilon_{\mathrm{F}}} + \ldots \end{aligned}$$

where in the second term $\mu = \epsilon_{\mathrm{F}}$ is used. Now we need to plug in μ,

$$\mu^{5/2} = \epsilon_{\mathrm{F}}^{5/2} \left[1 - \frac{\pi^2}{12} \left(\frac{kT}{\epsilon_{\mathrm{F}}} \right)^2 + \ldots \right]^{5/2} = \epsilon_{\mathrm{F}}^{5/2} \left[1 - \frac{5\pi^2}{24} \left(\frac{kT}{\epsilon_{\mathrm{F}}} \right)^2 + \ldots \right].$$

Finally,

$$U \approx \frac{3}{5} N \epsilon_{\mathrm{F}} \left[1 - \frac{5\pi^2}{24} \left(\frac{kT}{\epsilon_{\mathrm{F}}} \right)^2 \right] + \frac{3\pi^2}{8} \frac{N(kT)^2}{\epsilon_{\mathrm{F}}} = \frac{3}{5} N \epsilon_{\mathrm{F}} + \frac{\pi^2}{4} \frac{N(kT)^2}{\epsilon_{\mathrm{F}}}.$$

A.7 Blackbody Radiation

A.7.1 The Planck Distribution

Classical electromagnetism predicts electromagnetic waves as propagating fields through all space. If those fields are confined to a box with fixed temperature (like in an oven), they are actually infinite superpositions of standing waves with different frequencies. Each standing wave is like an harmonic oscillator: there are two quadratic degrees of freedom, so each has an energy of kT. Since there are infinite number of standing waves, the energy within the box is infinite, which obviously disagrees with experiments. This disagreement is called the **ultraviolet catastrophe**. Historically, it led to the birth of quantum mechanics.

In quantum mechanics, a harmonic oscillator cannot have arbitrary frequency or energy. The allowed energies are $E_n = (n + 1/2)\hbar\omega$. The partition function for one oscillator is

$$Z = 1 + e^{-\beta\hbar\omega} + e^{-2\beta\hbar\omega} + \cdots = \frac{1}{1 - e^{-\beta\hbar\omega}}.$$

This gives the average energy,

$$\langle E \rangle = -\frac{1}{Z} \frac{\partial Z}{\partial \beta} = \frac{\hbar\omega}{e^{\hbar\omega/kT} - 1}.$$

If we see the energy as in units of $h\nu = \hbar\omega$ (conventional notation), then the average number of units of energy in an oscillator is

$$\bar{n}_{\mathrm{Pl}} = \frac{1}{e^{h\nu/kT} - 1}. \tag{A.10}$$

This is called the **Planck distribution**. When $h\nu \gg kT$, the probability is exponentially suppressed. By making use of the quantized energy, the ultraviolet catastrophe is resolved.

The "units" of energy can be thought of as particles called **photons**. Photons are bosons, so they should obey the Bose-Einstein distribution:

$$\bar{n}_{\mathrm{BE}} = \frac{1}{e^{(\epsilon - \mu)/kT} - 1}.$$

With $\epsilon = h\nu$, the Planck distribution (A.10) says that the chemical potential for photons should be $\mu = 0$. There are two explanations for this:

1. At thermal equilibrium with T and V fixed, we know that the Helmholtz free energy is minimized. Since photons number is not conserved (as it can be absorbed by an electron), a system of N photons would take a specific N such that F is minimized. Then at equilibrium, by the partial derivative relation of F and μ:

$$\left(\frac{\partial F}{\partial N}\right)_{T,V} = \mu = 0.$$

2. Consider the reaction with a photon (γ) is emitted or absorbed by an electron:

$$e^- \longleftrightarrow e^- + \gamma.$$

The equilibrium condition is $\mu_e = \mu_e + \mu_\gamma$, so the chemical potential for photons is zero.

A.7.2 Total Particle Number and Energy

The Planck distribution provides the occupancy of photons. We shall compute the total number of photons in a box and the total energy of the system. Again, consider first the "particle in a 1D box" of length L. Photons have momentum

$$p = \frac{hn}{2L}, \quad n = 1, 2, 3, \ldots$$

Unlike electrons or other particles, photons are relativistic—their energies are given by

$$\epsilon = pc = \frac{hcn}{2L}.$$

Generalizing to a 3D box, the energy is

$$\epsilon = c\sqrt{p_x^2 + p_y^2 + p_z^2} = \frac{hc}{2L}\sqrt{n_x^2 + n_y^2 + n_z^2} = \frac{hcn}{2L}, \quad n_x, n_y, n_z = 1, 2, 3, \ldots$$

The next step is similar to the one in the last section. Summing over n, the total number of particles and energy are

$$N = 2\sum_{n_x,n_y,n_z} \bar{n}_{\mathrm{Pl}}(\epsilon) \approx 2\int_0^\infty dn \int_0^{\pi/2} d\theta \int_0^{\pi/2} d\phi \, \frac{n^2 \sin\theta}{e^{hcn/2LkT} - 1}.$$

$$U = 2\sum_{n_x,n_y,n_z} \epsilon(n)\bar{n}_{\mathrm{Pl}}(\epsilon) \approx \int_0^\infty dn \int_0^{\pi/2} d\theta \int_0^{\pi/2} d\phi \, \frac{hcn}{L}\frac{n^2 \sin\theta}{e^{hcn/2LkT} - 1}.$$

The factor of 2 accounts for the two polarizations of photons. Note that this time n runs from zero to infinity. Using the substitution $x = hcn/2LkT$, the N-integral is

$$N = \pi\int_0^\infty \frac{n^2}{e^{hcn/2LkT} - 1}\, dn = \pi\left(\frac{2LkT}{hc}\right)^3 \int_0^\infty \frac{x^2}{e^x - 1}\, dx = 8\pi V\left(\frac{kT}{hc}\right)^3 \int_0^\infty \frac{x^2}{e^x - 1}\, dx.$$

This integral does not have an analytical solution, the numerical value is $2\zeta(3) \approx 2.404$ where $\zeta(z)$ is the Riemann-zeta function.

The Planck Spectrum

Changing variables to $\epsilon = hcn/2L$, the U-integral becomes

$$\frac{U}{V} = \int_0^\infty \frac{8\pi\epsilon^3/(hc)^3}{e^{\epsilon/kT} - 1}\, d\epsilon.$$

The integrand is interpreted as the energy density per photon energy, or the **spectrum** of the photons:

$$u(\epsilon) = \frac{8\pi}{(hc)^3}\frac{\epsilon^3}{e^{\epsilon/kT} - 1}. \tag{A.11}$$

To evaluate the integral analytically, make a substitution $x = \epsilon/kT$,

$$\frac{U}{V} = \frac{8\pi(kT)^4}{(hc)^3}\int_0^\infty \frac{x^3}{e^x - 1}\,dx.$$

Note that the U-integral can also be written as

$$U = \int_0^\infty \epsilon\left[\frac{8\pi V\epsilon^2}{(hc)^3}\right]\bar{n}_{\mathrm{BE}}(\epsilon)\,d\epsilon.$$

The term in the square bracket is thus the density of states for photon (or other kinds of ultra-relativistic particles):

$$g(\epsilon) = \frac{8\pi V\epsilon^2}{(hc)^3}. \tag{A.12}$$

Back to the spectrum (A.11). By numerical calculation, the spectrum peaks at $\epsilon \approx 2.82kT$. **Wien's law** says that higher temperature gives a higher photon energy. Figure A.3 shows the integrand as a function of $x = \epsilon/kT$.

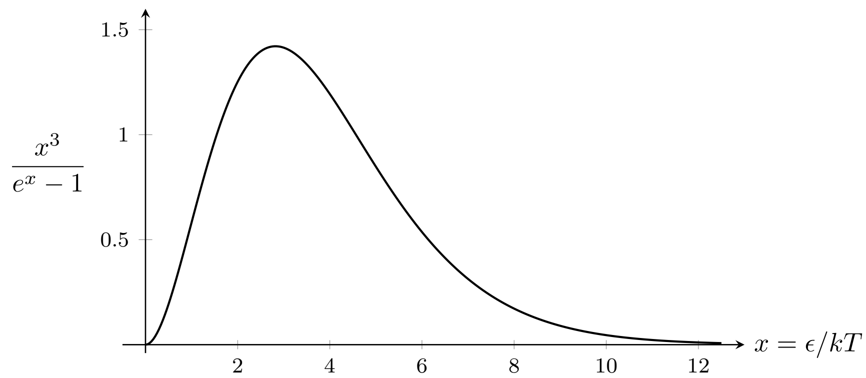


Figure A.3: The Planck spectrum, plotted in terms of the dimensionless variable $x = \epsilon/kT = h\nu/kT$. The area under any portion of this graph, multiplied by $8\pi(kT)^4/(hc)^3$, equals the energy density of electromagnetic radiation within the corresponding freqeuncy (or photon energy) range.

The total energy is just the integral. It has an analytical solution:

$$\frac{U}{V} = \frac{8\pi(kT)^4}{(hc)^3}\int_0^\infty \frac{x^3}{e^x - 1}\,dx = \frac{8\pi^5(kT)^4}{15(hc)^3}.$$

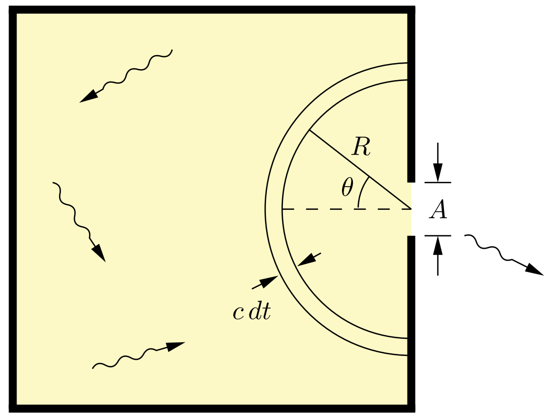
The total energy is proportional to the *fourth* power of the temperature. It is also possible to the spectrum in terms of wavelength $\lambda = hc/\epsilon$:

$$u(\lambda) = \frac{8\pi hc}{\lambda^5}\frac{1}{e^{hc/kT\lambda} - 1}. \tag{A.13}$$

It peaks at $\lambda = hc/4.97kT$. One may ask: why is the peak not at $\lambda_{\text{peak}} = hc/\epsilon_{\text{peak}} = hc/2.82kT$? Let's look at the two photon spectrum: (A.11) shows the energy density per unit photon energy ϵ, while (A.13) shows the energy density per unit wavelength λ. The problem is the photon energy and the wavelength are related to each other in a nonlinear way. For example, consider a 1 eV of photon energy unit $\delta\epsilon$. If we choose the photon energy range from 1 eV to 2 eV, this corresponds to a wavelength range of 621 nm to 1242 nm. If it is from 11 eV to 12 eV, the corresponding wavelength range is from 104 nm to 112 nm, which is a much smaller $\delta\lambda$.

A.7.3 Stefan-Boltzmann Law

Consider a box of photon gas, but with a small hole of area A. Photons have the same speed in vacuum, so low-energy photons will have the same probability escaping the hole as energetic photons. The spectrum inside the box should be the same as the one outside. We want to know the total amount of radiation coming out. Suppose there are some photons that can escape through the hole within a time interval dt. We know that for a time t prior to their escape, they were once located at some positions within a shell of radius $R = ct$ and thickness $c\,dt$.



The volume element of the shell is

$$dV = R^2 \sin\theta \, d\theta \, d\phi \, (c \, dt).$$

The energy in this volume element is

$$dE = \frac{U}{V} c R^2 \sin\theta \, d\theta \, d\phi \, dt.$$

However, not all energy in this volume will escape the hole because most of the photons are pointing in the wrong direction. The probability of pointing in the right direction is the apparent area of the hole seen from the volume, divided by the total possible area of the flow:

$$\mathcal{P} = \frac{A\cos\theta}{4\pi R^2}.$$

The true energy escaping from the volume element is

$$dE_{\text{esc}} = \frac{A\cos\theta}{4\pi} \frac{U}{V} c \sin\theta \, d\theta \, d\phi \, dt.$$

The power escaping is then

$$\frac{dE}{dt} = \int_0^{2\pi} d\phi \int_0^{\pi/2} d\theta \frac{A\cos\theta}{4\pi} \frac{U}{V} c \sin\theta = \frac{A}{4} \frac{U}{V} c.$$

From this we can obtain the flux, or power per unit area:

$$F_{\text{rad}} = \frac{c}{4} \frac{U}{V}.$$

Plug in (2.12), we obtain the **Stefan-Boltzmann law**:

$$F_{\text{rad}} = \frac{2\pi^5}{15} \frac{(kT)^4}{h^3 c^2} = \sigma T^4,$$

where $\sigma = 2\pi^5 k^4 / 15 h^3 c^2 \approx 5.67 \times 10^{-8}\,\mathrm{W/m^2\,K^4}$ is known as the **Stefan-Boltzmann constant**. The Stefan-Boltzmann law not only describes photons leaking through a hole, but also applies to radiation from non-reflecting surface at temperature T. This kind of radiation is called the **blackbody radiation**.

Here is the reasoning. Suppose there is a perfect blackbody surface that matches the size of the hole of the box. The photon gas in the box are at the same temperature as the blackbody. Then they must emit the same power. If one of them radiates more, then the other will heat up *when they are at thermal equilibrium*! That violates the second law of thermodynamics. Therefore, the blackbody must emit the same power as the photon gas does. Furthermore, if we have some filter that only allow a certain wavelength to pass through, then the same reasoning above holds for that wavelength. This is saying that the *whole* spectrum of the blackbody is the same as that of the hole.