The MIT Astrophysics 168

Michael A. Reefe[1]

[1]Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA

Last updated August 19, 2024

Contents

Introduction

These notes go over the 168 questions that PhD students in the Astrophysics Division of MIT's Physics Department are expected to know in preparation for their oral qualifying exam. The list of questions is updated for the 2023–2024 academic year, and covers all areas of astrophysics from the solar system to large scale structure and cosmology. I have written these notes primarily for my own sake while studying these questions, and have borrowed a lot of materials from previous generations of students' notes, but perhaps they may be useful for others studying in the future. The appendices in particular (§12.1–§12.12) may be useful as a quick reference for various information like constants, acronyms, telescopes, filters, notations, and a few topics that I felt warranted further expansion beyond what was covered in the questions themselves.

Conventions

Unit System:

Throughout these notes, I utilize the CGS (centimeter-gram-second) system of units since it is the most commonly used system in the current astrophysics community. This includes electromagnetism, where I use the CGS standard Gaussian units instead of SI units. This means that some E&M formulae may look foreign to students who have primarily worked in SI units (which use MKS, meter-kilogram-second, as a base). I personally find the Gaussian system more elegant, as it removes the factors of ε_0 and μ_0 that pop up frequently in SI units, and it also puts the E and B fields on equal grounds, measuring them both in the same units. These units are also supplemented by commonly used units in astrophysics, such as the solar mass ($M_\odot$), parsec (pc), gigayear (Gyr), kilometer per second ($\mathrm{km\,s^{-1}}$), ångström (Å), Jansky (Jy), etc. For definitions and conversions for some of these non-standard units, see §12.1 in the Appendix.

Notations and Normalizations:

I adopt the common convention of denoting vectors as bold, non-italicized characters. For example, $\mathbf{v}$ is a vector which has a scalar magnitude of $v \equiv |\mathbf{v}|$.

For the Fourier transform, I adopt the symmetric normalization:

$$\tilde{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}t\, e^{-i\omega t} f(t) \quad \longleftrightarrow \quad f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}\omega\, e^{i\omega t} \tilde{f}(\omega)$$

Relativity Conventions:

Explicit relativistic calculations do not pop up frequently in these notes, but for the few instances where they do, I always write 4-vectors with the index notation, i.e. x^μ. The time component of 4-vectors is always the 0-component, and I use the "mostly positive" (−,+,+,+) Mike Wazowski metric[1]

$$\eta_{\mu\nu} = \eta^{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

For the index notation on tensors, the first index always labels the **rows** of the matrix and the second index always labels the **columns** of the matrix.

[1] More commonly known by the boring name of the Minkowski metric

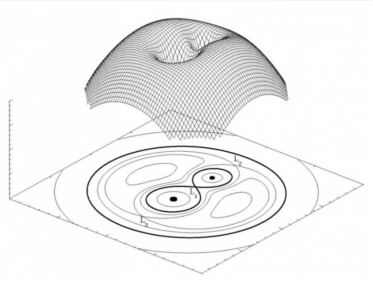
Before continuing, I also must warn you to turn back now, for your own sanity.

STOP DOING ASTROPHYSICS

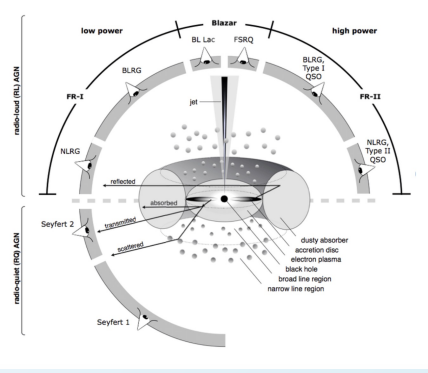
- THE UNIVERSE WAS NOT SUPPOSED TO BE UNDERSTOOD BY MORTALS
- CENTURIES OF OBSERVING yet NO REAL-WORLD USE FOUND for any object that is LIGHTYEARS AWAY or any law beyond an EMPIRICAL POWER LAW
- Wanted to look at further objects anyways for a laugh? We had a tool for that: It was called “STONEHENGE”
- “This star has a brightness of -23 magnitudes”; “An early-type galaxy is older than a late-type galaxy”; “Yes, radio jets can appear to move faster than the speed of light”; “Man, SDSS J0944-0038 sure looks beautiful today!”; ”Over 70% of the matter in the universe is some kind of ‘dark matter’ that we can’t see.” – Statements dreamed up by the utterly deranged.

LOOK at what Astrophysicists have been demanding your respect for all this time, with all the telescopes and satellites we built for them

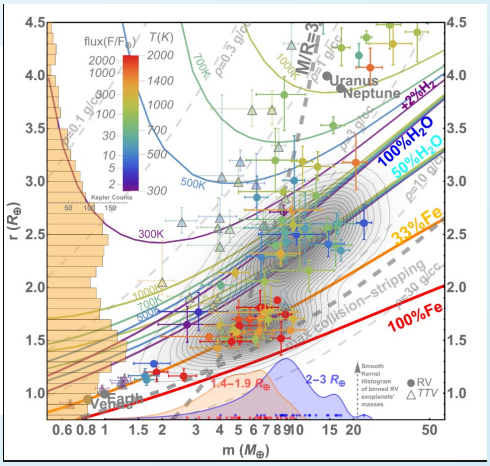
(This is REAL Astrophysics, done by REAL Astrophysicists):



???



???????



????????????????????

“Hello I would like $G_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$ curvature please”

They have played us for absolute fools

If you’re still undeterred, then continue at your own risk.

1 The Solar System

1. Describe, qualitatively, the standard model for the formation of the solar system, and discuss observational evidence for this model. Describe observations of exoplanets that have challenged this simple picture as a universal explanation for planet formation.

Formation Model[1]:

The Sun, Protoplanetary Disk, and Planets

1. A dense cloud of molecular Hydrogen becomes Jeans-unstable and collapses under its own weight. Since the cloud is denser at the center, it collapses inside-out, forming a **protostar** which contains most of the mass of the cloud.
2. Released gravitational energy increases temperatures in the protostar until it reaches hydrostatic equilibrium. Eventually it gets hot enough to start fusing deuterium (^{2}H), and then Hydrogen (^{1}H).
3. The outer parts of the cloud can be prevented from reaching the center if they have some amount of rotation, which causes them to instead form a **protoplanetary disk** around the protostar. Even a very slowly rotating cloud would contain much more angular momentum than the star could contain without breaking up, so most of the angular momentum remains in the disk.
4. As the protoplanetary disk cools, μm-sized silicate, iron, and water-ice grains condense out of the gas at different temperatures (i.e. different distances from the star). The star's gravity causes these grains to fall toward the midplane of the disk, causing collisional growth up to mm–cm. Further growth up to km sized **planetesimals** then occurs through collisions with other grains in the disk.
5. Motions of sub-cm grains are strongly coupled to the gas, which rotates sub-Keplerian due to an outward pressure gradient in the disk. Larger planetesimals will be at Keplerian speeds and thus encounter a headwind, causing them to lose angular momentum and migrate inwards.
6. Growth of planetesimals into **planetary embryos** (the largest planetesimals in different regions of the disk) is dominated by gravitational interactions, which depend on both the relative velocities and escape velocities of planetesimals. If the relative velocities are small relative to the escape velocities, the largest members of the population grow most rapidly, leading to **runaway growth**.
7. For the terrestrial planets, protoatmospheres can form once they are large enough ($\sim$0.01 $M_\oplus$) by degassing of accreted planetesimals. Optically thick protoatmospheres can trap heat in a blanketing effect, melting the planetary surface and causing differentiation. The present-day atmospheres were probably formed near the end of the accretionary epoch from outgassing of the hot planet and small planetesimal impacts. Solar winds prevent them from accreting massive atmospheres like the giant planets.
8. For the giant planets, once they become massive enough (several $M_\oplus$), they accumulate substantial envelopes of gas from the surrounding protoplanetary disk. Runaway gas accretion can occur once the protoplanet is massive enough to gravitationally compress its gas envelope.
9. Planetary orbits can migrate towards or away from their star through angular momentum exchanges:

- **Type I migration** occurs when a planet is too small to clear a gap around its orbit and is linearly proportional to the planet's mass.
- **Type II migration** occurs when a planet has cleared its orbit, causing it to be dragged by the disk as it viscously evolves.

Satellites

- **Regular satellites** (with low-eccentricity prograde orbits) formed via accretion in a **circumplanetary disk** of gas/dust around the planet's equatorial plane.
- **Irregular satellites** (high-eccentricity, high-inclination orbits, lower masses) were captured from heliocentric orbits.
- **Ring systems** form because tidal forces prevent accretion within an object's Roche limit.
- **Earth's Moon** is peculiar, as it has a large mass ratio compared to Earth and its composition more closely resembles that of Earth's mantle than it does solar composition. Therefore, the **collision ejection theory** states that a collision between Earth and a Mars-sized planetary embryo ejected material into Earth's orbit which then cooled and formed the Moon.

Asteroids and Comets

- The **asteroid belt** between Mars and Jupiter likely formed due to resonant perturbations from Jupiter which excited eccentricities and inclinations of asteroid zone planetesimals, preventing them from accreting. Much of the material may have been scattered into Jupiter-crossing orbits and ejected from the Solar System.
- **Trojan asteroids** (60° ahead of and behind Jupiter in its orbit) likely were moved there from the asteroid belt due to changes in Jupiter's orbit over time.
- Comets' highly volatile composition means they formed in the outermost parts of the protoplanetary disk. The **Kuiper belt** consists of Kuiper belt objects which likely formed close to their present locations.
- The **Oort cloud** likely consists of planetesimals that formed closer to the Sun and were then ejected by gravitational perturbations from the giant planets. Some Oort cloud comets may have formed around other stars in the Sun's birth cluster before being captured by the Solar System.

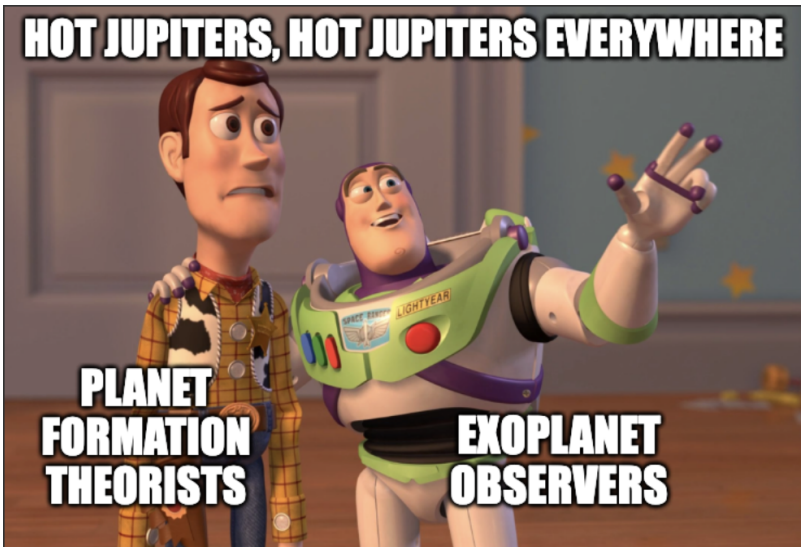
Observational evidence[1]:

- In the Solar System today, $> 98\%$ of the angular momentum is contained in planetary orbits, while 99.8% of the mass is in the Sun, supporting the molecular cloud collapse formation channel.
- The low-eccentricity, coplanar orbits of the Solar System planets are well explained by dynamical models of planetesimal accretion within a protoplanetary disk.
- Prograde planetary rotations of Jupiter and Saturn are a deterministic result from gas accretion.
- Cratering of planetary surfaces as evidence for the sweeping up of debris in planetary formation and an early high bombardment rate.
- The masses and bulk compositions of the planets all generally align with expectations for formation in a disk with decreasing temperature and surface density as a function of radius from the Sun.

- Volcanic and tectonic activity as evidence of hot protoplanets and differential heating during the accretionary epoch. Atmospheres of the rocky planets from outgassing.
- Radioisotope dating with $^{207}Pb/^{206}Pb$ puts the age of the Solar System at 4.568 Gyr.

Exoplanet challenges[1]:

- So-called "hot Jupiters" are giant planets in close-in orbits to their stars, which is difficult to explain with current models. Migration speeds are expected to increase as a planet approaches the star, so the chances of a giant planet migrating so close to its star without being completed destroyed seem very unlikely. These orbits can also be inclined, which cannot be explained by simple planet-disk interactions. It's possible that they transitioned from highly eccentric orbits to nearly circular ones through tidal interactions with the star near periapsis.



2. Explain the steps that you would take to show that both the orbital period and the total energy of a Keplerian orbit depend only on the semimajor axis, and not on the orbital eccentricity.

Total Energy[2]: The total energy of an orbit can be written as

$$E = T + U = \frac{1}{2}\mu v^2 - \frac{GM\mu}{r} \tag{1.1}$$

$$= \frac{1}{2}\mu \dot{r}^2 + \frac{1}{2}\mu r^2 \dot{\phi}^2 - \frac{GM\mu}{r} \tag{1.2}$$

$$= \frac{1}{2}\mu \dot{r}^2 + \frac{\ell^2}{2\mu r^2} - \frac{GM\mu}{r} \tag{1.3}$$

where $M = m_1 + m_2$ is the total mass, $\mu = m_1 m_2/(m_1 + m_2)$ is the reduced mass, and $\ell = \mu r^2 \dot{\phi}$ is the angular momentum. Since energy is conserved, the energy at periapsis and apoapsis should be equal. Using the equation of an ellipse with semimajor axis a and eccentricity e (see §12.9.1):

$$r(\phi) = \frac{a(1-e^2)}{1 + e\cos\phi} \tag{1.4}$$

And

$$r_{\min} = \frac{a(1-e^2)}{1+e} = a(1-e) \quad , \quad r_{\max} = \frac{a(1-e^2)}{1-e} = a(1+e) \tag{1.5}$$

Plugging these in and equating them (with $\dot{r} = 0$):

$$E(r_{\min/\max}) = \frac{\ell^2}{2\mu}\frac{1}{a^2(1\pm e)^2} - \frac{GM\mu}{a(1\pm e)} \tag{1.6}$$

$$\frac{\ell^2}{2\mu a^2}\left[\frac{1}{(1+e)^2} - \frac{1}{(1-e)^2}\right] = \frac{GM\mu}{a}\left[\frac{1}{1+e} - \frac{1}{1-e}\right] \tag{1.7}$$

$$\frac{\ell^2}{2a}\frac{-4e}{(1+e)^2(1-e)^2} = GM\mu^2\frac{-2e}{1-e^2} \tag{1.8}$$

Which, using $(1+e)^2(1-e)^2 = (1-e^2)^2$, gives

$$\boxed{\ell^2 = GM\mu^2 a(1-e^2)} \tag{1.9}$$

Then plugging this result back into $E(r_{\min})$:

$$E = \frac{1}{2\mu}\frac{GM\mu^2 a(1-e^2)}{a^2(1-e)^2} - \frac{GM\mu}{a(1-e)} \tag{1.10}$$

$$= GM\mu\left[\frac{1}{2}\frac{1+e}{a(1-e)} - \frac{1}{a(1-e)}\right] \tag{1.11}$$

The dependence on eccentricity all cancels out:

$$\boxed{E = -\frac{GM\mu}{2a}} \tag{1.12}$$

Orbital period[2]: The area that an object sweeps out over time in its orbit is

$$\mathrm{d}A = \frac{1}{2} r v \mathrm{d}t \quad \longrightarrow \quad \frac{\mathrm{d}A}{\mathrm{d}t} = \frac{1}{2} r v = \frac{\ell}{2\mu} \tag{1.13}$$

The period, therefore, can be written as

$$P = \frac{A}{\mathrm{d}A/\mathrm{d}t} = \frac{2\pi a b \mu}{\ell} \tag{1.14}$$

where the semiminor axis $b = a\sqrt{1-e^2}$. Squaring both sides and using the expression for ℓ from (1.9):

$$P^2 = \frac{4\pi^2 a^4 (1-e^2)\mu^2}{\ell^2} = \frac{4\pi^2 a^4 (1-e^2)\mu^2}{a(1-e^2)GM\mu^2} \tag{1.15}$$

Once again, the dependence on e cancels out and we are left with Kepler's 3rd law:

$$\boxed{P^2 = \frac{4\pi^2}{GM} a^3} \tag{1.16}$$

3. Calculate the approximate distance to the heliopause. Does the local interstellar medium begin at this boundary? Explain.

The **heliosphere** is the space around the Solar System that contains magnetic fields and plasma of solar origin. The **heliopause** is the boundary where the heliosphere meets with the interstellar medium (ISM)[1].

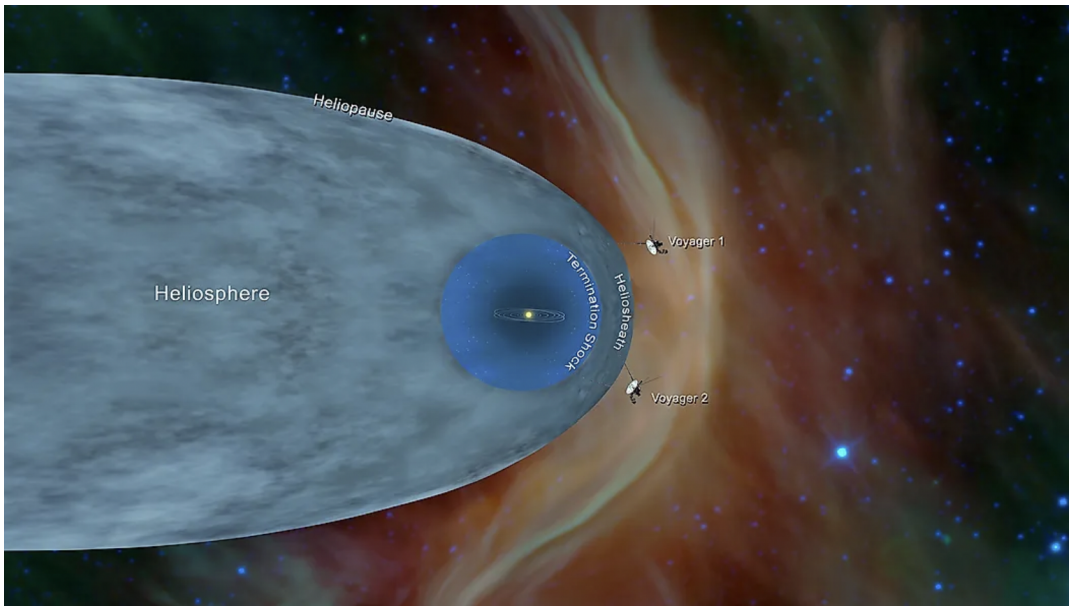


Figure 1.1: Schematic diagram of the heliosphere (credit: NASA)

This boundary can be found by solving for where the solar wind pressure is equal to the ISM pressure. The solar wind pressure is given by ram pressure:

$$P_{\odot}(r) = \rho(r)v^2(r) \tag{1.17}$$

The only solar wind particles that will make it out of the heliopause will be at or above the escape velocity of the Sun:

$$E = \frac{1}{2}mv_{\mathrm{esc},\odot}^2 - \frac{GmM_{\odot}}{R_{\odot}} = 0 \tag{1.18}$$

$$v \sim v_{\mathrm{esc},\odot} = \sqrt{\frac{2GM_{\odot}}{R_{\odot}}} \sim 600\ \mathrm{km\,s^{-1}} \tag{1.19}$$

We can use the mass loss rate of the Sun, $\dot{M} \sim 10^{-14}\ M_{\odot}\ \mathrm{yr}^{-1}$, and conservation of mass to determine the density of the solar wind at some radius:

$$\frac{\mathrm{d}M}{\mathrm{d}r} = \frac{\mathrm{d}M}{\mathrm{d}t}\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^{-1} = \frac{\dot{M}}{v(r)} = 4\pi\rho(r)r^2 \tag{1.20}$$

Which simplifies to

$$\rho(r) = \frac{\dot{M}}{4\pi r^2 v(r)}. \tag{1.21}$$

And the ISM pressure is mostly thermal:

$$P_{\text{ISM}} = n_{\text{ISM}} k T_{\text{ISM}} \tag{1.22}$$

Typical values in the ISM are $n_{\text{ISM}} \sim 10^{-1}$ cm^{-3} and $T_{\text{ISM}} \sim 10^4$ K for the warm phase, but the phase doesn't really matter as they all have roughly the same pressure of $P/k \sim 3\times10^3$ K cm^{-3} (see §5.Q72). Equating the two pressures and solving for r:

$$\frac{\dot{M}}{4\pi r^2} v(r) = P_{\text{ISM}} \tag{1.23}$$

$$\boxed{r = \sqrt{\frac{\dot{M} v}{4\pi P_{\text{ISM}}}}} \tag{1.24}$$

Doing a rough order of magnitude calculation

$$r \sim \left(\frac{10^{-14+33-7} \times 10^{2+5}}{10 \times 10^{-16} \times 10^3} \right)^{1/2-13} \sim 10^{2.5} \sim 100 \text{ AU} \tag{1.25}$$

Note, as in Figure 1.1, the heliopause is not spherical, so this is a very rough order of magnitude approximation.

4. Describe the physics involved in the Earth-Moon interaction whereby the Earth's rotation rate is slowing and the orbital separation is increasing.

The Moon's gravitational force on the Earth creates a tidal force, which is a differential gravitational force between the equator and the poles. This creates a bulge in the oceans around the center where the force is strongest (see Figure 1.2).

$$\mathbf{F}_{\text{tidal}} = \mathbf{F}_g(\mathbf{R}) - \mathbf{F}_g(0) \tag{1.26}$$

Where

$$\mathbf{F}_g(\mathbf{0}) = \frac{GM_\oplus M_M}{r^2}\hat{\mathbf{x}} \ , \ \mathbf{F}_g(\mathbf{R}) = \frac{GM_\oplus M_M}{s^2}(\cos\phi\hat{\mathbf{x}} - \sin\phi\hat{\mathbf{y}}) \tag{1.27}$$

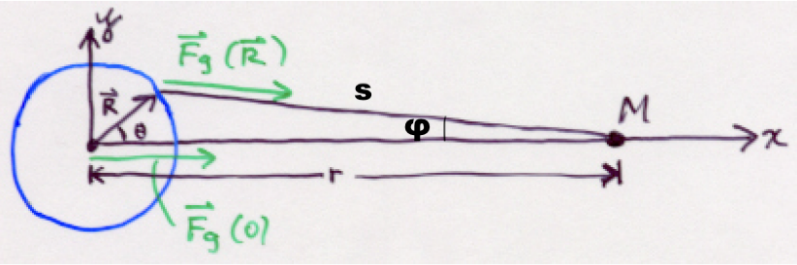


Figure 1.2: The tidal force creating bulges in the Earth's oceans. Credit: 3.

Using the law of cosines to get s^2:

$$s^2 = r^2 + R_\oplus^2 - 2rR_\oplus\cos\theta \approx r^2\left(1 - \frac{2R_\oplus\cos\theta}{r}\right) \tag{1.28}$$

R^2 is small relative to r and can be ignored. Plugging this back in and using the binomial approximation:

$$|\mathbf{F}_g(\mathbf{R})| \approx \frac{GM_\oplus M_M}{r^2}\left(1 + \frac{2R_\oplus\cos\theta}{r}\right) \tag{1.29}$$

Then, using the law of sines:

$$\frac{\sin\phi}{R_\oplus} = \frac{\sin\theta}{s} \approx \frac{\sin\theta}{r} \quad \longrightarrow \quad \sin\phi \approx \frac{R_\oplus}{r}\sin\theta \tag{1.30}$$

$$\cos\phi = \sqrt{1-\sin^2\phi} \approx 1 - \frac{R_\oplus^2}{2r^2}\sin^2\theta \approx 1 \tag{1.31}$$

Thus,

$$\mathbf{F}_g(\mathbf{R}) \approx \frac{GM_\oplus M_M}{r^2}\left(1 + \frac{2R_\oplus\cos\theta}{r}\right)\left(\hat{\mathbf{x}} - \frac{R_\oplus\sin\theta}{r}\hat{\mathbf{y}}\right) \tag{1.32}$$

And the tidal force becomes proportional to r^{-3} (dropping terms of order r^{-4} or higher):

$$\boxed{\mathbf{F}_{\text{tidal}}(\theta) = \frac{GM_\oplus M_M R_\oplus}{r^3}(2\cos\theta\hat{\mathbf{x}} - \sin\theta\hat{\mathbf{y}})} \tag{1.33}$$

However, Earth rotates faster than the Moon orbits, so it drags the tidal bulge ahead of the line from the Earth to the Moon. Thus, the Moon exerts a net torque on the tidal bulge, causing so-called "**tidal friction**" (see Figure 1.3).

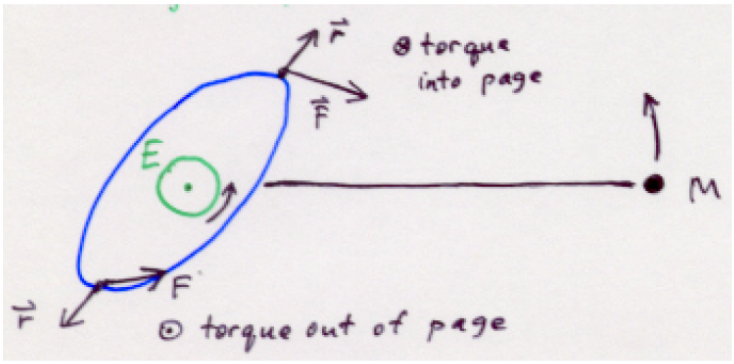


Figure 1.3: The Earth-Moon tidal interaction[3].

The torque on the bulge is transferred to the Earth, which slows down its rotation rate. And by Newton's 3rd law, the Earth also exterts a net torque on the Moon, which increases its orbital separation (the affect on the rotation rate of the Moon is higher-order since it is also differential and depends on the size of the Moon, which is much smaller than the Earth).

There should be negligible net external torque on the whole system, so we can use conservation of angular momentum to relate the change in Earth's rotation rate to the change in the Moon's orbital distance:

$$\dot{\ell}_\oplus + \dot{\ell}_M = 0 \tag{1.34}$$

$$\ell_\oplus = I_\oplus \omega_\oplus \quad \longrightarrow \quad \dot{\ell}_\oplus = 0.33 M_\oplus R_\oplus^2 \dot{\omega}_\oplus \tag{1.35}$$

$$\ell_M \approx \sqrt{GM\mu^2 a} \quad \longrightarrow \quad \dot{\ell}_M = \frac{1}{2}\sqrt{GM\mu^2/a}\,\dot{a} \tag{1.36}$$

Notes:

1. The factor 0.33 Earth's moment of inertia $I_\oplus$ is slightly smaller than that for a uniform sphere (2/5), since the Earth's density changes as a function of radius.
2. I have assumed, for simplicity, that the Moon's orbital eccentricity is 0, and that the Moon's orbital plane equals the Earth's rotational plane (i.e. its inclincation is also 0).

We can solve for $\dot{a}$ or $\dot{\omega}_\oplus$ in terms of each other, i.e.

$$\dot{\omega}_\oplus = -\frac{1}{2(0.33)M_\oplus R_\oplus^2}\sqrt{\frac{GM\mu^2}{a}}\dot{a} \tag{1.37}$$

But in general we cannot solve for both of them from first principles since we would need to know exactly how the Earth deforms in response to the tidal force and how strong the tidal friction mechanism is. Modern estimates of $\dot{a}$ are ~ 4 cm yr^{-1}, which gives $\dot{\omega}_\oplus \approx -2 \times 10^{-14}$ rad s^{-1} yr^{-1}. Thus, the day is getting longer by (see Ref. 3)

$$\boxed{\dot{P}_\oplus = -P_\oplus^2 \frac{\dot{\omega}_\oplus}{2\pi} \sim 24\ \mu\text{s yr}^{-1}} \tag{1.38}$$

5. Describe the Oort cloud and the Kuiper belt, and theories of their origins.

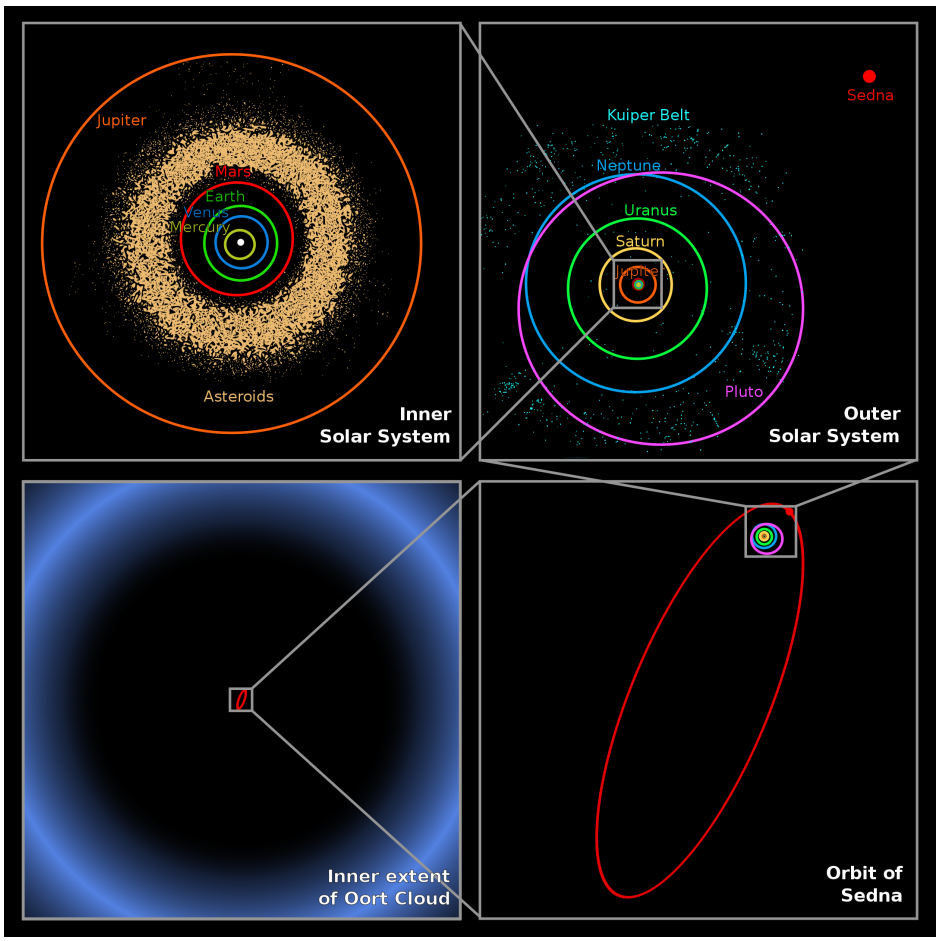


Figure 1.4: Zoom-out diagram of the Kuiper belt and Oort cloud. Credit: NASA/JPL.

Kuiper belt[1]: A thick disk of icy/rocky bodies beyond the orbit of Neptune, at ∼35–50 AU, that is far more massive than, and contains more massive/more icy objects than, the asteroid belt. This is the primary reservoir for short-period comets. It is also home to Pluto and other dwarf planets. The estimated number of Kuiper belt objects with radii $\gtrsim$ 1 km is $\sim 10^8$–10^{10}. Dave Jewitt and Jane Luu (MIT graduate student) found the first trans-Neptunian/Kuiper belt object.

Oort cloud[1]: A nearly spherical region at heliocentric distances of ~ 1–5×10^4 AU that is a giant reservoir for long-period, dynamically new comets. The estimated number of comets in the Oort cloud with radii $\gtrsim$ 1 km is $\sim 10^{12}$–10^{13}.

The "Nice" Formation model[4,5,6]:

1. The giant planets initially formed much closer to the Sun (∼5–17 AU) along with a massive disk of planetesimals out to ∼35 AU.
2. Planetesimals from the inner edge of the disk are gravitationally perturbed by the giant planets, scattering them inwards while the planets are nudged outwards.
3. The planetesimals are then scattered by Jupiter, sending them on highly eccentric, inclined orbits and sometimes ejecting them entirely.

4. After a few Myr of slow migration, Jupiter and Saturn approach a 1:2 orbital resonance that causes them to rapidly increase in eccentricity and destabilize the whole system. They drift outwards, interacting with Uranus and Neptune and causing them to become eccentric and drift outwards as well.
5. The ice giants plow through the planetesimal disk, displacing 99% of its mass into the outer Solar System.
6. The giant planets eventually reach their present-day distances and through dynamical friction they settle down into mostly circular orbits.

6. Explain why solar system bodies are often found with integer ratios of orbital periods.

This is due to the formation of **mean motion resonances**[1]. This can be thought of like a driven harmonic oscillator where the driving frequency is very close to the natural frequency, causing large amplitude effects:

$$\ddot{x} + \omega_0^2 x = f\cos\big(\omega_f t\big) \;\longrightarrow\; A \propto \frac{1}{\omega_0^2 - \omega_f^2}\ . \tag{1.39}$$

In most cases, these resonances drive *unstable* interactions where momentum is exchanged between bodies until the resonance no longer exists. But there are certain situations in which they can be held in place (self-correcting) by stable "locks" that result from nonlinear effects. Differential tidal recession (see §1.Q4) can bring Moons into resonance, with nonlinear interactions locking them there. "First-order" resonances of the form $(N+1)/N$ are usually the strongest.

Examples:

- Io:Europa:Ganymede orbit Jupiter in a 4:2:1 ratio (this specific ratio is called a **Laplace resonance**)
- Trojan asteroids are locked on a 1:1 resonance with Jupiter
- Neptune:Pluto have a 3:2 resonance
- Saturn's Moons have several pairs of resonances (Janus:Epimetheus, Mimas:Tethys, Enceladus:Dione)
- Seen in the asteroid belt (with Jupiter) and Kuiper belt (with Neptune)

There are also **secular resonances**[1] which occur when the apses/nodes of orbits precess at the same rate. In a restricted 3-body problem where the mass of the 3rd body is small relative to the other 2, secular perturbations can change the small body's orbital inclination and eccentricity, exchanging one for the other. For high inclinations, the **Kozai–Lidov mechanism** forces the argument of periapse to oscillate around a constant value, producing large periodic variations in eccentricity and inclination as they are exchanged between each other. Specifically, the quantity $L_z = \sqrt{1-e^2}\cos i$ is conserved and remains constant.

7. How did the Greeks determine the size of the Earth and the distance to the Moon?

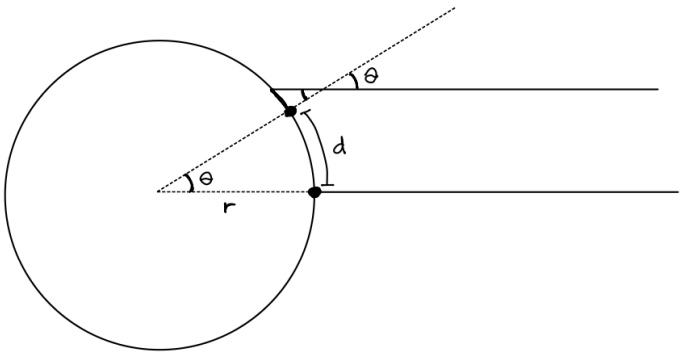


Figure 1.5: Geometry to calculate the size of the Earth

Size of the Earth[7]: The Greek mathematician **Eratosthenes** (c. 276–195 BC) was able to calculate the circumference of the Earth by using the difference in the angles of shadows cast in two different cities at the same time.

Calling the distance between the cities d and the angle of the shadow θ (see Figure 1.5), the circumference C can be found:

$$C = 2\pi r \quad \text{and} \quad \theta = \frac{d}{r} \tag{1.40}$$

$$\therefore\ C = \frac{2\pi}{\theta} d \tag{1.41}$$

Eratosthenes used the cities of Syene (modern day Aswan), Etypt, and Alexandria, Egypt, which were measured to be ∼5,000 stadia apart. When the Sun was directly overhead in Syene, casting no shadows, the angle of shadows on sundials in Alexandria was measured to be 1/50th of a circle. He therefore measured the circumference to be ∼250,000 stadia. The length of the stadion that Eratosthenes used is thought to be ∼157 m, which gives a circumference of **39,250 km**. This is remarkably accurate compared to the modern value of **40,075 km** (within ∼2%).

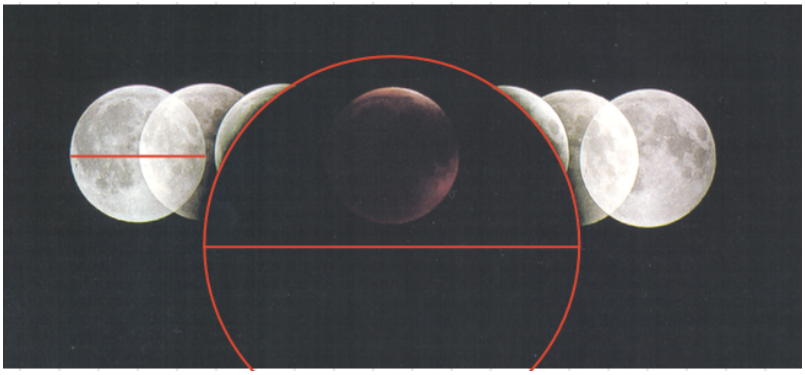


Figure 1.6: Measurement of the relative size of the Moon and the Earth's umbral shadow

Distance to the Moon[8]: The Greek mathematician **Aristarchus** (c. 310–230 BC) took advantage of lunar eclipses to directly measure the size of the Moon relative to Earth's umbral shadow (see Figure 1.6), obtaining a ratio of $\sim 1/3$. From there, he extrapolated an estimate for the distance to the Moon using geometry.

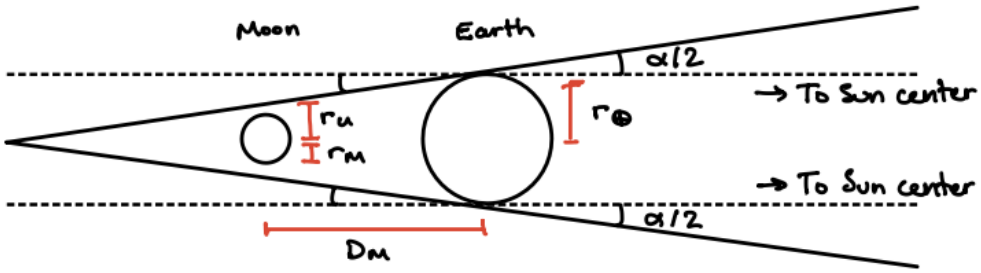


Figure 1.7: Geometry during a lunar eclipse

The size of the Earth's umbral shadow projected onto the Moon is smaller than the true size of the Earth. Note that in the figure, since $r_\odot \gg r_\oplus$, the total angular size of the Sun is given by $\sim 2 \times \alpha/2 = \alpha$. Since the angular size of the Sun is approximately the same as that of the Moon ($\alpha \sim \alpha_M \sim 0.5°$, which was observed at the time by eye), the Earth's umbral shadow radius is smaller by exactly 1 Moon radius (see Figure 1.7). Thus,

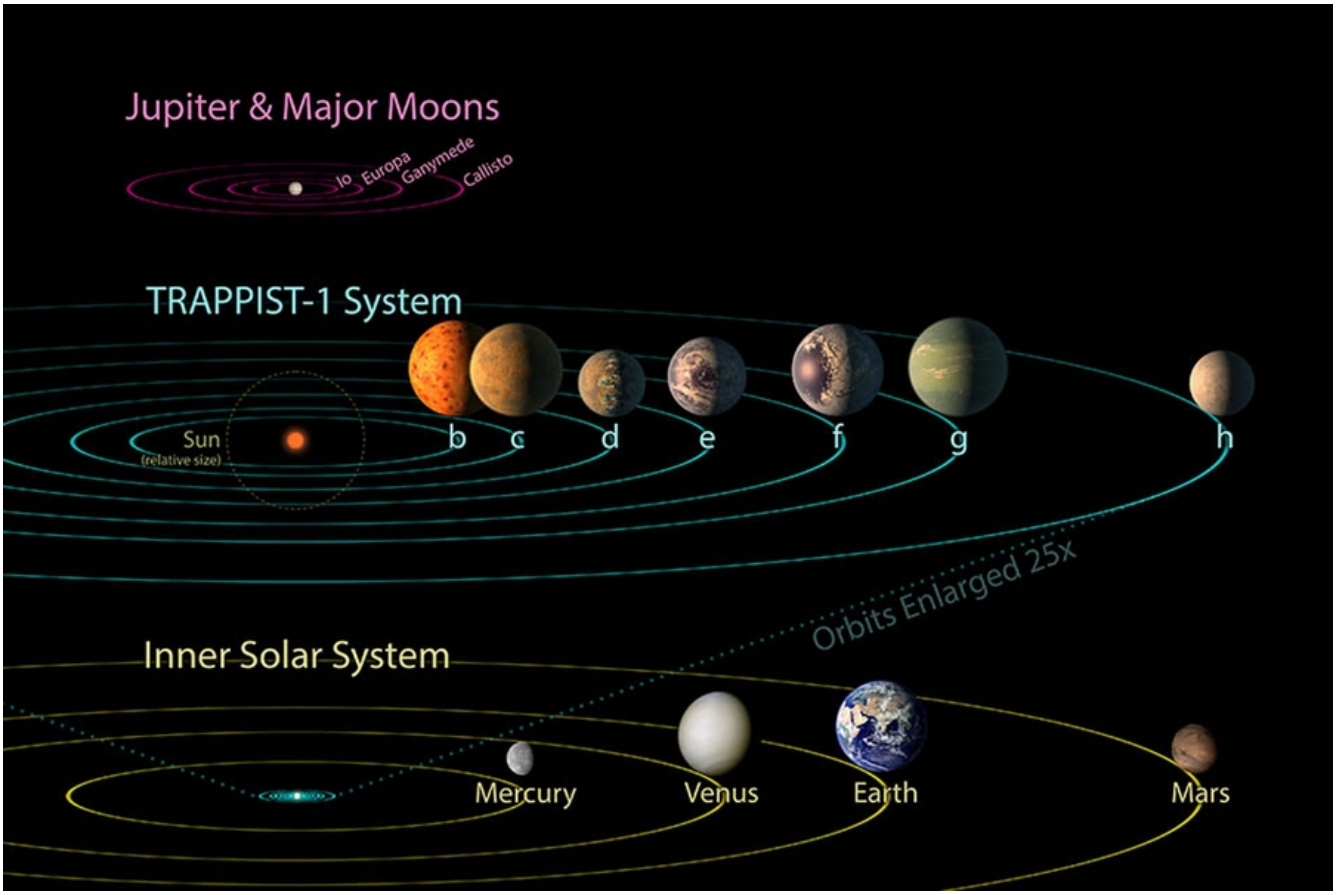
$$\xi = \frac{r_u}{r_M} \sim 3 \tag{1.42}$$

$$r_u = r_\oplus - D_M \frac{\alpha}{2} \quad \longrightarrow \quad \xi = \frac{r_\oplus}{r_M} - \frac{\alpha}{\alpha_M} \quad \longrightarrow \quad r_M = \frac{r_\oplus}{1+\xi} \tag{1.43}$$

$$\tan\frac{\alpha_M}{2} = \frac{r_M}{D_M} \quad \longrightarrow \quad D_M \approx \frac{2}{\alpha_M} r_M = \frac{2}{\alpha_M(1+\xi)} r_\oplus \tag{1.44}$$

where the last line utilizes the small angle approximation. This gives an estimate of $D_M \sim 60 r_\oplus$, which is once again surprisingly close to the modern value of $60.3 r_\oplus$ (given in units of Earth radii, as Aristarchus predates Eratosthenes and his measurements of the size of the Earth).

2 Exoplanets



8. *TESS* finds an exoplanet. What can one directly determine from this measurement? What additional observations are needed to determine the mass of this planet?

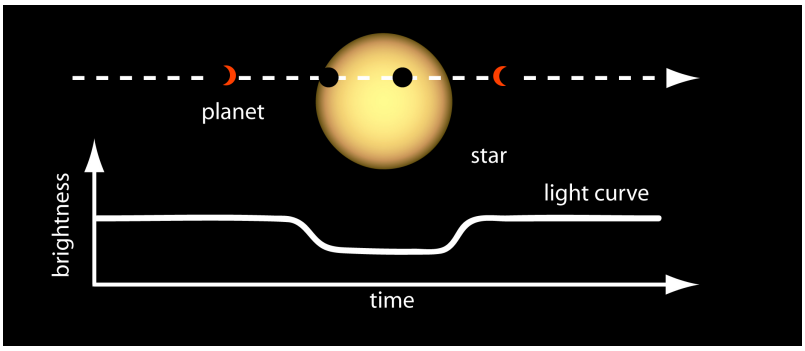


Figure 2.1: Exoplanet transit diagram (Credit: NASA Ames)

Direct observables[3,9]: *TESS* uses the transit method to detect exoplanets (see §2.Q13), which measures the dip in brightness as a planet passes in front of its star (Figure 2.1). Measurables: P, R_p/R_*, a/R_*, i (inclination) / b (impact parameter), ρ_*.

The orbital period P can be directly determined by measuring the time between repeated transits. The magnitude of the dip (or "depth") can be directly related to the planet's radius R_p relative to the star's R_*, since the amount of light the planet blocks is proportional to its cross-sectional area:

$$F_{\text{baseline}} \propto \pi R_*^2 \quad , \quad F_{\text{in-transit}} \propto \pi(R_*^2 - R_p^2) \tag{2.1}$$

$$\delta = \frac{F_{\text{baseline}} - F_{\text{in-transit}}}{F_{\text{baseline}}} \longrightarrow \boxed{\delta = \left(\frac{R_p}{R_*}\right)^2} \tag{2.2}$$

Thus, δ is directly observable from the transit depth, and the stellar radius R_* can be independently modeled using isochrones with stellar photometry or spectra, leading to an estimate for R_p. If the planet has a large enough atmosphere, differences in the transit depth in different wavelength bands can be used to measure broad spectroscopic features. If the atmosphere has mean molecular mass μm_p, temperature T, and gravitational acceleration g, then the atmospheric scale height H increases the "effective" planet radius:

$$\delta(\lambda) = \left(\frac{R_p + N_H H}{R_*}\right)^2 \quad \text{where} \quad H = \frac{kT}{\mu m_p g} \tag{2.3}$$

The duration of the transit T_{dur} can be related to a few different orbital parameters. In the simplest case, assuming a circular orbit with no inclination (Figure 2.2, left):

$$\sin\theta = \frac{R_*}{a} \quad , \quad \frac{T_{\text{dur}}}{P} = \frac{2\theta}{2\pi} \longrightarrow T_{\text{dur}} = \frac{P}{\pi}\arcsin\left(\frac{R_*}{a}\right) \tag{2.4}$$

We can include the effects of inclination with only a small adjustment (Figure 2.2, right):

$$\sin\theta = \frac{L}{a} \quad \text{where} \quad L = \sqrt{(R_* + R_p)^2 - (bR_*)^2} \quad , \quad b = \frac{a}{R_*}\cos i \tag{2.5}$$

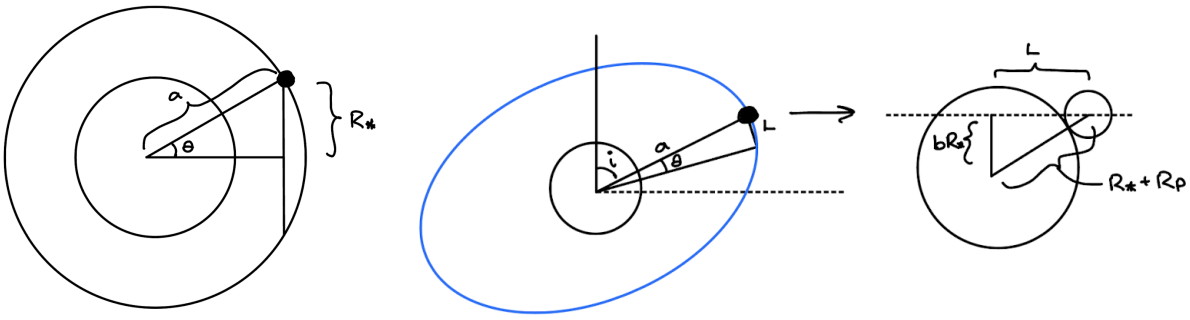


Figure 2.2: Transit duration geometry with and without inclination

$$\therefore \quad \boxed{T_{\text{dur}} = \frac{P}{\pi} \arcsin\left(\frac{R_*}{a}\sqrt{(1+\sqrt{\delta})^2 - b^2}\right)} \tag{2.6}$$

One can also get the stellar density purely from transit observables P and a/R_* by using Kepler's 3rd law (1.16):

$$M_p + M_* = \frac{4\pi^2 a^3}{GP^2} \tag{2.7}$$

$$\left(\frac{R_p}{R_*}\right)^3 \rho_p + \rho_* = \frac{3\pi}{GP^2}\left(\frac{a}{R_*}\right)^3 \tag{2.8}$$

$$\boxed{\rho_* \approx \frac{3\pi}{GP^2}\left(\frac{a}{R_*}\right)^3} \tag{2.9}$$

Getting the mass[3]: To get the mass of the planet, spectroscopic RV follow-up is necessary (see §2.Q13). The amplitude of the radial velocities imposed on a star by a planet are directly related to the planet mass by the **binary mass function**. This is derived in §3.Q20 in the context of a binary star system, but it can be easily applied to a star-planet system by just replacing all of the quantities associated with star 2 to the planet:

$$K = \left(\frac{2\pi G}{P}\right)^{1/3} \frac{M_p \sin i}{(M_* + M_p)^{2/3}} \frac{1}{\sqrt{1-e^2}} \tag{2.10}$$

Notice it is proportional to $M_p \sin i$, the planet mass degenerate with the inclination angle i.

9. With what velocity precision must one measure the reflex motion of a sun-like star to detect an Earth-mass planet in a 1-year orbit about it? What astrophysical processes complicate the measurement of radial velocities with this precision?

We can plug in $M_p \sin i = M_\oplus$, $P = 1$ yr, and $M_* = M_\odot$ into the RV amplitude equation (3.26) to see that it generates a signal of $\sim$**10 cm s^{-1}**. This is beyond the current limits of RV precision which can only get down to $\sim$1 m s^{-1}, 0.5 m s^{-1} in the absolute best cases.

The main astrophysical process that bottlenecks sensitivity is **stellar activity**[1]. This is includes stellar rotation, granulation from near-surface convection, and intrinsic variability (i.e. starspots). These sources of noise are most prominent in **younger and cooler** stars, and tend to go away as they age and lose angular momentum to stellar winds. However, there are other problems with measuring RVs for hot stars (i.e. A, B, and O-types): they have relatively few absorption lines to measure RVs from, and the ones they do have tend to be Doppler broadened much more due to higher rotational velocities ($\sim$200 km s^{-1} compared to $\sim$2 km s^{-1} for Sun-like stars), which makes the line velocities harder to measure precisely. Additional important sources of systematics include:

- Earth's time-varying rotation and revolution, and the Sun's motion relative to the star
- Atmospheric seeing and turbulence, which requires stable reference spectra

10. What is meant by the "obliquity" of an exoplanet orbit about its host star? How can obliquity be measured, and what are typical values?

The **obliquity** in this case is the angle between the rotational axis of the star and the orbital axis of the planet. It can be measured using the **Rossiter-McLaughlin Effect**: A rotating star will emit slightly redshifted light from the side moving away from us and slightly blueshifted light from the side moving towards us. When a planet passes in front of the star and partially blocks some light from the blueshifted side, it looks like a small anomalous redshift in the radial velocities, and vice versa when it passes over the redshifted side of the star (see Figure 2.3). However, this can only measure the angle between the projections of the vectors on the sky. The true angle is usually poorly understood because i_* is not well constrained. Typical values of obliquity are between $\sim$10°–30°.

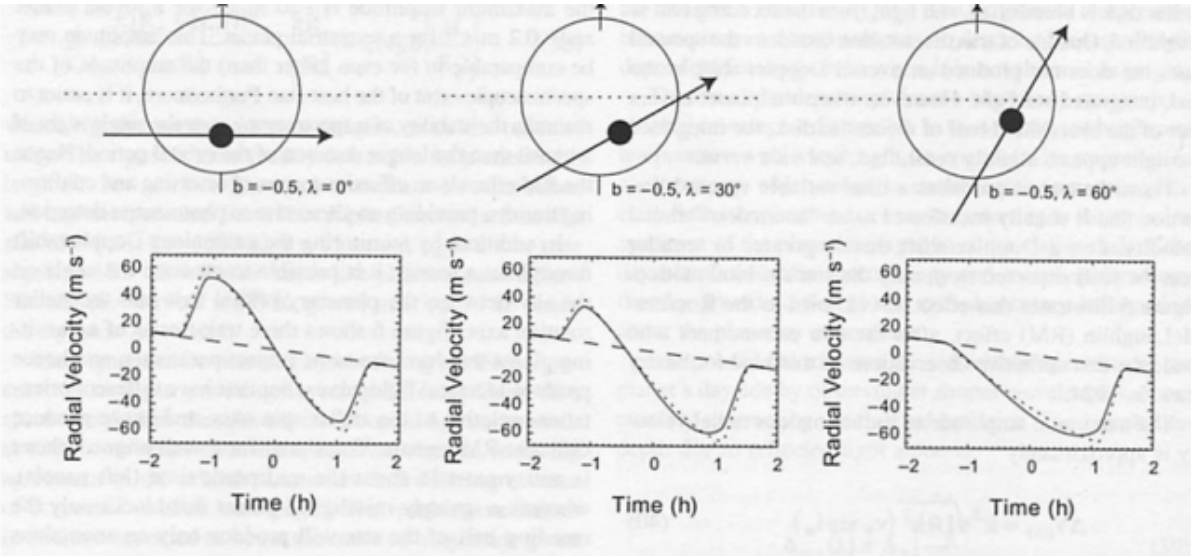


Figure 2.3: A diagram showing the Rossiter-McLaughlin effect[10]

In other contexts, obliquity can also refer to the angle between a planet's rotational axis and its orbital axis, but with current techniques it is not possible to measure this angle for any planets outside of our own Solar System. It might be possible to do it with a version of the Rossiter-McLaughlin effect using an exomoon's orbit around the exoplanet, but we don't currently have the sensitivity to be able to do it (we would need to be able to measure exomoon-induced RVs which will be $\ll$ than the previously estimated exo-Earth induced RVs of 10 cm s^{-1}, which are already too small for current techniques).

11. What is meant by tidal locking between an exoplanet and its host star? What are the circumstances where planets are likely to be tidally locked?

Tidal locking[1,11] occurs when the rotation period of the exoplanet is the same as its orbital period, meaning that the same hemisphere of the planet always faces the same hemisphere of the star. This is the stable end state of a system undergoing tidal friction, see §1.Q4. The tidal force is proportional to $1/r^3$, so it falls off quickly with distances, meaning only the most close-in planets are likely to be tidally locked. In most cases, the end state is **synchronous rotation** where there is a 1:1 ratio between the rotational and orbital period. However, there is also **asynchronous rotation** where it reaches another ratio, i.e. 3:2 for Mercury.

12. Describe qualitatively one of the theories explaining the presence of hot Jupiters – Jupiter mass planets orbiting at separations of less than 0.05 AU from the star. Why was the discovery of these planets such a surprise?

The discovery of hot Jupiters was surprising because it was not thought that giant planets could have formed so close to their host stars[1]. Close to the star in a protoplanetary disk, local feeding zones for planetesimals are small enough that it would be difficult to grow a large enough core ($\sim 10 M_\oplus$). If gravitational instability (i.e. like the collapse of a molecular cloud) plays any role in giant planet formation, the gas conditions close to the star would prevent such a collapse from occurring. Thus, it is thought that these planets must have formed further out and migrated inwards.

There are mechanisms that can cause planets to migrate inwards: disk-planet interactions, planet-planet perturbations and scatterings, and tidal forces from the central star. However, with these mechanisms, migration speeds are expected to increase as a planet approaches its star, so it would be highly unlikely for a Jupiter-like planet to get so close to its star without falling completely onto it. There are two mechanisms that have been proposed to counteract this effect:

- **Tidal torques** from the central star counteract torques from the disk, or the disk torques themselves are reduced when the planet enters a nearly empty zone close to the star. This method does not explain the observed close-in giant planets that do not experience significant tidal forces, which have been observed.
- A **dynamical mechanism** where planet-planet perturbations place planets into high-eccentricity orbits with periapses very close to their star. When the planets approach periapse, extreme tidal torques from the star can then circularize them. A side effect of these interactions is that they can also produce high inclinations via the Kozai–Lidov mechanism, which are observed in a lot of hot Jupiters.

Observational signatures in support of these mechanisms include:

- High inclination angles due to the Kozai–Lidov mechanism
- Orbital resonances due to tidal effects

13. Describe four methods for detecting an exoplanet, and the potential biases in the inferred populations of planets associated with each method. Roughly how many planets have been detected using each?

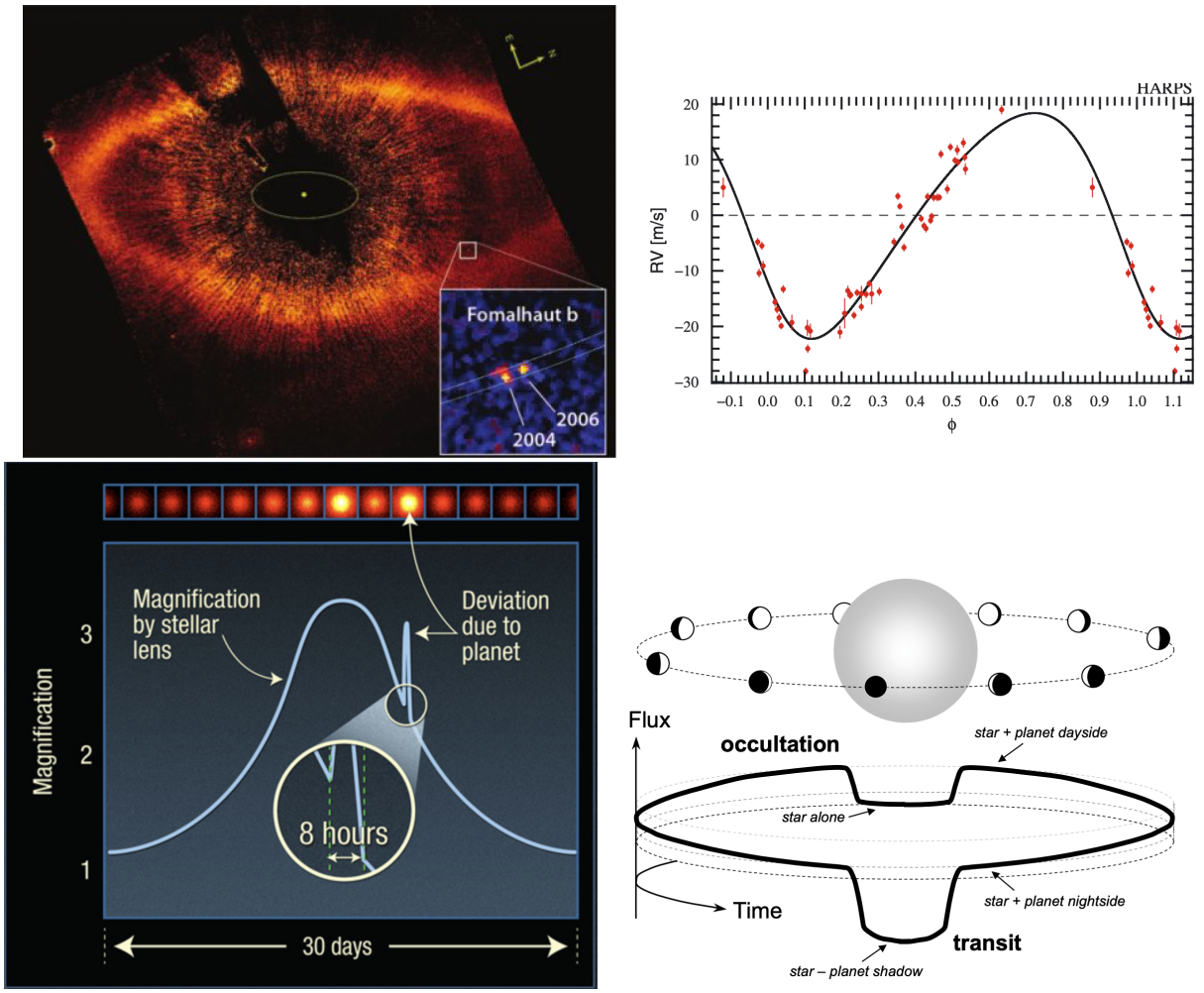


Figure 2.4: *Top Left*: Direct imaging example with a coronagraph on *HST*[12]. *Top Right*: RV curve for an eccentric orbit[13]. *Bottom Left*: Schematic of a microlensing event (Credit: NASA, ESA, and K. Sahu (STScI)). *Bottom Right*: Schematic of a transit light curve[9].

Method 1 - Direct Imaging[3,14]: The most conceptually straightforward method is to directly image the flux from the exoplanet. However, this method is made difficult because the contrast between starlight and planet light is so large, and the separations are so small, that in most cases it's impossible to identify planet light. The contrast can be reduced by utilizing coronagraphs, which block the starlight, interferometry, or by observing at IR wavelengths.

- **∼70** confirmed as of Jan 2024
- Observables: Orbital radius, period, planet temperature
- Biased towards **larger orbital separations**, **larger planet radii**, and **hotter planets** (i.e. more blackbody flux)

Method 2 - Radial Velocities[3,14]: The RV method is done by measuring periodic variations in the radial velocity of a star as it wobbles around its barycenter due to the gravitational influence of the planets orbiting it. The radial velocities are measured from stellar absorption lines using high resolution spectroscopy. The amplitude of these variations is proportional to the mass of the planets and the star (degenerate with the orbital inclination).

- **∼1000** confirmed as of Jan 2024
- Observables: Planet mass (degenerate with inclination), orbital period, orbital eccentricity
- Biased towards **larger planet masses** and **smaller orbital separations**

Method 3 - Transits[3,14]: The transit method is done by measuring the periodic dip in starlight as a planet passes in front of its star and eclipses it. This is achieved by performing photometric measurements of a star over time. The magnitude of the dip (AKA "depth") of the transit is proportional to the planet radius relative to the star radius.

- **∼4000** confirmed as of Jan 2024
- Observables: Planet radius, orbital period, inclination
- Limited to exoplanets with **low enough inclinations** that they pass in front of the star along our line of sight, the chances of which decrease as the distance to the star increases
- Biased towards **larger planet radii** and **smaller orbital separations**

Method 4 - Astrometry[3,14]: The astrometry method uses the reflex motion of the star, similar to the RV method, but instead measures the changes in the position of the star on the sky over time.

- **∼3** confirmed as of Jan 2024
- Observables: Orbital radius, inclination, planet mass
- Limited to only the **closest stars** since the angular size of the exoplanet's orbit decreases with distance to the star, thus also decreasing the amount the star moves in response
- Biased towards **larger orbital separations** since this also increases the star's semimajor axis

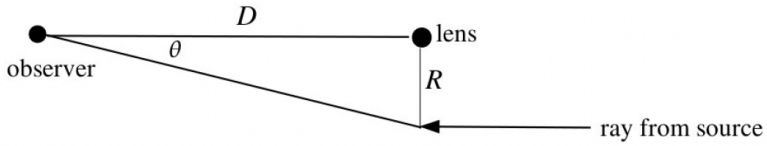


Figure 2.5: A simple example of gravitational lensing geometry

Method 5 - Microlensing[3,14]: When a massive object passes in front of a background star, it acts as a gravitational lens, bending the light from the background star (which is essentially parallel at the location of the lens; see Figure 2.5). The lensing images cannot be resolved at this scale, but there is a characteristic increase in brightness in the light curve that can be measured. If there are planets around the object acting as the lens, they can also act as lenses and introduce structure into the light curve that can be measured. This works best when the planets are close to the Einstein radius.

- **∼200** confirmed as of Jan 2024
- Observables: Orbital radius, planet mass

- Biased towards **larger orbital separations** that are close to the Einstein radius

Method 6 - Pulsar Timing[3,14]: The arrival time of pulsar pulses can deviate from perfect periodicity if the pulsar moves about its barycenter due to the motions of orbiting planets. This is most easily measured around millisecond pulsars since they have the highest sample rate. The first two exoplanets ever discovered utilized this method[15]. The origins of planets around pulsars is unknown, as they should not survive when their star goes supernova.

- ∼7 confirmed as of Jan 2024
- Observables: Orbital period, planet mass (degenerate with inclination)
- Biased towards **larger masses**

Method 7 - Transit Timing Variations (TTVs)[3,14]: The timing of an exoplanet's transit is determined precisely by its period. However, the presence of other gravitating bodies, such as other planets, in a system can introduce perturbations that cause slight variations in the timing of a planet's transits. Even if these other bodies do not themselves transit, their effects on the bodies that do transit can be used to find their masses and periods.

- ∼30 confirmed as of Jan 2024
- Observables: Orbital period, planet mass
- Biased towards **larger masses**

In summary

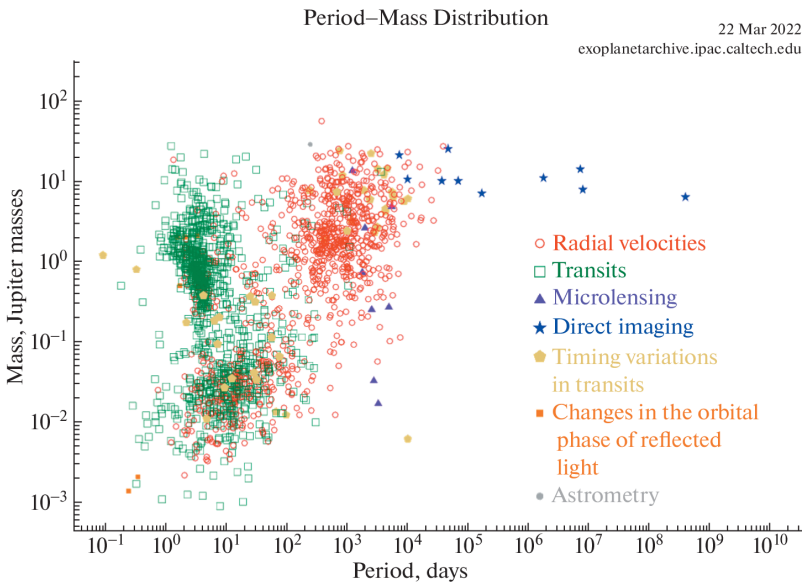


Figure 2.6: Summary of confirmed planets using different detection methods[16].

The **first** confirmed exoplanet seems to depend on who you ask. *Technically* the first discovery was of two rocky exoplanets orbiting the pulsar PSR B1 257+12, using the pulsar timing method, in 1992 by Wolszczan and Frail[17]. However, the first discovered exoplanet around a *main sequence star* was 51 Pegasi b, using the radial velocity method, in 1995 by Mayor and Queloz[18], who won the Nobel Prize for their discovery in 2019.

14. Describe a sequence of observations to identify an exoplanet, to determine that it has a rocky composition, and to measure the chemical constituents of its atmosphere. What have we learned about exoplanet interiors and atmospheres from these observations?

1. **Identification**: Use the transit method to identify an exoplanet. This method has been very successful in identifying exoplanets (∼4000 confirmed detections) and is easy to perform with large time-domain surveys like *TESS*. From these observations we can obtain the planet radius R_p, orbital period P, and orbital inclination i (from the transit duration).
2. **Composition**: We can get a mass for the exoplanet by performing spectroscopic RV follow-up observations. This gives $M_p \sin i$, which in combination with i from the transit gives us M_p unambiguously. Then we can get a bulk density $\rho_p = M_p/(4\pi R_p^3/3)$, which can tell us if the planet has a rocky ($\rho_p \sim 5\ \mathrm{g\,cm^{-3}}$) or gassy ($\rho_p \sim 2\ \mathrm{g\,cm^{-3}}$) composition.
3. **Atmosphere**: We can perform **transmission spectroscopy**. This technique involves taking a spectrum of the planet's atmosphere as it passes in front of its host star and measuring the absorption features in the spectrum. Or one can take transit observations in different filters and measure the difference in transit depth to get broad spectral features ($\Delta\delta^2 \approx 2R_p N_H H/R_*^2$).

What have we learned?[10]

- There is a large population of "super-Earths" and "sub-Neptunes" with masses $\sim$1–20$M_\oplus$ on realtively close-in orbits. These lie close to the detection thresholds, so they are likely very common types of planets.
- Planets tend to follow a pretty tight mass-radius relation (sometimes called the **Chen-Kipping relation**[19]) of the form $R/R_\oplus = (M/M_\oplus)^b$ where $b \sim 0.274$ for $M < M_\oplus$. b decreases for larger planet masses since the effect of compression due to pressure becomes more prominent. The larger pressure gradient means these planets have larger cores and mantles and smaller upper mantles. The radius of the core depends a lot on the equation of state of Fe and silicates which is not well known at the high temperatures and pressures expected in these planet cores.
- The interiors of **solar system** planets can generally be described by 4 structure equations, similar to the 4 equations of stellar structure (see §3.Q30). For planets, the equivalents are: (1) mass conservation, (2) hydrostatic equilibrium, (3) energy transport, and (4) equation of state (which is very much NOT the ideal gas law—we're dealing with solid/plastic materials here in the case of terrestrial planets / giant planet cores). It's nearly impossible currently to do this analysis for exoplanets, where we only have 2 data points (mass and radius).
- Giant planet atmospheres are dominated by molecular hydrogen (H_2), followed by Helium. They have temperatures ranging from 2500 K at birth to 50 K for older exoplanets in wide orbits. There are no internal sources of energy, so after formation the atmosphere cools and shrinks over time (modulo stellar irradiation). Hot Jupiter atmospheres typically sit at 1000–2000 K and have significant variation in day-side and night-side composition, suggesting nonequilibrium chemistry.
- Exoplanet atmospheres *reflect* stellar radiation in the optical/NIR (based on their albedo) and *emit* thermal radiation in the NIR/MIR. A higher albedo can indicate the presence of clouds, and thermal radiation can trace surface temperatures.

- Terrestrial planet atmospheres display a wide variety of characteristics. There are a few different general categories:
    1. Super-Earths ($M \sim 1$–$10 M_\oplus$) with no atmosphere, often found to be tidally locked and with no atmosphere due to escape processes.
    2. Super-Earths with a hot atmosphere ($\gtrsim$1500 K) that have lost H, C, N, O, and S through hydrodynamic escape.
    3. Super-Earths that have outgassed their atmospheres.
    4. Outgassed H-rich atmospheres that have retained against atmospheric escape and include observable signatures of H_2O, CH_4, or CO.
    5. Young, cold, more massive super-Earths that have retained a primordial atmosphere from the protoplanetary disk, dominated by H and He.

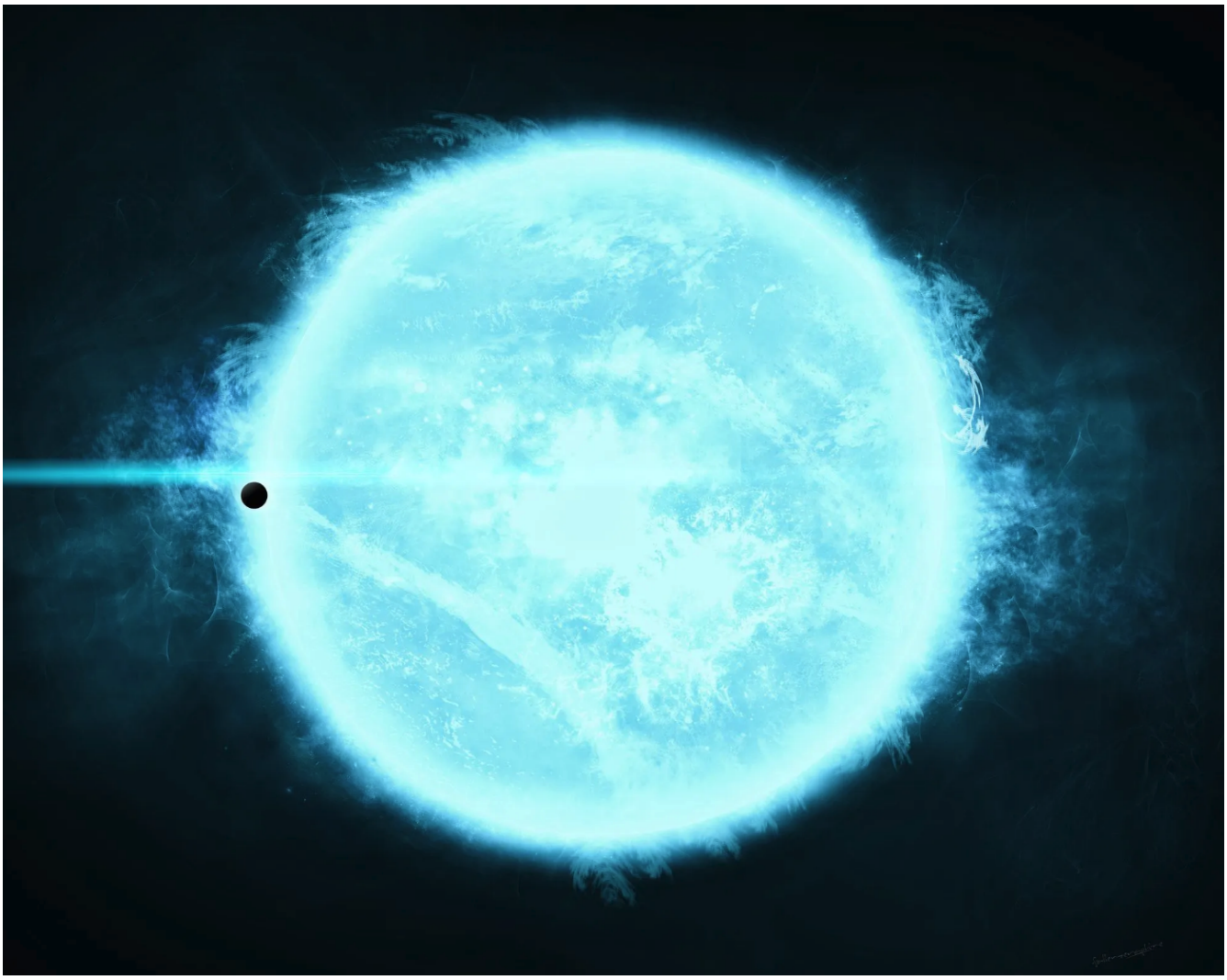
15. Why are M Dwarfs thought to be promising candidates for identifying Earth-sized, temperate (but not necessarily habitable) planets?

M dwarfs have a few properties that make them appealing candidates for detecting an Earth-sized, temperate planet[10]:

- They are less massive, so the RV amplitude $K \propto M_*^{-2/3}$ becomes larger.
- They are relatively cool, making their habitable zones closer in, which in turn leads to an increased likelihood that an Earth-sized planet in the habitable zone will be detectable since $K \propto a^{-1/2}$. The transit method is also biased towards closer-in planets since there is a higher likelihood that their inclination will produce a transit along our line of sight.
- M dwarfs are the most common type of star in the Milky Way ($\sim$75%)

However, the prospects of actually finding life on one of these planets is less promising, since M dwarfs exhibit a lot of stellar activity (flares, strong magnetic fields, etc.) that can affect the atmospheres of nearby planets.

3 Stars and Stellar Evolution



16. Describe the spectral classification scheme for stars: O, B, A, F, G, K, M. What are the characteristic effective temperatures for stars of each class? What are the characteristic absorption lines observed in the spectra of stars of each class? What are the characteristic luminosities for main-sequence stars of each class? What are the approximate masses for stars in each class?

There are a few different important classification systems for stars. The one mentioned in the question is the **Harvard Spectral Classification** system, which, as the name suggests, primarily differentiates stars based on spectral properties.

Type	T_{eff} (K)	M ($M_\odot$)	L ($L_\odot$)	R ($R_\odot$)	Characteristics
O	$\geqslant 33,000$	$\geqslant 16$	$\geqslant 30,000$	$\geqslant 6.6$	Hottest blue-white Strong He II absorption He I absorption Few lines
B	10,000–33,000	2.1–16	25–30,000	1.8–6.6	Hot blue-white Strongest He I absorption (B2) Weak H I absorption
A	7,300–10,000	1.4–2.1	5–25	1.4–1.8	White Strongest H I absorption (A0) Weak Ca II H & K absorption
F	6,000–7,300	1.04–1.4	1.5–5	1.15–1.4	Yellow-white Ca II H & K absorption Weaker H I absorption Neutral metal (Fe I, Na I D) absorption
G	5,300–6,000	0.8–1.04	0.6–1.5	0.96–1.15	Yellow Solar-type spectra Stronger Ca II H & K absorption Stronger metal (Fe I, Na I D) absorption
K	3,900-5,300	0.45–0.8	0.08–0.6	0.7–0.96	Cool orange Strongest Ca II H & K absorption (K0) Dominant metal (Fe I, Na I D) absorption
M	2,300–3,900	0.08–0.45	$\leqslant 0.08$	$\leqslant 0.7$	Cool red Dominant molecular absorption (Particularly TiO and VO) Strong metal (Fe I, Na I D) absorption
L	∼1,300–2,300	$\lesssim 0.08$	$\lesssim 0.08$	$\lesssim 0.7$	Very cool, dark red Stronger in IR than visible Strong molecular absorption (CrH, FeH, H_2O, CO, Na, K, ...)
T	∼550-1,300	$\lesssim 0.08$	$\lesssim 0.08$	$\lesssim 0.7$	Very cool, infrared Strong methane (CH_4) bands Weakening CO bands
Y	$\lesssim 550$	$\lesssim 0.08$	$\lesssim 0.08$	$\lesssim 0.7$	Coolest, infrared Ammonia (NH_3) absorption Jupiters?

A few notes:

- An outdated and iffy acronym that I was taught to memorize this is “Oh Be A Fine Girl/Guy, Kiss Me! (Less Tongue, Yo)” An alternative, less troublesome one I’ve seen is “Oh Be A Fine Goat, Kick Me!”

- Each type can be subdivided further into 10 subtypes ranging from 0–9 (i.e. A0, F5, K9) based on their temperature, with 0 being the hottest and 9 being the coolest.
- The L, T, and Y classes are separated from the rest since they are mostly representative of brown dwarfs and substellar objects. These have an unresolvable overlap in luminosity/mass/radius since brown dwarfs aren't fusing and thus cool through different spectral classes over time.
- The table above is taken mostly from 11 and 20.
- For clarification on the Ca II "H & K" and Na I "D" line notation, see the appendix section on line notations (§12.7.4).

In addition to the spectral classifications, there is also the **Morgan-Keenan (MK) Luminosity Classification** system, which divides stars mainly based on their luminosity, which is highly correlated with their evolutionary stage[11,20].

Type	Description
Ia-O	Extreme, luminous supergiants
Ia	Luminous supergiants
Iab	Intermediate-luminosity supergiants
Ib	Less luminous supergiants
II	Bright giants
III	Giants
IV	Subgiants
V	Dwarfs (main sequence stars)
VI, sd	Sub-dwarfs
VII, D	White dwarfs

More notes:

- These classes are determined observationally by looking at the widths of absorption lines in the star's spectrum. Stars with larger radii have a smaller pressure broadening effect, so their lines are narrower.
- Once again, we have some classses (VI and VII) separated from the others because they are on the borderline between being actual stars vs. something closely related to stars (in this case, stellar remnants)
- The Sun is a **G2V** star with a $T_{\text{eff}} \approx 5,700$ K.

17. Sketch the HR diagrams for a typical globular cluster and open cluster. Identify the various observed populations and interpret them on the basis of stellar evolution theory. How would you go about constructing lines of constant stellar radius on the diagram?

The **Hertzsprung-Russell (HR) Diagram** shows stars in temperature-luminosity space, making it a useful tool for differentiating evolutionary stages of stars. Note that due to traditions, the x-axis is usually flipped so hotter stars are on the left and cooler stars are on the right (this is because this maintains the same spatial relationship as plotting the $B-V$ color on the x-axis, which puts redder stars on the right and bluer stars on the left). See what the HR diagram in general looks like, where stars of different spectral and luminosity classes occupy different regions, in Figure 3.1.

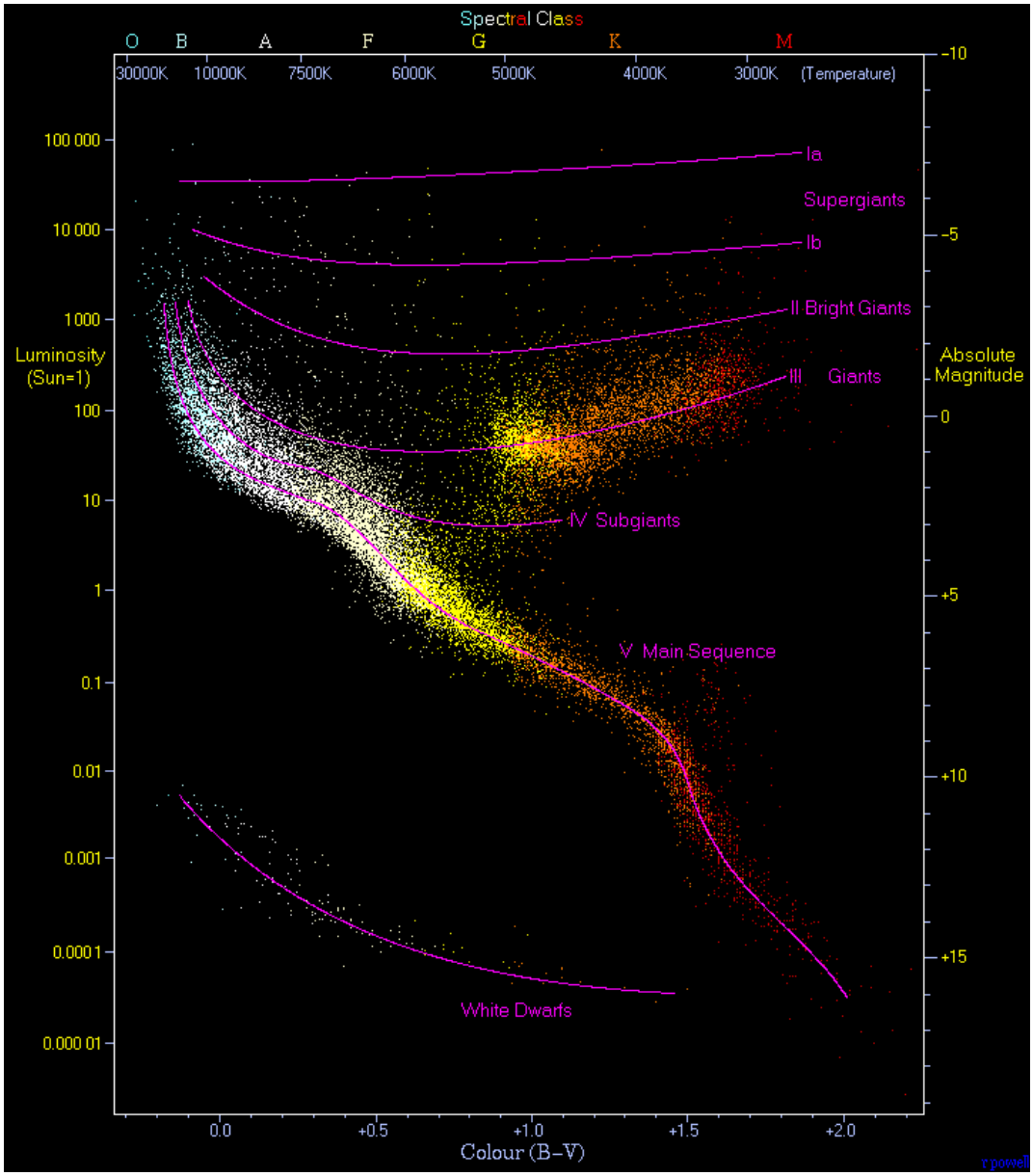


Figure 3.1: The HR diagram for a general population of stars throughout the whole Galaxy. Just used an illustrative example of what the HR diagram looks like[21].

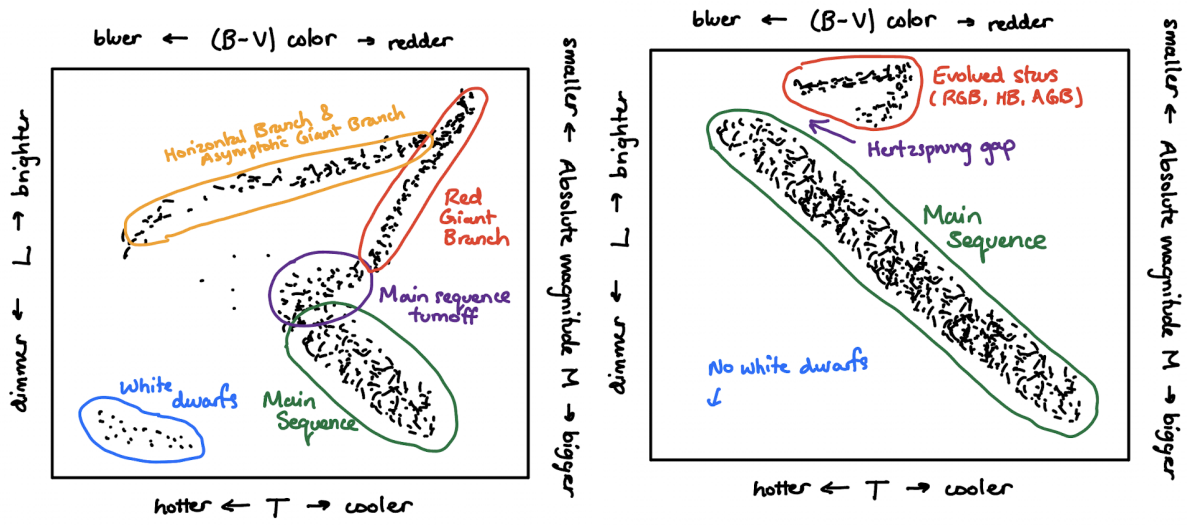


Figure 3.2: The HR diagrams of a globular cluster (left) and open cluster (right) labeled with different evolutionary stages.

Globular Clusters

Globular clusters are **old**, so the evolution of stars on the HR diagram becomes important for studying them. Main sequence stars tend to fall along a diagonal line, maintaining a constant temperature-luminosity relationship. As they age, they inflate into red giants, becoming both brighter and cooler at the same time, moving to the top-right of the diagram. Hotter stars exhaust their fuel quicker, so they evolve faster than cooler stars. Thus, for a globular cluster where all of the stars formed at around the same time, we can see a distinct turn-off in the main sequence where the hotter stars have started evolving into red giants while the cooler stars are still on the main sequence. Therefore, the HR diagram of a globular cluster looks something like the left panel of Figure 3.2. The location of the turn-off point can be used to estimate the age of the globular cluster, since the older the cluster is, the further down the main sequence the turn-off point moves.

Open Clusters

Open clusters are **young**, so their main sequence turn off occurs much higher up—only the most massive/luminous stars have begun evolving off the main sequence. There is a lack of white dwarfs, since the evolved massive stars turn into neutron stars or black holes. The Hertzsprung gap is also apparent here since the subgiant branch evolution is very quick for these stars. See Figure 3.2.

Lines of constant radius

Stars are blackbodies, so we can use the **Stefan-Boltzmann Law** at the surface of the star to relate the luminosity to the temperature and radius:

$$F(R) = \frac{L}{4\pi R^2} = \sigma_{\text{SB}} T_{\text{eff}}^4 \quad \longrightarrow \quad \boxed{\log L = \log(4\pi\sigma_{\text{SB}}) + 2\log R + 4\log T_{\text{eff}}} \tag{3.1}$$

Thus we have a power law with $L \propto R^2 T_{\text{eff}}^4$, appearing on the log-log HR diagram as a straight line with a downwards slope (since T increases to the left). The main sequence roughly (but not exactly) follows a line of constant $R = 1 R_\odot$, which is why the dynamic range of the radii of different stellar classes in §3.Q16 is much smaller than some of the other properties ($\sim$ 1 order of magnitude, as opposed to L which varies by $\sim$ 6 orders of magnitude).

18. From a physics perspective, how does the quantity $(B-V)$ help to determine a star's effective (surface) temperature?

The radiation from a star is approximately blackbody radiation, given by Planck's Law (see §12.11 for a derivation):

$$\boxed{B_\nu(\nu,T)=\frac{2h\nu^3}{c^2}\frac{1}{\exp(h\nu/kT)-1}} \tag{3.2}$$

Where B_ν is a specific intensity (in units of $\mathrm{erg\,s^{-1}\,cm^{-2}\,Hz^{-1}\,sr^{-1}}$). Or, we could write it in terms of wavelength:

$$\boxed{B_\lambda(\lambda,T)=\frac{2hc^2}{\lambda^5}\frac{1}{\exp(hc/\lambda kT)-1}} \tag{3.3}$$

For derivations of these formulas, see §12.11.14 in the appendix. See what this looks like for different temperatures in Figure 3.3.

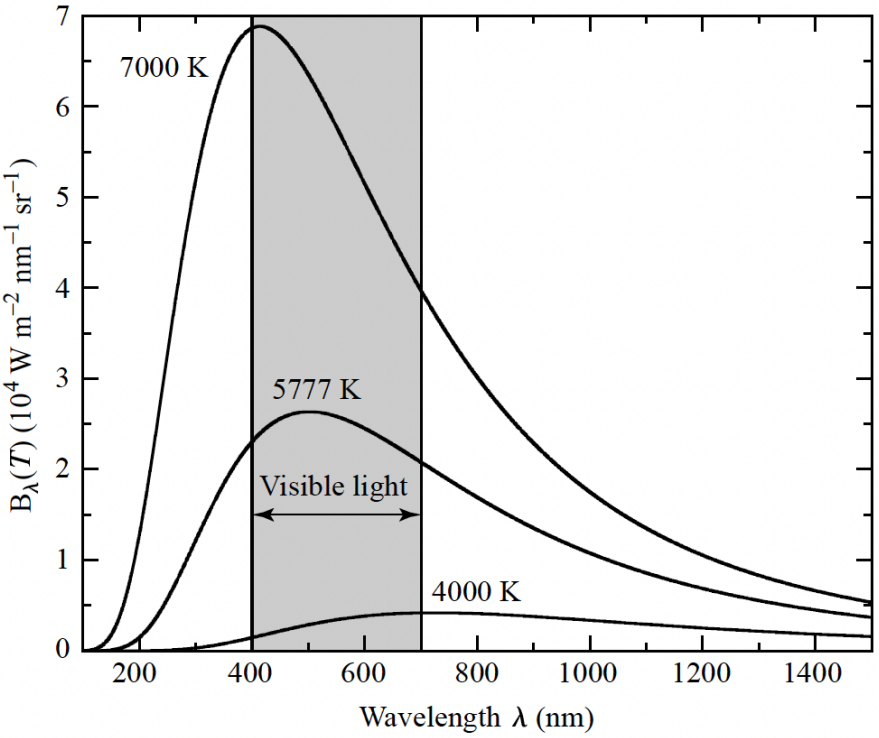


Figure 3.3: The Planck function at different temperatures[11].

Then, if we observe a star at two filters in the blue B and visual V bands, which have central wavelengths of $\sim 4400\,$Å and $\sim 5500\,$Å, respectively (see §12.6), the quantity $B-V$ is effectively a crude measurement of the **slope** of the blackbody spectrum between these two wavelengths:

$$B-V\equiv m_B-m_V=-2.5\log\left(\frac{F_B}{F_V}\right) \tag{3.4}$$

The slope strongly depends on the temperature. Notice in Figure 3.3 that the hotter blackbodies are much brighter in the blue band than the visual band, leading to a *smaller* $B-V$ color (because remember, smaller magnitudes correspond to brighter fluxes, §12.11). Conversely, cooler blackbodies have a shallower slope, leading to a a *larger* $B-V$ color. Thus, T_{eff} and $B-V$ have an inverse relationship! This is why the temperature axis on the HR diagram has to be inverted—it was originally plotted with $B-V$ on the x axis, so people got used to the hotter (bluer) stars being on the left[22].

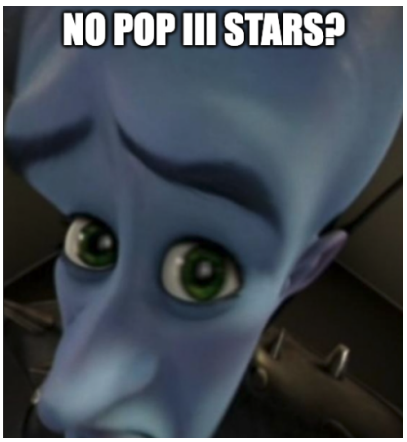
Note: the relationship between the temperature T and the peak wavelength λ_{peak} is known as **Wein's displacement law** (or just simply Wein's law), and has the form

$$\boxed{\lambda_{\text{peak}} = \frac{2898\,\mu\text{m K}}{T}} \tag{3.5}$$

19. What are stellar populations? Give several examples of Pop. I and Pop. II objects in our galaxy.

Stellar populations are a general categorization of stars based on their properties, originally introduced by Baade (1944)[23] to distinguish between the two observed groups of stars in the Milky Way[24].

1. **Population I**
    - Metal-rich
    - Bright O/B stars
    - Young—thought to have formed more recently from metal-enriched gas
    - Typically found in the disks of spiral galaxies
    - Examples (stars): The Sun (Sol), Vega, Altair, Deneb, Polaris Aa, Betelgeuse, Arcturus
    - Examples (open clusters): Pleiades, Hyades, Butterfly Cluster, M7
2. **Population II**
    - Metal-poor
    - No O/B stars
    - Old—thought to have formed earlier than Pop I stars before the ISM had been enriched by the deaths of stars
    - Typically found in the galactic bulge and halo (globular clusters)
    - Examples (globular clusters): M5, M13 (Hercules Globular Cluster), M22, ω Centauri
3. **Population III** (theoretical—not yet observed)
    - No metals at all
    - The oldest populations that would have formed at a $z \sim 20$
    - Hard to detect because we cannot resolve stars this far away
    - Examples: lol



Like many things in astrophysics, the numbering of the populations seems to be reversed from what we would expect based on their relative ages (III is the first formed, I is the latest formed).

20. What is the "mass-function" of a binary star system and how is it determined?

Consider two stars orbiting in a binary system, where some component of their relative motion is along our line of sight. Then, they will have some radial velocity v_r, which we can relate to their masses using the **binary mass function**.

To derive this, let's start by establishing some quantities. Star 1 has a mass M_1 and position $\mathbf{r}_1$ and star 2 has a mass M_2 and position $\mathbf{r}_2$. We can represent the 2-body motion as a reduced 1-body problem where a reduced mass orbits the center of mass of the system. We do this by setting the center of mass position $\mathbf{R}$ at the origin, and define the following quantities:

$$M \equiv M_1 + M_2 \tag{3.6}$$

$$\mu \equiv \frac{M_1 M_2}{M_1 + M_2} \tag{3.7}$$

$$\mathbf{R} \equiv \frac{M_1 \mathbf{r}_1 + M_2 \mathbf{r}_2}{M} = 0 \tag{3.8}$$

$$\mathbf{r} \equiv \mathbf{r}_2 - \mathbf{r}_1 \tag{3.9}$$

See a diagram in Figure 3.4. Then we can relate the positions of each star to the relative position vector $\mathbf{r}$:

$$\text{from (3.8)} \longrightarrow \mathbf{r}_2 = -\frac{M_1}{M_2}\mathbf{r}_1 \tag{3.10}$$

$$\text{from (3.9)} \longrightarrow \mathbf{r} = -\mathbf{r}_1(1 + M_1/M_2) \longrightarrow \mathbf{r}_1 = -\frac{M_2}{M}\mathbf{r} \tag{3.11}$$

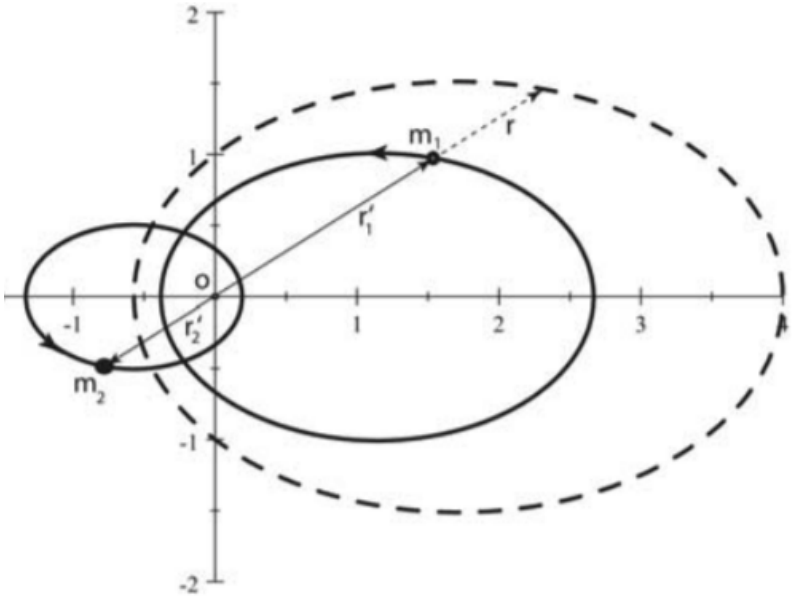


Figure 3.4: Reducing a 2-body orbit into an equivalent 1-body problem, with labeled quantities[25].

Circular orbits

Assuming circular orbits ($e = 0$) makes things easier, so let's examine that case first. We can use a force balance equation to obtain:

$$F_{\text{cf}} = F_g \quad \longrightarrow \quad \mu\frac{v^2}{r} = \frac{GM\mu}{r^2} \quad \longrightarrow \quad v^2 = \frac{GM}{a} \tag{3.12}$$

Then we convert to the velocity of star 1 by differentiating (3.11) and multiplying by $\sin i$ (where i is the inclination angle between the orbital plane and the plane of the sky) to get the maximum/minimum radial velocity when the star is moving directly towards/away from us:

$$v_{r,1} = \pm\frac{M_2}{M}v\sin i = \pm\frac{M_2}{M}\sqrt{\frac{GM}{a}}\sin i \tag{3.13}$$

To get the radial velocity amplitude K, we take $(\max(v_{r,1}) - \min(v_{r,1}))/2$ and use Kepler's 3rd law to get:

$$\boxed{K = \frac{M_2}{M}\sqrt{\frac{GM}{a}}\sin i = \left(\frac{2\pi G}{P}\right)^{1/3}\frac{M_2\sin i}{(M_1+M_2)^{2/3}}} \tag{3.14}$$

Which shows that $K \propto M_2 \sin i$, the mass of the second star degenerate with the inclination angle.

Eccentric orbits

Considering eccentric orbits requires a more careful examination. I'm fairly certain this is not something that is necessary to know for the exam. Nevertheless, I figured I would include a derivation here for completeness's sake.

Instead of just worrying about the minimum/maximum $v_{r,1}$, which is less obvious in the eccentric case than it was in the circular case, let's derive a full expression for $v_{r,1}$ as a function of the true anomaly ϕ. We'll work in a coordinate system where the orbit is in the x-y plane, in which case

$$x = r\cos\phi \quad \longrightarrow \quad v_x = \dot{r}\cos\phi - r\dot{\phi}\sin\phi \tag{3.15}$$

$$y = r\sin\phi \quad \longrightarrow \quad v_y = \dot{r}\sin\phi + r\dot{\phi}\cos\phi \tag{3.16}$$

The unit vector pointing from our line of sight to the barycenter is (see Figure 3.5):

$$\hat{\mathbf{s}} = \sin i\sin\omega\hat{\mathbf{x}} + \sin i\cos\omega\hat{\mathbf{y}} + \cos i\hat{\mathbf{z}} \tag{3.17}$$

Notes: $\hat{\mathbf{x}}$ is the direction to periapse, $\hat{\mathbf{z}}$ is the normal to the orbital plane, and $\hat{\mathbf{s}}$ is the normal to the reference plane. The inclination i is the angle between the orbital plane and the reference plane, and the argument of periapse ω is the angle from the ascending node to the periapse.

Therefore, the velocity of star 1, which is $\mathbf{v}_1 = (-M_2/M)(v_x, v_y, 0)$, corresponds to a radial velocity

$$v_{r,1} = \mathbf{v_1}\cdot\hat{\mathbf{s}} = -\frac{M_2}{M}(v_x\sin\omega + v_y\cos\omega)\sin i \tag{3.18}$$

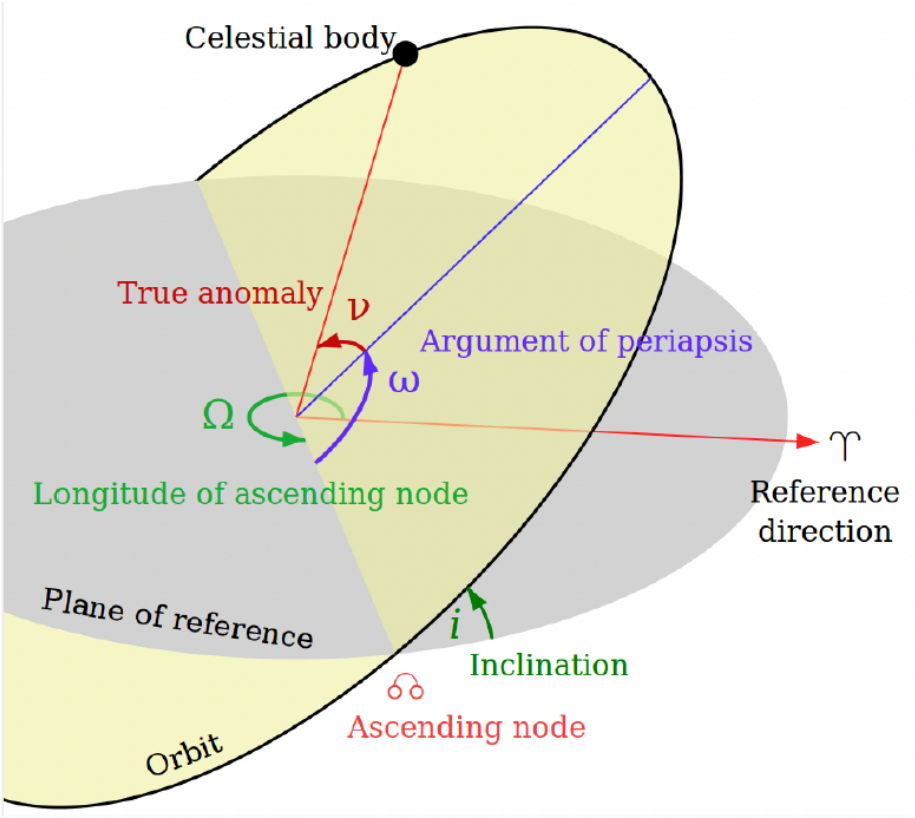


Figure 3.5: Orbital elements (line of sight is looking down from above, i.e. plane of reference is the plane of the sky)

Differentiating the ellipse equation (1.4) and using conservation of angular momentum:

$$r(\phi) = \frac{\ell^2/GM\mu^2}{1+e\cos\phi} \longrightarrow (\ell^2/GM\mu^2)r^{-1} = 1 + e\cos\phi \tag{3.19}$$

$$\longrightarrow -(\ell^2/GM\mu^2)r^{-2}\dot{r} = e\dot{\phi}\sin\phi \tag{3.20}$$

$$\frac{1}{2}r^2\dot{\phi} = \frac{\ell}{2\mu} \longrightarrow \dot{\phi} = \frac{\ell}{\mu r^2} \tag{3.21}$$

$$\longrightarrow \dot{r} = \frac{GM\mu}{\ell}e\sin\phi \tag{3.22}$$

$$\longrightarrow r\dot{\phi} = \frac{GM\mu}{\ell}(1+e\cos\phi) \tag{3.23}$$

Thus, plugging in $\dot{r}$ and $r\dot{\phi}$ to our v_x and v_y, and then plugging those into $v_{r,1}$:

$$v_x = -\frac{GM\mu}{\ell}\sin\phi \quad , \quad v_y = \frac{GM\mu}{\ell}(e+\cos\phi) \tag{3.24}$$

$$v_{r,1}(\phi) = -\frac{M_2}{M}\sqrt{\frac{GM}{a(1-e^2)}}(\cos(\phi+\omega)+e\cos\omega)\sin i \tag{3.25}$$

Notice the $e\cos\omega$ term is just a constant, so taking the difference of the max and min as before gives

us the same result, just with an extra factor of $1/\sqrt{1-e^2}$. Then we use Kepler's 3rd law, like before, to obtain:

$$\boxed{K = \left(\frac{2\pi G}{P}\right)^{1/3} \frac{M_2 \sin i}{(M_1 + M_2)^{2/3}} \frac{1}{\sqrt{1-e^2}}} \tag{3.26}$$

This is the full eccentric binary mass function[3].

How is it determined?

We can get spectroscopic measurements of star 1 over time, measuring the Doppler shifts of the spectral lines, to get a radial velocity curve, which should be described by $v_{r,1}(\phi)$. This allows us to get an estimate of the total mass $M_1 + M_2$ and the secondary mass (degenerate with the inclination angle) $M_2 \sin i$. The shape of the radial velocity curve determines the eccentricity of the orbit, with circular orbits being perfectly sinusoidal. An example of an eccentric radial velocity curve is shown in the top-right panel of Figure 2.4. A cool animation for this is shown here, where 2 frames at opposite peaks have been extracted and shown in Figure 3.6.

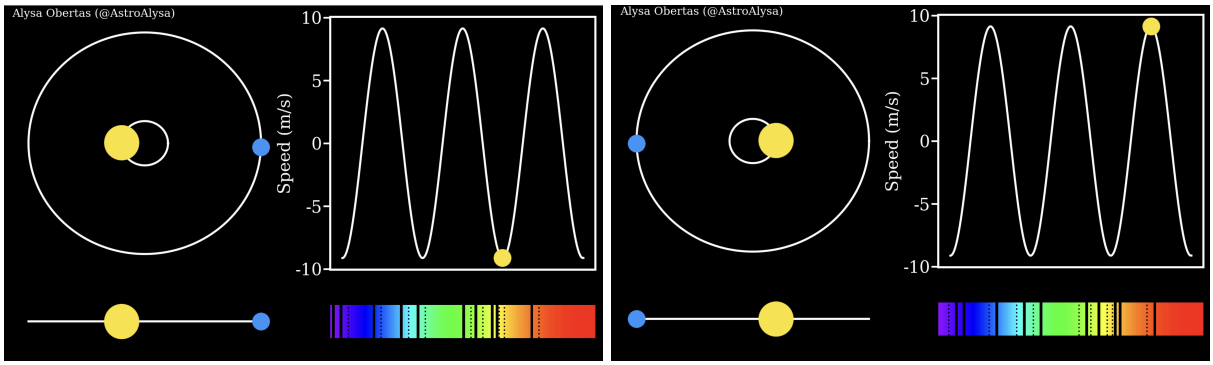


Figure 3.6: The binary orbit of 2 bodies, where the top-left shows a top-down view of the orbit, the bottom-left shows an edge-on view of the orbit, the top-right shows the radial velocity curve over time, and the bottom-right shows the Doppler shift of the spectral lines. Credit: Alysa Obertas (@AstroAlysa)

21. What is meant by the "gravothermal collapse" of a globular cluster, and what can save the cluster from complete collapse?

For a purely self-gravitating system (like a globular cluster), if we use the **virial theorem** (see §12.9.2 for a derivation)

$$\boxed{2K + U = 0} \quad \longrightarrow \quad E = K + U = -K \tag{3.27}$$

And then we treat the system as an ideal gas with temperature T and N particles (i.e. stars) such that the total kinetic energy can be written

$$K = \frac{3}{2} NkT \tag{3.28}$$

Then we notice that the heat capacity is negative

$$\frac{\partial E}{\partial T} = -\frac{3}{2} Nk \tag{3.29}$$

In other words, as the system loses energy, the temperature gets *hotter*. This should make sense since we're talking about *dynamical* heat—as the system loses energy, stars' orbits must fall in closer to the gravitational center, which causes them to start orbiting faster.

With this in mind, the **gravothermal collapse** of a globular cluster can take place as follows. Consider the cluster to be composed of 2 components: a central core, and a halo surrounding the core.

1. The core starts out hotter than the halo
2. Heat flows outwards from the core into the halo (e.g. via two-body scattering)
3. The core loses energy, which causes its temperature to increase due to its negative heat capacity
4. The increased temperature of the core causes the energy loss rate to increase
5. This causes runaway acceleration of the collapse of the core

In reality, the core and the halo are not physically separate components, but these kinds of instabilities *have* been observed in real systems[26].

To save the cluster from complete collapse, one needs some mechanism to inject energy into the core to offset the energy losses into the halo. This can be accomplished through scatterings involving binary systems[26].

1. When a binary system scatters with a third star, the binary system loses energy and becomes more tightly bound, while the lone star increases in energy (to conserve total energy)
2. Thus, energy can be injected into the core through these interactions if there are enough binary systems
3. If there aren't any binary systems, they can be created as the core collapses and 3-body encounters become common, which can end up leaving 2 stars bound while the third one escapes

22. Describe the various evolutionary phases of a low-mass ($1\,M_\odot$) star and those of a high-mass (e.g. $12\,M_\odot$) star. Show the corresponding evolutionary tracks on an HR diagram

First we'll examine the case of the low-mass star. The following evolutionary track will work for any star between 0.08–1.5 $M_\odot$. The timescales shown next to each step are for a [1 $M_\odot$ /12 $M_\odot$] star.

1. **Molecular Cloud Collapse** [∼10 kyr / 10 kyr]

   An optically thin cloud of molecular hydrogen (H_2) gets perturbed by inhomogeneities in the cloud or by nearby supernovae. It will collapse if it is bigger than the Jeans length λ_J

   $$\lambda_J = c_s\sqrt{\frac{\pi}{G\rho}} \tag{3.30}$$

   (see §3.Q22). The sound speed, assuming an ideal gas, is

   $$c_s^2 = \frac{\partial P}{\partial \rho} = \frac{\partial}{\partial \rho}\left(\frac{\rho}{\mu m_p}kT\right) = \frac{kT}{\mu m_p} \tag{3.31}$$

   Thus

   $$\lambda_J = \sqrt{\frac{\pi kT}{G\rho\mu m_p}} \tag{3.32}$$

   Note that this ignores the effects of rotation, external pressure from the ISM, and magnetic fields. The collapse is typically on timescales $t \sim 1/\sqrt{G\rho} \sim 10^4$ yr.

2. **Protostar** [-]

   As the cloud collapses, it becomes optically thick (the opacity at this stage is dominated by H^- ions). A protostar forms at the center with a radius of $R \sim 100R_\odot$ and temperature of $T \sim 10^{3.5}$ K, while the rest of the gas will collapse into a disk due to conservation of angular momentum. The protostar itself also continues collapsing, which converts gravitational energy into thermal energy, and thus outwards pressure. At this point, the dominant energy transport mechanism is convection (as opposed to radiation) because it is optically thicc. In these initial stages, the temperature gradient between the core and the surface of the protostar is so large that it emits energy through convection faster than it gains energy through contraction, so we are *not* in hydrostatic equilibrium. The high energy flux from the convection rapidly heats up the outer layers of the protostar faster than the gravitational contraction heats up the core, which decreases the temperature gradient to adiabatic levels, moves the star left on the HR diagram (increased T_{eff}), and establishes hydrostatic equilibrium. At this point, the protostar lands on the **Hayashi track**.

3. **The Hayashi Track** [∼10 Myr / 100 kyr]

   This is defined as the boundary of the "forbidden region" of the HR diagram where temperature gradients are super-adiabatic, causing instability and preventing protostars from being in hydrostatic equilibrium. Once a star reaches the Hayashi Track, it will contract on **Kelvin-Helmholtz timescales** at a roughly constant T_{eff}:

   $$t_{\text{KH}} = \frac{GM^2/R}{L} \tag{3.33}$$

This means moving roughly straight down on the HR diagram. This track is shown in Figure 3.7.

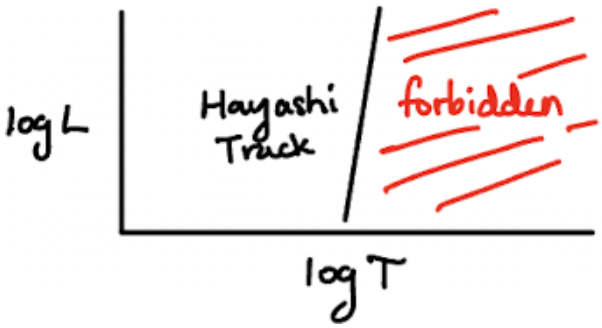


Figure 3.7: The Hayashi track separating the permitted and forbidden regions due to hydrostatic equilibrium.

Why is the T_{eff} almost constant on the Hayashi track? We can derive the actual form of the Hayashi track if we assume a protostar is fully convective at this stage, in which case we have an adiabatic equation of state, along with the ideal gas law

$$P = K\rho^{\gamma} = \frac{\rho}{\mu m_p} kT \tag{3.34}$$

where $\gamma = 5/3$ for an ideal gas. Since K is a constant, we can get it in terms of the total mass M and radius R of the star by noting that the *average* density $\rho_{\text{avg}} \propto MR^{-3}$ and the *central* pressure $P_c \propto M^2R^{-4}$. The *central* density is related to the average density by a constant, so we can use the same proportionality:

$$K = P\rho^{-\gamma} \propto M^{2-\gamma}R^{3\gamma-4} \tag{3.35}$$

Then, we can also write K in terms of P and T:

$$K = P\rho^{-\gamma} = P(PT^{-1})^{-\gamma} = P^{1-\gamma}T^{\gamma} \tag{3.36}$$

Equating these two reveals

$$P \propto (M^{2-\gamma}R^{3\gamma-4}T^{-\gamma})^{1/(1-\gamma)} = M^{-1/2}R^{-3/2}T^{5/2} \tag{3.37}$$

Now we want to eliminate P. On the Hayashi track, the protostar should be in hydrostatic equilibrium, in which case we can use

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{GM(r)\rho(r)}{r^2} \longrightarrow \frac{\mathrm{d}P}{\mathrm{d}\tau} = \frac{GM(r)}{r^2}\frac{1}{\kappa_R} \longrightarrow P_0 = \frac{GM(r)}{r^2}\frac{\tau_0}{\kappa_R} \tag{3.38}$$

The opacity at this stage is approximately given by

$$\kappa_R \propto PT^3 \tag{3.39}$$

Therefore we have

$$P \propto M^{1/2}R^{-1}T^{-3/2} \tag{3.40}$$

One more relationship we consider is that of the luminosity

$$L = 4\pi R^2 \sigma_{\rm SB} T^4 \propto R^2 T^4 \tag{3.41}$$

Finally, equating (3.37) with (3.40) and plugging in the luminosity gives

$$L \propto T^{20} M^{-4} \tag{3.42}$$

Notice the incredibly steep dependence on (effective) temperature! Thus, the Hayashi track is a nearly vertical line on the HR diagram ($\log L$ vs. $\log T$). Can think of it like this: if T is nearly constant as the star contracts, then L must decrease like R^2.

During the contraction along the Hayashi track, the central density ρ_c and temperature T_c increase, while the opacity κ_R decreases (due to ionization of H^-). This allows the dominant energy transport mechanism to transition from convection to radiation, allowing the protostar to move onto the **Henyey Track**.

4. **The Henyey Track** [∼100 Myr / 1 Myr]

   Once the protostar develops a radiative zone, energy can escape much more easily than before. This causes the luminosity to slowly and steadily rise as thermal radiation escapes, but the change is slow enough that we can consider it roughly constant. Thus, as the protostar continues gravitationally contracting, its effective temperature increases as

$$L \propto R^2 T_{\rm eff}^4 \quad \longrightarrow \quad T_{\rm eff} \propto R^{-1/2} \tag{3.43}$$

   The opacity is now described by a Kramer's law:

$$\kappa_R \propto \rho T^{-7/2} \tag{3.44}$$

   On the HR diagram, the protostar moves horizontally to the left until the core temperature becomes hot enough that it can start fusing Hydrogen. At this point, it enters the **Zero-Age Main Sequence** and is officially considered a star. This is shown in Figure 3.8.

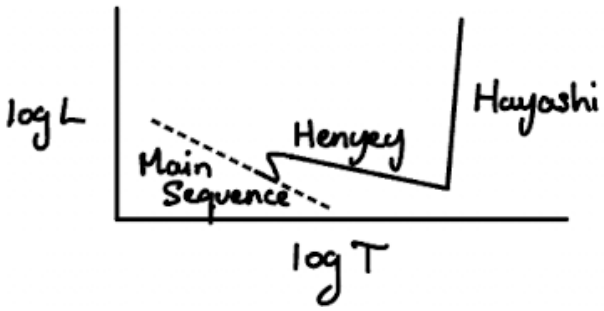


Figure 3.8: The Henyey track connecting the Hayashi track to the main sequence in an almost horizontal line.

   More massive protostars can reach burning temperatures in the core much more quickly, so they leave the Hayashi and Henyey tracks faster than low mass protostars.

5. **The Main Sequence** [∼10 Gyr / 10 Myr]

   On the main sequence, the star is fusing Hydrogen in its core. When it initially enters the main sequence, it is at a point called the **Zero-Age Main Sequence (ZAMS)**. During its time on the

main sequence, it barely moves, maintaining a constant L and T. Thus, at the end of its main sequence lifetime, its location is nearly the same as it was at the beginning. Nevertheless, it is still sometimes called the **Terminal-Age Main Sequence (TAMS)** to differentiate it.

How long does a start spend on the main sequence? See §3.Q33 for how this is calculated. A typical lifetime for a G-type star like the Sun is 10 Gyr, with cooler stars lasting much longer and hotter stars lasting much shorter. In all stars though, main sequence is the longest evolutionary stage by far. After this point, the star runs out of Hydrogen fuel to fuse, and it enters the **subgiant branch**.

Less massive stars primarily fuse through the pp chain and have an inner radiative layer surrounded by an outer convective layer. More massive stars, on the other hand, primarily fuse through the CNO cycle and have an inner convective layer surrounded by an outer radiative layer. More massive stars also have shorter main sequence lifetimes due to their higher luminosities, and they have significant mass loss throughout all stages of evolution, even on the main sequence.

6. **Subgiant Branch (SGB)** [∼2 Gyr / kyr]

   Once the Hydrogen fuel starts to run out, we have a core of inert helium (the core is not hot enough to fuse helium...yet) surrounded by a shell that is still fusing Hydrogen and depositing more helium "ash" into the core. The core is now isothermal, with the stellar structure equations

$$\frac{\mathrm{d}L}{\mathrm{d}r} = 4\pi r^2 \rho \varepsilon_m = 0 \tag{3.45}$$

$$\frac{\mathrm{d}T}{\mathrm{d}r} = -\frac{3\kappa_R \rho L(r)}{16\pi a c T^3 r^2} = 0 \tag{3.46}$$

   With no more nuclear fusion in the core, there is no longer sufficient pressure support against gravity, so it can only handle so much mass before it starts to contract and get hotter. Once it does, this releases gravitational energy that propagates outwards and causes the outer layers of the star to inflate. So, while the core contracts, the outer layers expand. This causes the surface (effective) temperature to get cooler, and the star to get redder. The radius also starts to increase, which increases the luminosity ($L \propto R^2$), so the star moves up and to the right on the HR diagram. A slice of what the star now looks like as a function of radius is shown in Figure 3.9.

   We can find exactly when the core starts to contract by using hydrostatic equilibrium, multiplying by $4\pi r^3$ and integrating up to the core radius r_c

$$\int_0^{r_c} \mathrm{d}r\, 4\pi r^3 \frac{\mathrm{d}P}{\mathrm{d}r} = -\int_0^{r_c} \mathrm{d}r\, 4\pi r^3 \frac{GM(r)\rho}{r^2} \tag{3.47}$$

   The LHS can be integrated by parts, and RHS can be done if we assume we have a uniform sphere such that $M(r) = (4/3)\pi r^3 \rho$

$$4\pi r^3 P(r)\Big|_0^{r_c} - 3\int_0^{r_c} \mathrm{d}r\, 4\pi r^2 P(r) = -\frac{16}{3} G\pi^2 \rho^2 \int_0^{r_c} \mathrm{d}r\, r^4 \tag{3.48}$$

$$4\pi r_c^3 P(r_c) - \frac{4\pi \rho k T r_c^3}{\mu m_p} = -\frac{16}{15} G\pi^2 \rho^2 r_c^5 \tag{3.49}$$

   Where we used the ideal gas law, assuming ρ and T are constant, to solve the integral over

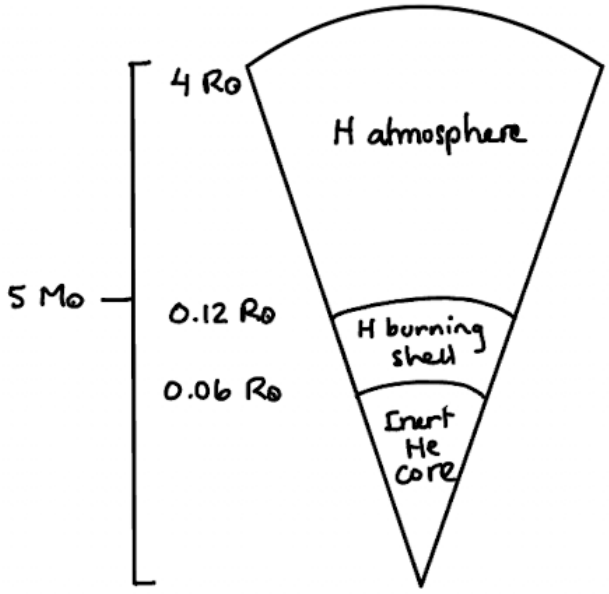


Figure 3.9: A slice of a star on the subgiant and red giant branches, showing the different layers inside the star and their relative radii for a 5 solar mass star.

$P(r)$. Then, the pressure at the boundary of the core is

$$P(r_c) = \frac{3kT_c m_c}{4\pi\mu m_p r_c^3} - \frac{\alpha G m_c^2}{4\pi r_c^4} \tag{3.50}$$

where T_c, m_c, and r_c are the temperature, mass, and radius of the core, and α is a constant equal to 3/5 for a uniform sphere. The first term originates from thermal energy while the second term is from gravitational energy. We can find the core radius r_c by finding where $\mathrm{d}P/\mathrm{d}r$ goes to 0, which allows us to eliminate r_c in favor of the other constants. Skipping the math, this ends up getting a proportionality of

$$P_c \propto \frac{T_c^4}{\mu_c^4 m_c^2} \tag{3.51}$$

Our limiting condition for the core can now be examined by looking at these values just below and just above the core radius. For stability, we need $P_c(r_c) \geqslant P_{\mathrm{env}}(r_c)$, which corresponds to

$$\frac{m_c}{M} \lesssim 0.37\left(\frac{\mu_{\mathrm{env}}}{\mu_C}\right)^2 \longrightarrow \frac{m_c}{M} \lesssim 0.1 \tag{3.52}$$

This is the **Schönberg-Chandrasekhar Limit**, and it means that once the inert helium core exceeds $\sim 10\%$ of the star's total mass, the core will begin collapsing gravitationally on Kelvin-Helmholtz timescales, as discussed at the beginning of this section.

Note: If the star has a mass $M \lesssim 2\,M_\odot$, this limit does not apply because electron degeneracy pressure can kick in and prevent the core from contracting before reaching the Schönberg-Chandrasekhar limit. But degenerate cores have $R \propto M^{-1/3}$, so they will still contract, just more slowly than non-degenerate cores. This is why the subgiant branch lasts quite long for

low-mass stars (up to a couple Gyrs), but can be very short (thousands of years) for high-mass stars. This causes an apparent gap between the main sequence and red giant branch called the **Hertzsprung gap**.

7. **Red Giant Branch (RGB)** [∼500 Myr / 1 Myr]

   As the T_{eff} at the surface of the star continues decreasing, H^- opacity starts becoming important again, and the optically thick regime causes the star to develop an outer convective layer. It eventually hits the Hayashi track again, at which point it stops cooling because any cooler temperature would create a super-adiabatic temperature gradient causing it to leave hydrostatic equilibrium. The core is still contracting and getting hotter, but now the star can't decrease T_{eff} anymore to account for the increased energy, so instead it has to increase in radius. It moves up the Hayashi track in the reverse process to what happened in the pre-main sequence phase, causing its luminosity to rapidly increase. This is known as the **red giant branch**.

   Due to the convection, some nuclear helium ashes can be mixed into the surface layers in a process called the **first dredge up**. This alters the surface conditions of the star, particularly in the ratios of $^{12}C/^{13}C$ and C/N (from the CNO cycle).

   Because of the much higher luminosity, the lifetime of the red giant phase is much shorter than the main sequence phase. Typically $L_{\text{RG}} \sim 50 L_{\text{MS}}$, so $t_{\text{RG}} \sim 0.02 t_{\text{MS}}$.

8. **Horizontal Branch** [∼100 Myr / 1 Myr]

   Throughout the red giant phase, the core keeps getting hotter. At a certain point, a few different things can happen depending on the mass of the star.

   - **High mass** ($\gtrsim 1.5\ M_\odot$): The contracting core gets hot enough to start helium fusion. This creates a new source of energy/pressure that prevents the core from collapsing any further. The H burning shell continues, but is no longer driven out of equilibrium by the core contraction, so the outer layers of the star stop expanding, causing an increase in the effective temperature (moving left on the HR diagram) and a *slight* decrease in the radius/luminosity until we reach a new equilibrium point along the **horizontal branch**. This can be thought of like a "helium burning main sequence." Instabilities on this branch can lead to the development of variable stars (RR Lyrae).
   - **Low mass** (0.5–1.5 $M_\odot$): In this case, the core will have become degenerate before reaching the Schönberg-Chandrasekhar limit in the subgiant phase. Thus, when the core does eventually get hot enough to start helium burning, it does so in degenerate conditions. This means that the energy injected by the start of helium fusion cannot expand the core fast enough to shut off helium fusion, leading to a runaway reaction called a **helium flash**. This causes a sudden rapid rise in luminosity up to $\sim 10^{11}\ L_\odot$ for a few seconds, creating a spike in the HR diagram called the **tip of the red giant branch (TRGB)**. Afterwards, the degeneracy in the core is released, allowing it to expand and reach an equilibrium state similar to the high mass case.
   - **Very low mass** (0.08–0.5 $M_\odot$): The core will never get hot enough to start burning helium. Theoretically these should become pure He white dwarfs. These have not been observed because their lifetimes are longer than the current age of the universe. As such, we will not examine this case any further.

   In both the high and low mass cases, working from the inside outwards, we have a core burning He into C and O, a layer of inert He, a shell burning H into He, and a rich H envelope. For solar-mass stars, this phase lasts on the order of 100 Myr.

9. **Asymptotic Giant Branch (AGB)** [∼10 Myr / 100 kyr]

   Eventually we run out of He fuel to burn into C and O in the core. Just like in the subgiant phase, this causes our core of inert C and O to contract and release gravitational energy, driving the now shell of He burning into disequilibrium, which causes the outer evelope to expand and cool, getting redder and more luminous. This temporarily causes the H-burning shell to cease. This stage looks very similar to the subgiant and RGB phases and is called the **asymptotic giant branch** because the star asymptotically approaches the Hayashi track on the HR diagram.

   The envelope once again becomes convective because of the decrease in temperature (increase in H^- opacity), which causes a **second dredge up** of core material (^{4}He, ^{14}N, ^{13}C).

   Later in the AGB phase, the H-burning shell reignites and we have double shell burning, with the H-burning shell depositing He into the inert layer from the top, and the He burning shell eating up the inert He layer from the bottom. See the different layers in Figure 3.10.

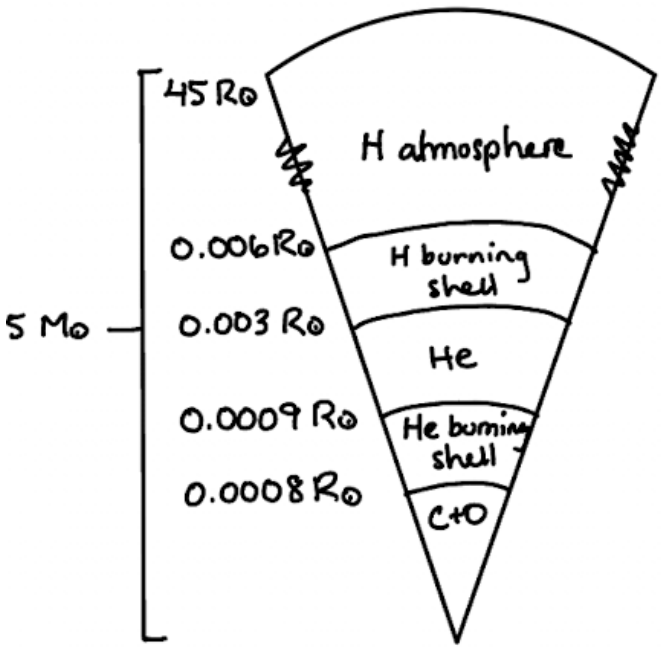


Figure 3.10: A slice of a star on the AGB branch, analogous to figure 3.9.

This is not a steady state—we go through cycles where the dominant shell switches. These are called **thermal AGB pulses**. The cycle generally goes:

- H shell active, He shell dormant
- Inert He layer grows in mass and contracts, heats up
- He burning ignites non-degenerately in thin shell ("shell flash")
- The increased energy expands and cools the envelope, extinguishing the H shell
- He burning eats through the He layer until it reaches the layer of H
- He burning runs out of fuel and stops, but its energy reignites the H burning layer

These cycles repeat on timescales of 10^3–10^5 yr. They can also lead to periodic dredge-ups and very strong stellar winds. The next stages of evolution depend heavily on the mass of the star, so we break this up into two separate cases.

10. A. ($M \lesssim 8\,M_\odot$) **Planetary Nebula** [~100 kyr]

    The strong stellar winds driven by thermal AGB pulses cause significant mass loss in late-stage AGB stars, as high as $10^{-4}\,M_\odot\,\mathrm{yr}^{-1}$. This causes their outer envelopes to be completely blown off until eventually shell burning stops. High energy photons can then ionize the surrounding gas shell, causing it to fluoresce and creating a **planetary nebula**. For more info, see §3.Q27. The luminosity of the star itself drops rapidly because of its decreased radius, while its temperature rises rapidly because we are now seeing deeper into the layers of the star, closer to the core. So during this phase it moves left and down on the HR diagram.

11. A. ($M \lesssim 8\,M_\odot$) **White Dwarf** [$\gtrsim$10–100 Gyr]

    As the outer layers are stripped off, we're left with the inert CO core of the star, supported by electron degeneracy pressure, called a **white dwarf**.

The evolution of a low-mass star on the HR diagram is shown in figure 3.11.

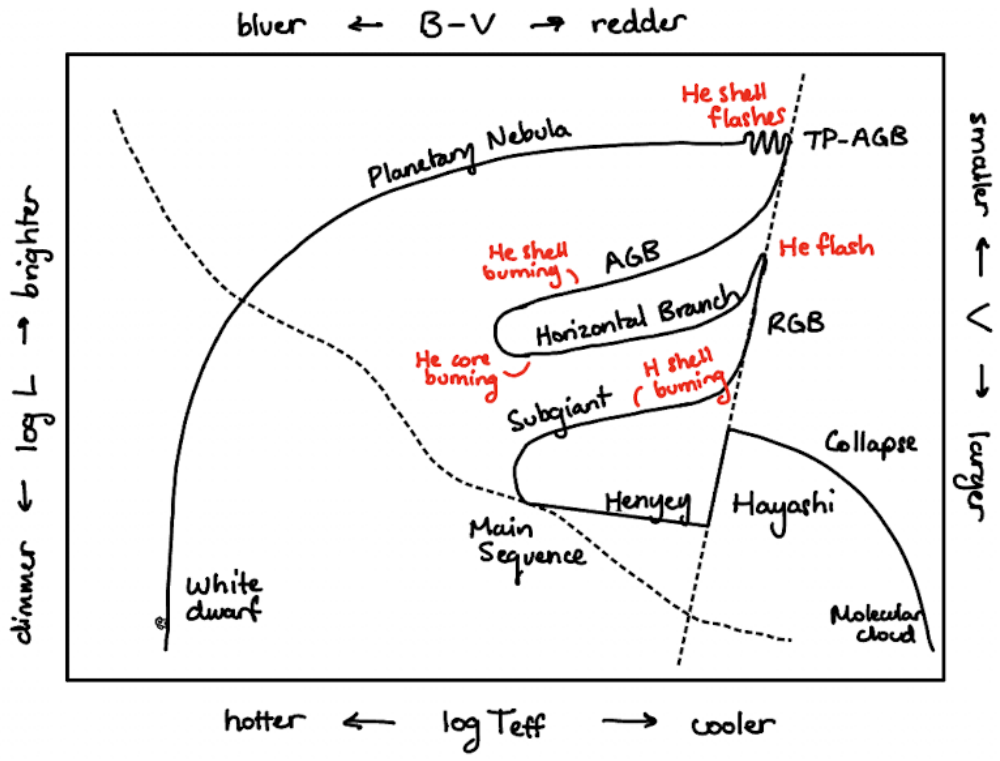


Figure 3.11: The evolution of a low-mass star on the HR diagram.

10. B. ($M \gtrsim 8\,M_\odot$) **Post-AGB / Supergiant** [~300 yr]

    At the end of the AGB phase, once the degenerate CO core reaches a mass of $1.4\,M_\odot$, it can no longer support itself and begins to collapse. It eventually passes the threshold for C burning, which stabilizes it temporarily, but it will continue heating up and passing the thresholds for further burning of O and Si, creating many overlapping layers resembling an onion. As Shrek once said, "Ogres are like onions...Onions have layers." Stars in these late stages of evolution, much like ogres and onions, also have layers, so we can take Shrek's wise words to heart here. See all of these layers in Figure 3.12.

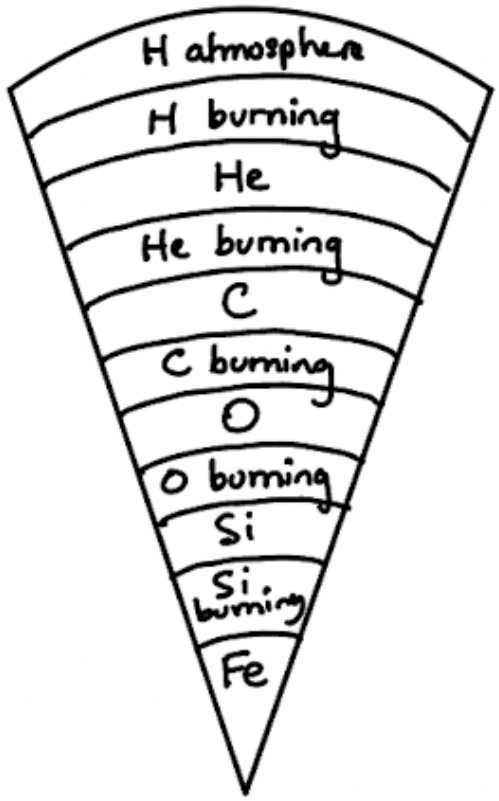


Figure 3.12: The different layers of a massive star fusing everything up to Fe.

These repeated cycles of fusion ignition cause the star to evolve back and forth in temperature on the HR diagram, essentially repeating faster and faster versions of the horizontal branch and AGB. At this stage, the star is sometimes referred to as a **post-AGB star** or a **supergiant**. Note: high-mass stars do not experience significant upwards evolution in the HR diagram like low-mass stars do, because they are already close to the Eddington limit during their early phases of evolution.

11. B. ($M \gtrsim 8\,M_\odot$) **Supernova** [$\sim$1 s]

    After the massive star starts burning Si, it builds up a core of inert ^{56}Fe. The star has been trying so hard to prevent its own gravitational demise by burning more and more elements, but at this point our poor boi is doomed because ^{56}Fe has the highest binding energy per nucleon. This means that if we tried burning Fe into anything else, we would actually *lose* energy in the process. Without any mechanism to push back against gravity, the core starts to collapse uncontrollably.

    During the collapse, the temperature gets hotter and hotter, and the Fe starts succumbing to photodisintegration ($T \sim 7 \times 10^9$ K) where it breaks up into its constituent protons and neutrons. Normally the neutrons would decay into protons by emitting an electron and a neutrino, but they can't here because the electron degeneracy pressure means there are no available states for an electron to be created in. Instead, the protons start absorbing free electrons and create more neutrons. As the collapse continues, eventually neutrons become degenerate and we get neutron degeneracy pressure. Once this pressure kicks in, the core suddenly becomes

incompressible (the neutrons are literally packed as close together as is possibly allowed), which causes the collapsing outer envelope to bounce back at an incredible pace, creating a **supernova explosion**.

The dynamical timescale of the core collapse is incredibly short, $t_{\rm dyn} \sim 1/\sqrt{G\rho} \sim 1$ second, during which it releases an incredible amount of energy $E \sim 10^{53}$ erg. The vast majority of the energy released is in the form of neutrinos (created when the protons combine with electrons to form neutrons).

12. B. ($M \gtrsim 8\,M_\odot$) **Neutron Star** [∼10 Myr–10 Gyr]

    In this case, the leftover core, composed entirely of neutrons and supported by neutron degeneracy pressure, is called a **neutron star**. These typically have radii of ∼ 10 km.

    If instead we have a really massive star ($M_{\rm ZAMS} \gtrsim 25\ M_\odot$) such that the core mass is $\sim 2\ M_\odot$, then even the neutron degeneracy pressure will not be enough to prevent it from completely collapsing. All the mass gets compressed into a single point with 0 volume and infinite density called a singularity. This creates a **black hole**.

The evolution of a high-mass star on the HR diagram is shown in figure 3.13.

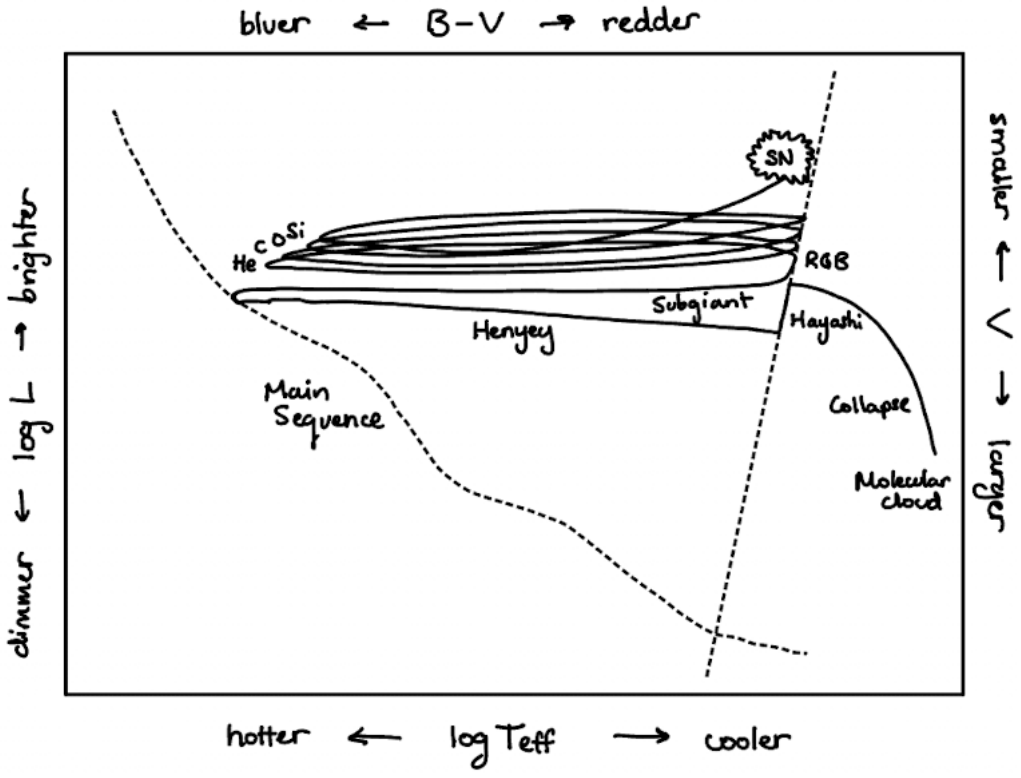


Figure 3.13: The evolution of a high-mass star on the HR diagram. Notice it is much more vertically constrained than the low-mass case since the star already starts close to the Eddington limit.

References: 11,22,27,28.

23. Describe the types of stellar evolution that lead to type Ia, type Ib and type II supernovae. What are the observational differences among these?

The observational classification of supernova types predates our knowledge of what these events actually *are* physically, so the distinctions don't actually make that much sense in terms of the physical scenarios going on (type Ib and type II are much more similar than type Ib and type Ia). Nevertheless, there is a helpful flowchart in figure 3.14 that describes how the classification is determined. All of the classifications depend on which species of absorption lines are detected in the spectrum at peak luminosity[22].

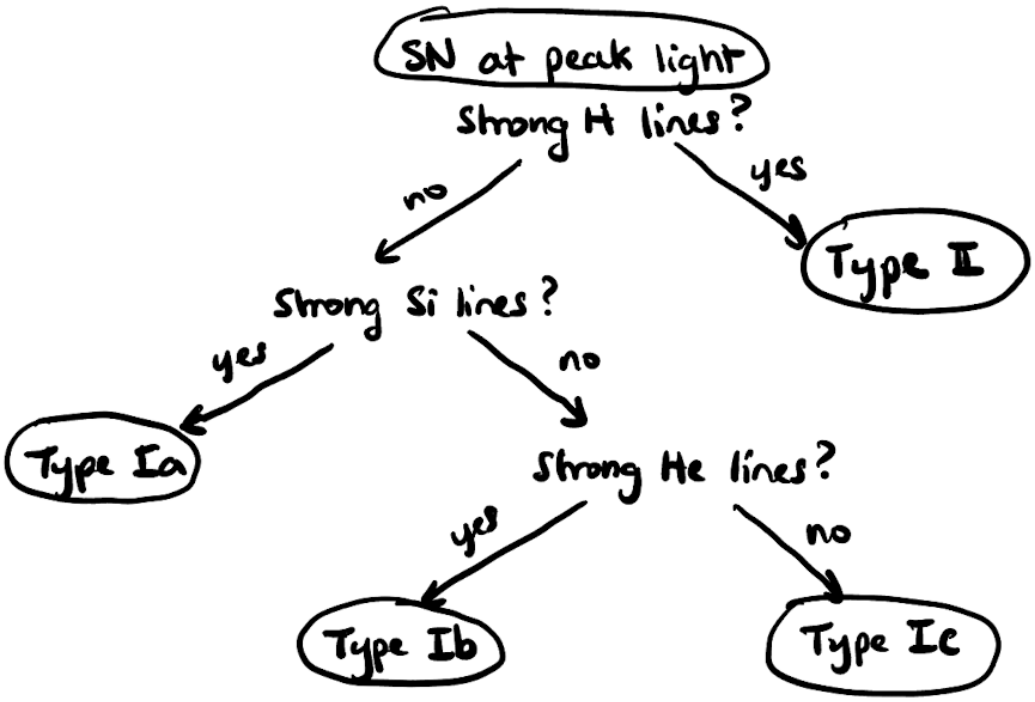


Figure 3.14: A flowchart to determine how supernovae are classified.

- **Type Ia**: CO white dwarf detonation

  Type Ia supernovae are caused by thermonuclear carbon detonation in white dwarfs, the end products of stellar evolution for low mass stars ($M \lesssim 8\ M_\odot$; see §3.Q22).

  White dwarfs, which are supported by electron degeneracy pressure, have a maximum stable mass of $\sim 1.4\ M_\odot$ called the **Chandrasekhar limit**. If a white dwarf were to somehow surpass this maximum mass, it can no longer be supported by electron degeneracy pressure and it starts to collapse gravitationally. This causes runaway carbon fusion because the degenerate WD cannot expand to regulate its temperature. Heavy elements get fused on timescales of a few seconds before an energetic explosion releases $\sim 10^{51}$ erg of energy.

  How can a white dwarf suddenly gain mass to surpass the Chandrasekhar limit? This can happen if the WD is in a binary system and accretes mass from its companion star ("singly degenerate"), or it can happen if two WDs merge with each other ("doubly degenerate").

Because the Chandrasekhar limit is the same for all WDs, all type Ia supernovae are approximately the same luminosity, so these are good **standard candles** and are an important rung on the **distance ladder**. This also makes them a good object of study for determining the dark energy content of the universe Ω_Λ (since this has an effect on the luminosity distance d_L, which is what we measure with standard candles)—for details see §8.Q140.

- **Type Ib and Ic**: Core collapse

  Type Ib and Ic supernovae are both caused by the core collapse of a star at the end of its evolutionary stages. This only happens if the star is massive enough ($M \gtrsim 8\,M_\odot$) to fuse heavy elements up to ^{56}Fe in its core, at which point nothing further can be fused to prevent gravitational collapse (see §3.Q22 for details).

  Type Ib and Ic supernovae are special because they do not exhibit any H absorption lines in their spectra, so they must have originated from extremely massive ($M \gtrsim 30\,M_\odot$) progenitors that had strong enough stellar winds in the AGB phase to completely blow away their H-rich envelopes before the core collapse phase. These are called **Wolf-Rayet stars**. The only difference between Ib and Ic is that Ic also do not exhibit He absorption lines, so they must have had their He envelopes blown away before core collapse too.

- **Type II**: Core collapse

  Like type Ib and Ic, type II supernovae originate from the core collapse of a $M \gtrsim 8\,M_\odot$ star at the end of its evolution. The difference is that these do exhibit H absorption, so they originate from less massive stars that have not lost their H envelopes before core collapse. See §3.Q22 for more details.

A few related classes of events:

- **Classical & Dwarf Novae**:

  Accreting white dwarfs in binary systems that experience instabilities in mass transfer rates and thermonuclear hydrogen detonation in their atmospheres that does *not* completely disrupt them. See §4.Q60.

- **Kilonovae**:

  Neutron star–Neutron star mergers that also produce short GRBs and gravitational wave transients. They are $\sim 1\%$ as bright as a typical supernova and their light curves are thought to be powered by r-process decay. See §4.Q68.

- **Hypernovae**:

  Core-collapse events of extremely massive ($\gtrsim 100\ M_\odot$) stars, which are linked to long GRBs and the formation of BH remnants. About $10\times$ as energetic as a typical supernova. See §4.Q68.

- **Pair-Instability Supernovae**:

  In extremely massive stars (130–250 $M_\odot$) with low metallicities, gamma rays in the core can become energetic enough that they produce electron-positron pairs. This reduces the radiation pressure in the core supporting the star against collapse, causing the core to collapse in a runaway process. This leads to a supernova. The rapidly contracting core has a runaway nuclear fusion reaction for several seconds which completely disrupts it—no black hole or other stellar remnant is left behind. This is predicted to contribute to a mass gap in the black hole population (see §4.Q38), but so far none of these events have been observed/confirmed.

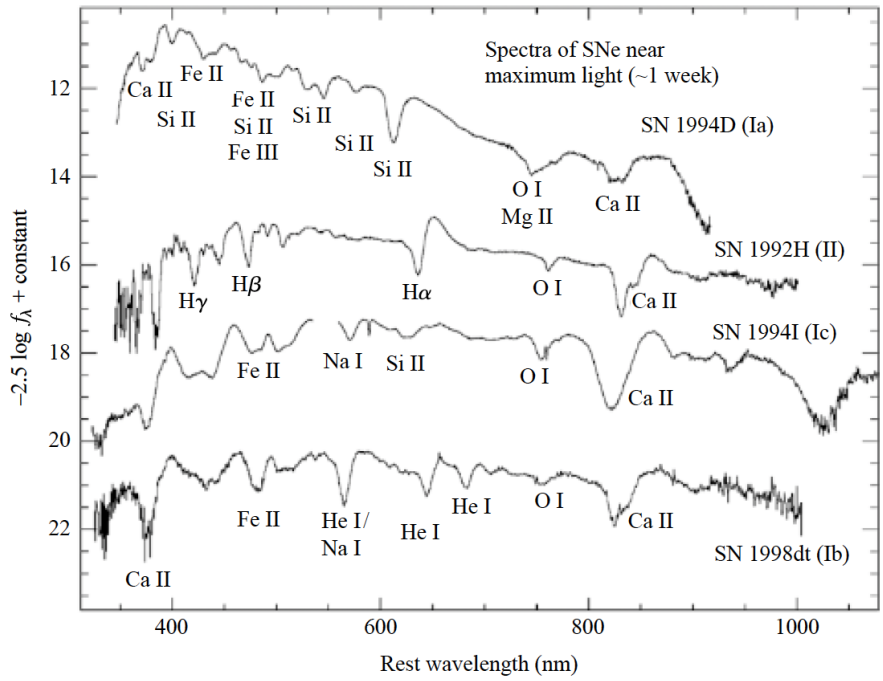


Figure 3.15: Spectra of supernovae for each type, with important absorption lines labeled[11].

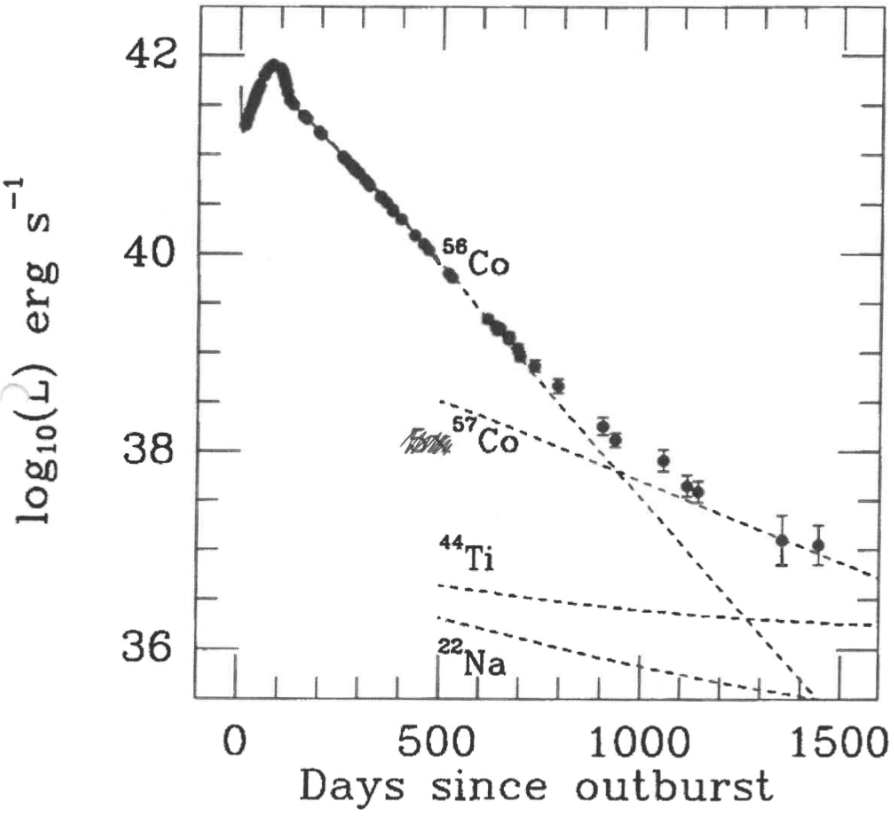


Figure 3.16: The light curve of a type II supernova[29].

Examples of spectra from supernovae of each type are shown in Figure 3.15. And an example of a light curve from a type II supernova is shown in Figure 3.16. This is powered by 2 sources:

- Shock-deposited energy from the initial explosion
- Radioactive decays of ^{56}Ni and its by-products:

$$^{56}\text{Ni} \longrightarrow {}^{56}\text{Co} + \text{e}^{+} + \nu_e + \gamma \tag{3.53}$$

$$^{56}\text{Co} \longrightarrow {}^{56}\text{Fe} + \text{e}^{+} + \nu_e + \gamma \tag{3.54}$$

  Most of the ^{56}Ni decays during the initial explosion, while the following downwards slope is due to the decay of ^{56}Co.

Zooming in on the first few hundred days, a comparison between a type II and a type Ia supernova is shown in Figure 3.17. Type II supernova experience a plateau thought to be caused by the expansion and cooling of the star's outer envelope as it's blown into space. Type Ia supernova have a more consistent decline.

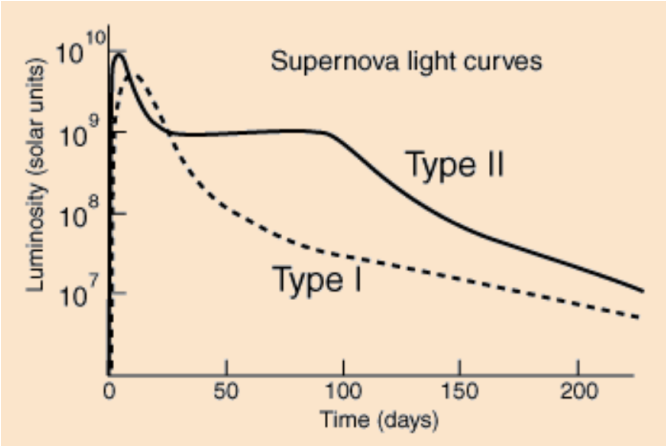


Figure 3.17: Light curves of type Ia and type II supernovae[30].

24. What is a Cepheid variable? Explain the underlying stellar physics involved. What is the Leavitt Law? Over what range of distances do Cepheids play a major role in the "cosmic distance ladder"?

Cepheid variables are a class of pulsating stars with a brightness that varies over time in a predictable way. They are named after the first discovered example, δ Cephei. This star's strangeness was discovered in 1784 by John Goodricke, who contracted pneumonia while observing it and died at the age of 21[11]. It do be like that sometimes.

It turns out that the period of pulsation for Cepheid variables is correlated with their intrinsic luminosity. Henrietta Swan Leavitt (1868—1921) singlehandedly discovered nearly 5% of the known population of Cepheid variables, and she was the first one to discover this relationship between the period and luminosity. As such, it's called the **Leavitt Law**:

$$\boxed{\log_{10}\left(\frac{\langle L\rangle}{L_\odot}\right) = 1.15\log_{10} P_d + 2.47} \tag{3.55}$$

where P_d is the pulsation period measured in days. Or, in magnitudes in the V band:

$$M_V = -2.81\log_{10} P_d - 1.43 \tag{3.56}$$

This relationship is shown in Figure 3.18[11]. Cepheids tend to be very massive and luminous evolved supergiant / bright giant stars of spectral class F–K, with $M \sim 4$–$20\ M_\odot$, $L \sim 10^3$–$10^5\ L_\odot$, $R \sim 50\ R_\odot$, and $P_d \sim 1$–50 days.

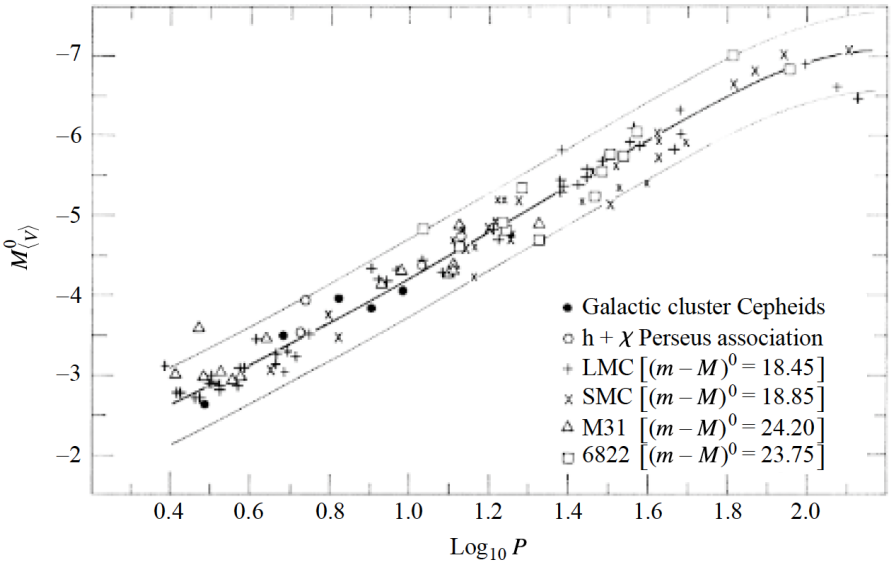


Figure 3.18: The correlation between pulsation period and luminosity for Cepheids in the Small Magellanic Cloud[11].

Why are the period and luminosity related?

The pulsations are related to radial oscillations in the star that result from sound waves resonanting in the interior. Thus, we can estimate the pulsation period Π by considering how long one of these sound waves would take to cross a star with radius R and constant density ρ. The adiabatic sound

speed is

$$c_s^2 = \frac{\partial P}{\partial \rho} = \gamma \frac{P}{\rho} \tag{3.57}$$

And the pressure P can be found by using hydrostatic equilibrium

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{GM\rho}{r^2} = -\frac{4}{3}\pi G\rho^2 r \tag{3.58}$$

Integrating using the boundary condition that $P = 0$ at the surface

$$P(r) = \int_R^r \mathrm{d}r \, \frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{4}{3}\pi G\rho^2 \int_R^r \mathrm{d}r' \, r' = \frac{2}{3}\pi G\rho^2 (R^2 - r^2) \tag{3.59}$$

Thus the pulsation period Π is

$$\Pi \approx 2\int_0^R \frac{\mathrm{d}r}{c_s} \approx 2\int_0^R \frac{\mathrm{d}r}{\sqrt{\frac{2}{3}\gamma\pi G\rho(R^2 - r^2)}} \approx \sqrt{\frac{6}{\gamma\pi G\rho}} \int_0^R \frac{\mathrm{d}r}{\sqrt{R^2 - r^2}} \tag{3.60}$$

This is another one of those annoying trig integrals, so I'll just skip to the answer (the integral is $\pi/2$)

$$\boxed{\Pi \approx \sqrt{\frac{3\pi}{2\gamma G\rho}}} \tag{3.61}$$

We see that $\Pi \propto \rho^{-1/2}$. Now, Cepheid variables occupy a narrow vertical strip on the HR diagram called the **instability strip**, meaning their T_{eff} is $\sim$ constant. I'll explain why this is in just a second, but taking this for granted we can see that this necessitates a proportionality between the period and the luminosity since this means the luminosity is proportional to the radius and hence the density[31]:

$$L \propto R^2 T_{\text{eff}}^4 \propto R^2 \;\;,\;\; \Pi \propto \left(\frac{M}{R^3}\right)^{-1/2} \propto M^{-1/2} R^{3/2} \;\longrightarrow\; \boxed{\Pi \propto M^{-1/2} L^{3/4}} \tag{3.62}$$

Why do Cepheid variables exhist in a narrow strip on the HR diagram?

This is because the mechanism thought to be responsible for creating the sound waves, the ***κ* mechanism** (AKA the **Eddington valve**), can only occur at a very specific temperature. In normal circumstances, an increase in compression of the atmosphere causes an increase in temperature and density, which decreases the opacity of the atmosphere (since $\kappa \propto \rho T^{-7/2}$), allowing more radiation to escape and regulating the temperature and density back down. This is a stable equilibrium. However, when the ***κ* mechanism** operates, an increase in temperature and density cause an *increase* in opacity. This can happen, for example, in a star's **partial ionization zone**, where some of the energy from the compression can go into ionizing more atoms instead of increasing the temperature. Then, our opacity $\kappa \propto \rho T^{-7/2}$ will go *up* since ρ increases and T does not change. Stars typically have a partial ionization zone for Hydrogen (H I $\leftrightarrow$ H II) at $\sim 10^4$ K, and for Helium (He II $\leftrightarrow$ He III) at $\sim 4 \times 10^4$ K. In these zones, the κ mechanism can create cycles if the circumstances are just right[32]. The cycles, conceptually, evolve something like this:

- The He partial ionization zone compresses and gets more dense, ionizing more He II → He III
- The opacity $\kappa \propto \rho T^{-7/2}$ goes up since ρ increases and T remains ∼constant
- The layers beneath the zone heat up because they cannot radiate their energy through the opaque layer
- The internal pressure increases and the layer expands, recombining He III → He II
- The opacity $\kappa \propto \rho T^{-7/2}$ drops since ρ decreases and T remains ∼constant
- The layers beneath the zone cool down because they can radiate their energy more efficiently
- Repeat ad infinitum

If the temperature is too low, convection starts dominating in the star's atmosphere, which mixes the gas and prevents He ionization from driving pulsations. If the temperature is too high, the He ionization zone is too far out in the star's atmosphere to drive pulsations. Therefore, we get the instability strip. See the shaded Cepheid region in Figure 3.19.

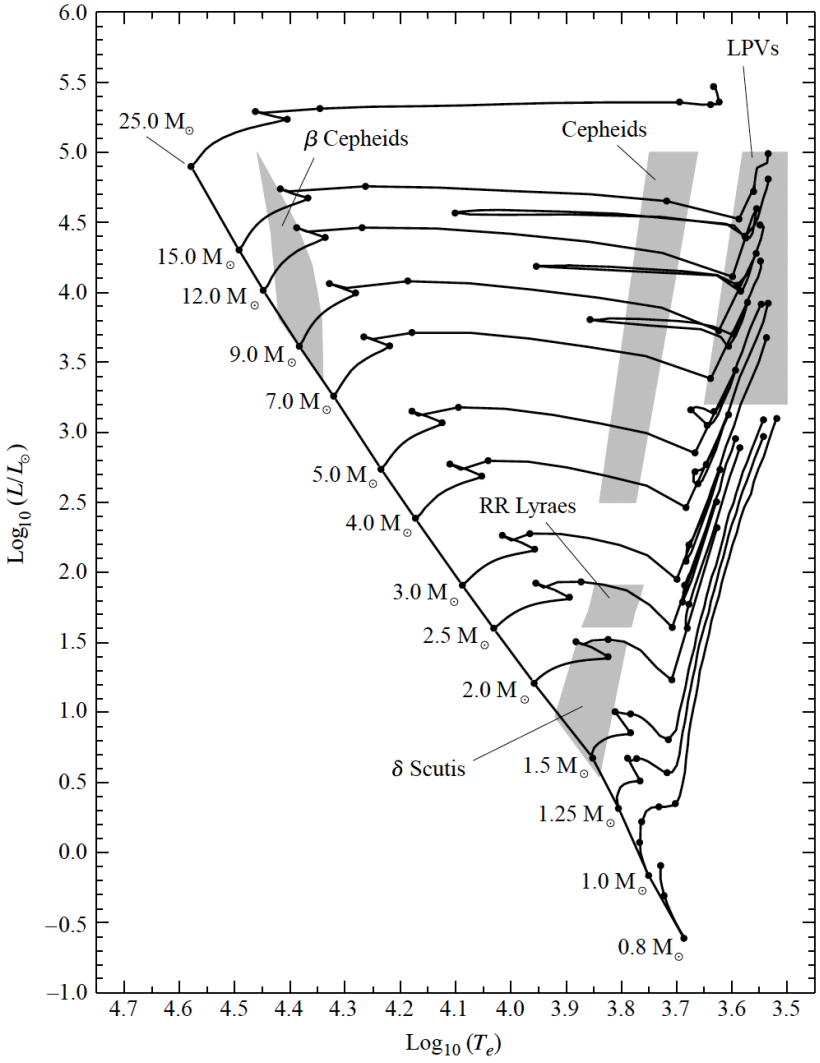


Figure 3.19: The HR diagram annotated with regions for different classes of variable stars[11].

The Distance Ladder

Since we can get absolute magnitudes from the pulsation period, Cepheids play an important role in the distance ladder. They can be used for distances much farther than parallaxes, since Cepheids are bright supergiant stars (luminosity class Ib) and can be seen at intergalactic distances, but they're still relatively local since they require one to be able to resolve individual stars. Parallaxes can only go out to about ∼kpc at most (and that's with a high level of uncertainty). Whereas the furthest known Cepheid currently is in NGC 3370 at a distance of ~ 29 Mpc[33]. This is closer than, for example, type Ia supernovae, which can go up to 100s of Mpc. This gets a bit more complicated because differences in metallicity and dust extinction can cause slight deviations in the period-luminosity relationship for Cepheids.

25. What are RR Lyrae stars, and how do they differ from Cepheids?

RR Lyrae stars are variable stars in the horizontal branch evolutionary phase that are found in globular clusters (Population II—metal poor, old). They occupy the same instability strip as Cepheids, so the mechanisms that drive their oscillations are the same (**κ mechanism**; see §3.Q24). Also like Cepheids, they are named after the prototypical example, RR Lyrae. But since they are subgiant stars with spectral types A–F ($L \sim 50\,L_\odot$, $M \sim 1\,M_\odot$), they are much dimmer than classical Cepheids, and their periods are much shorter, on the order of hours to 1 day. Their period-luminosity relation is also harder to calibrate and is typically less reliable than the one for Cepheids. See the shaded region in Figure 3.19 for reference.

Since RR Lyrae stars need multiple measurements to be averaged over time to obtain their "true" colors and luminosities, they are often omitted from the HR diagrams of clusters and stellar populations altogether, leading to an artificial gap where the instability strip should be. This is purely an artifact of our choice to omit these stars and is not indicative of any underlying physics.

26. What are HH objects? T Tauri stars? Bipolar flows? OH masers? Where are they all found?

Herbig-Haro (HH) Objects[11]

First discovered by George Herbig and Guillermo Haro in the 1950s in the Orion nebula, these are regions of gas around a young protostar that exhibit bright emission line spectra. The protostar emits narrow jets of gas that expand supersonically and collisionally excite the gas, creating the emission lines. The gas in the jets can be moving at 100s of $\mathrm{km\,s^{-1}}$. Continuous emission is also observed in some cases due to reflected light from the protostar. See a diagram in Figure 3.20.

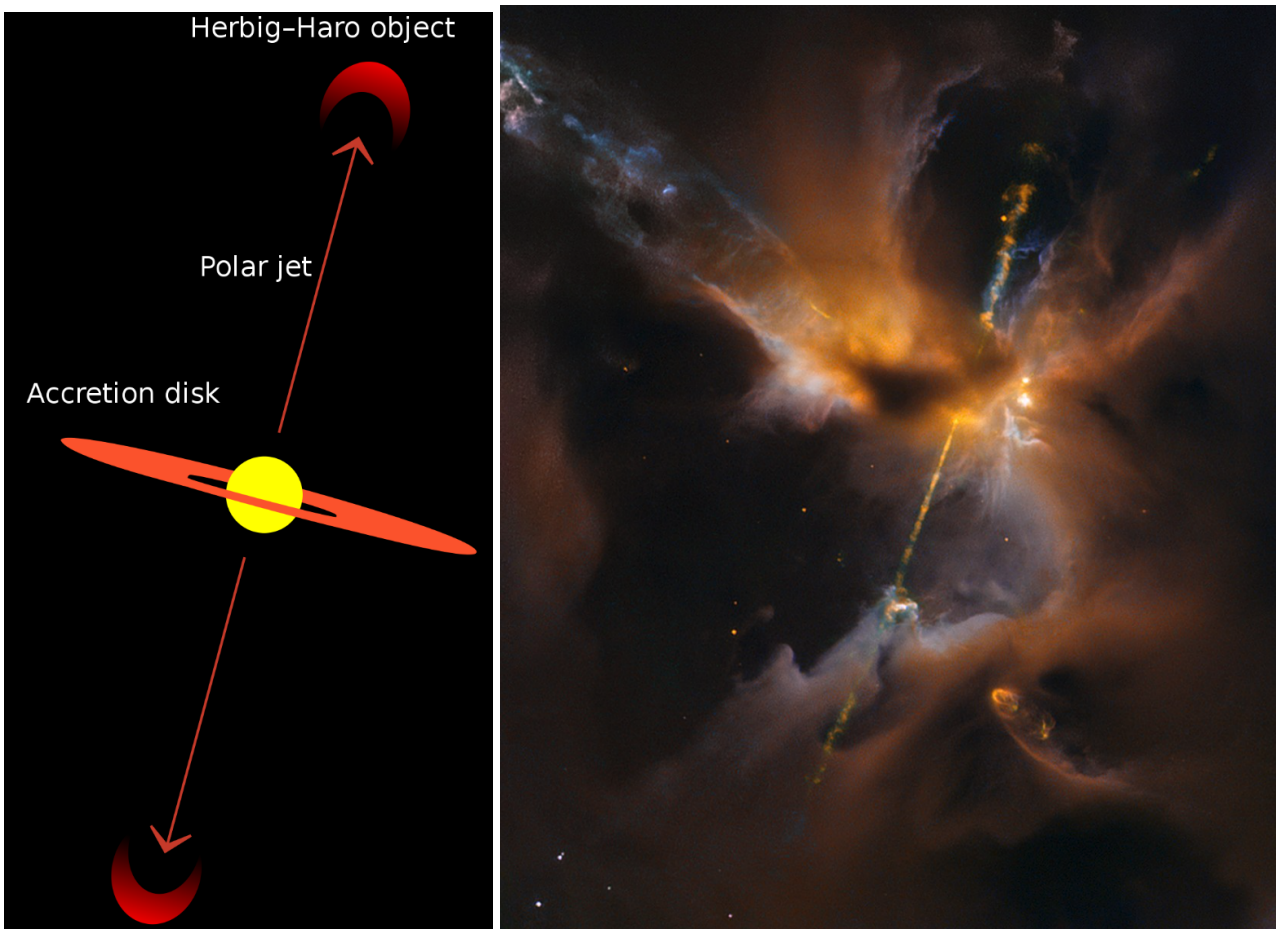


Figure 3.20: Left: Diagram of an HH object. Right: A real HH object observed with *Hubble*. Credit: *NASA*.

T Tauri Stars[11]

A class of low mass (0.5–2 $M_\odot$) pre-main-sequence stars, named after the prototypical example T Tauri. These represent a transition between stars that are still obscured by dust and main sequence stars, and they are characterized by unusual spectral features and rapid irregular variations in luminosity (timescales of days). They are often somewhere along the Hayashi or Henyey track, see Figure 3.21.

Many exhibit strong emission lines from the Balmer series, Ca II, and Fe, as well as absorption from Li. They also show the forbidden lines of [O I] and [S II]—the existence of forbidden lines is indicative

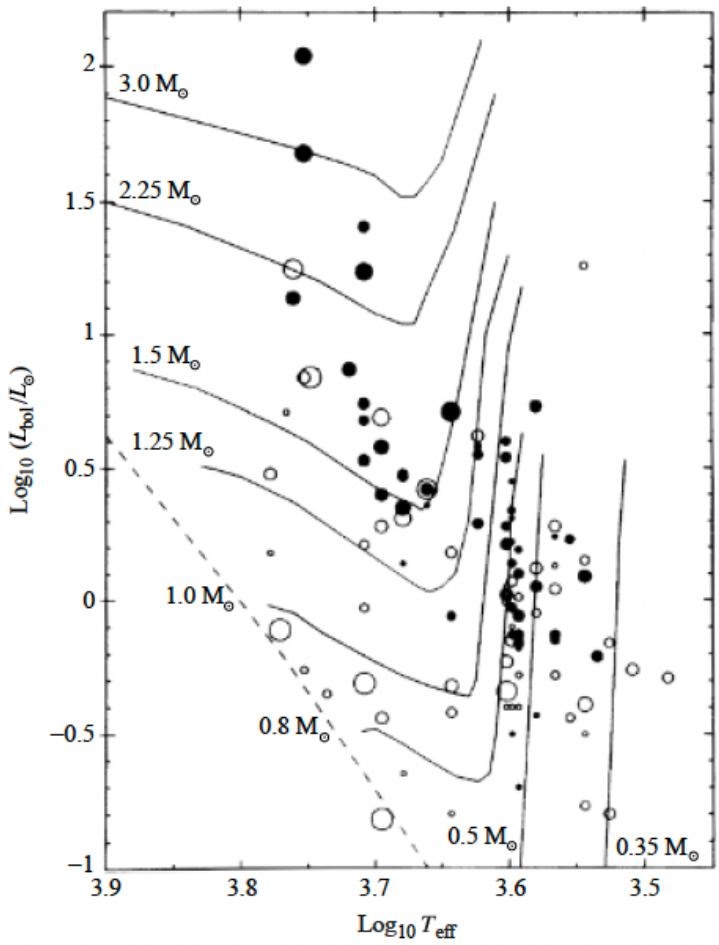


Figure 3.21: The positions of T Tauri stars on the HR diagram. The size of the circles indicates the rate of rotation. Stars with strong emission lines have filled in circles, while stars with weak emission lines have open circles[11].

of very low gas densities. The shapes of these lines is also often in the form of a **P Cygni profile** (see §3.Q28), indicating significant mass loss.

Bipolar flows

These are outflows of material launched by jets (narrow) or winds (wide) at each pole, which can be induced either by protostars or post-AGB stars. In the case of jets, these are highly collimated supersonic (100s of $\mathrm{km\,s^{-1}}$) streams of gas that create shock fronts that heat the surrounding gas to $\sim 10^4$ K (HH objects). The formation of the jet may be related to magnetic fields created by material in the disk (use the right hand rule—a circular current in the disk would lead to a magnetic field going vertically through the axis of rotation), which could allow the protostar to shed angular momentum and prevent itself from being torn apart. In the case of winds, the outflows are slower (10 $\mathrm{km\,s^{-1}}$) and cooler (10 K) and can be observed with molecular lines.

OH masers[11]

Like a laser, the emission from a maser is stimulated and monochromatic. Electrons are "pumped up" from a lower energy level into a higher, long-lived metastable energy state. When a photon with the right energy level comes along, it can stimulate the electron (see §5.Q88 on stimulated emission) to

move down and emit another photon with the same energy in the same direction and in phase with the first photon. This causes a chain reaction that amplifies the radiation of this one energy transition to be much brighter than would be expected in a purely thermal distribution.

In the late AGB stage of evolution, stars develop strong winds and experience significant mass loss (see §3.Q22). The outer layers of the star expand out and become very cool and optically thick. These clouds can form molecules that get photodissociated into OH (hydroxyl radicals), which develops the perfect conditions for an OH maser to form. These can also form around protostars that are still enshrouded by an optically thick molecular cloud, for similar reasons.

27. What is a planetary nebula? What is our current understanding of the formation of planetary nebulae?

A late stage AGB star produces strong stellar winds that strips it of its outer layers. These outer layers create a surrounding shell of gas that is irradiated by UV light from the now exposed white dwarf, causing it to fluoresce (mostly in the UV and visible) and produce a **planetary nebula**. The morphologies of these nebulae can tell us a lot about the angular momentum of the star, magnetic fields, possible companion stars, and more. They are only produced in stars with masses between 0.8–8 $M_\odot$. More massive stars will lose their atmospheres too quickly (like Wolf-Rayet stars), while less massive stars will never start fusing He into C and O in their cores, so they won't produce ionizing radiation to cause their ejected atmospheres to fluoresce. The name is obviously a misnomer, they have nothing to do with planets (they just sort of looked like gas giants when viewed through nineteenth century telescopes)[11,34]. See an example in Figure 3.22.

Figure 3.22: *Left*: An example of a planetary nebula: The Cat's Eye Nebula, in the optical and X-ray using data from *Hubble* and *Chandra*. Credit: NASA. *Right*: The Butterfly Nebula, using data from *Hubble*.

Typical characteristics:

- Length scale: 0.3 pc
- Expansion velocity: 10–30 $\mathrm{km\,s^{-1}}$
- Temperature: 10^4 K
- Age: 10^4 yr (max $\sim 5 \times 10^4$ yr before it dissipates into the ISM)

28. What is a P Cygni line profile, and what does it signify?

A **P Cygni profile** is a line profile which has an emission component superimposed with a blueshifted absorption component. It was first observed around the star P Cygni, hence the name. The presence of a profile like this indicates that the star is experiencing significant mass loss, with an expanding shell of cold gas that absorbs the emission from the star with a relative blueshift because of the expansion (see Figure 3.23)[11].

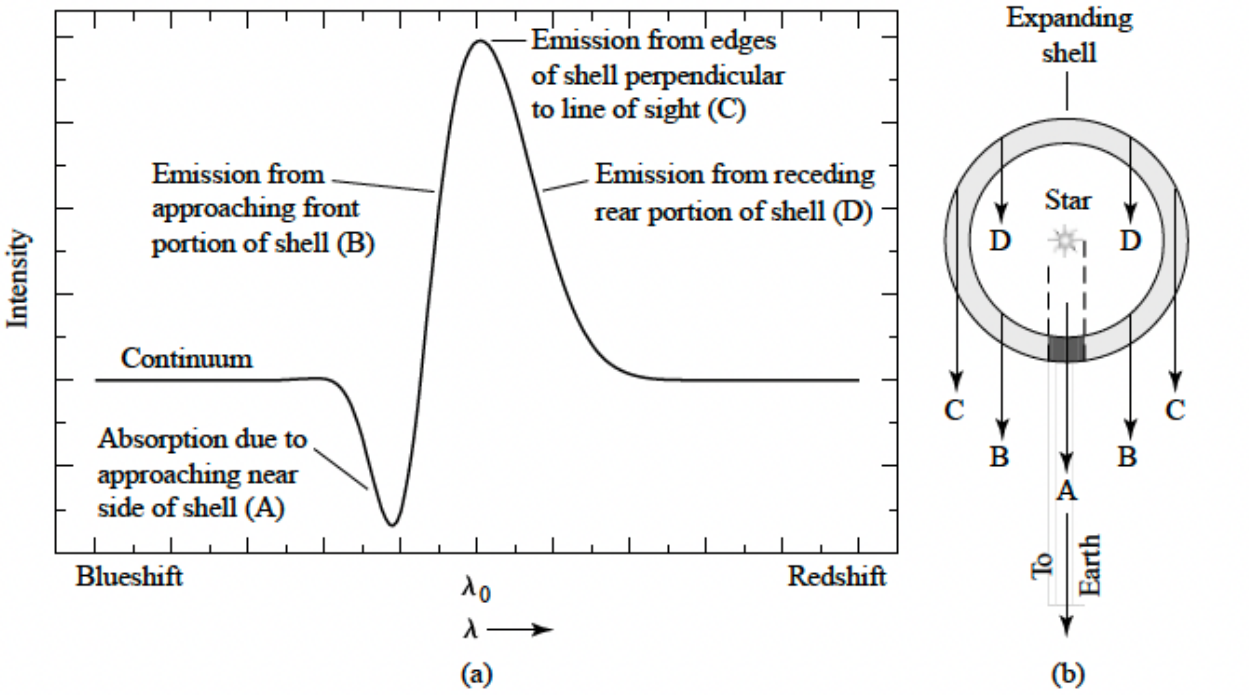


Figure 3.23: Left: A P Cygni profile. Right: A model of how the P Cygni profile is formed, labeling the different parts of the star that contribute to emission/absorption[11].

29. Two stars are observed to have the same color and brightness. One of them is a giant at a greater distance than the other, which is a main sequence star. How could these be distinguished from spectroscopic measurements?

If two stars have the same color, that means the slope of their blackbody curve is the same, so they have the same T_{eff} (vertical column on the HR diagram). If they have the same *apparent* brightness but are at different distances, then the giant will have a larger *intrinsic* brightness but will be further away.

We can distinguish these two cases by using spectra of each star and looking at the shapes of the absorption line profiles. This is because *pressure broadening* becomes important in the atmospheres of stars. The orbitals of atoms can be perturbed by collisions or close EM encounters with other atoms/ions, which causes a general broadening of the line profiles produced by these orbitals. This effect is proportional to the likelihood that one of these close encounters happens, which is the inverse of the mean free path:

$$\Delta\lambda \propto \frac{v}{\ell} \propto nv\sigma \tag{3.63}$$

The important part here is the proportionality to the number density n of particles. This, in turn, is proportional to the star's radius like $n \propto R^{-3}$. Therefore $\Delta\lambda \propto R^{-3}$. Obviously a giant star with the same T_{eff} as a main sequence star will have a larger radius:

$$L = 4\pi R^2 \sigma_{\text{SB}} T_{\text{eff}} \tag{3.64}$$

$$L_{\text{giant}} > L_{\text{MS}} \quad \longrightarrow \quad R_{\text{giant}} > R_{\text{MS}} \tag{3.65}$$

So a giant star will have narrower absorption lines than a main sequence star:

$$\boxed{\Delta\lambda_{\text{giant}} < \Delta\lambda_{\text{MS}}} \tag{3.66}$$

This is because the surface gravity ($g = GM/R^2$) of a star with a larger radius is smaller, so the atmosphere of a giant star is less compressed than the atmosphere of a main sequence star, and pressure broadening becomes less important.

This is how the **MK luminosity classes** (§3.Q16) are determined[11].

30. Write down and describe the four basic equations of stellar structure.

1. Conservation of mass

Say that the enclosed mass at a radius r is $M(r)$ and the density at this point is $\rho(r)$. Conservation of mass states that the mass within a thin spherical shell of volume $4\pi r^2 \mathrm{d}r$ should just be $\mathrm{d}M = 4\pi r^2 \rho(r)\mathrm{d}r$. Therefore, we get our first stellar structure equation:

$$\boxed{\frac{\mathrm{d}M(r)}{\mathrm{d}r} = 4\pi r^2 \rho(r)} \tag{3.67}$$

2. Hydrostatic equilibrium

For a star to be stable (i.e. not expanding or contracting), the thermal pressure $P(r)$ pushing outwards at some radius should be balanced by the inwards pull of gravity at the same radius. Say we have a small area element $\mathrm{d}A$ at r. The pressure pushes outwards from the bottom with a force of $P(r)\mathrm{d}A$, and it pushes inwards from the top with a force of $-P(r+\mathrm{d}r)\mathrm{d}A$. If we say $P(r+\mathrm{d}r) = P(r) + \mathrm{d}P$, then the *net* force from pressure on the area element is $-\mathrm{d}P\mathrm{d}A$. Then we just set this equal to the gravitational force:

$$-\mathrm{d}P\mathrm{d}A - \frac{GM(r)}{r^2}\rho(r)\mathrm{d}r\mathrm{d}A = 0 \tag{3.68}$$

The area elements cancel out, and we get the equation of hydrostatic equilibrium:

$$\boxed{\frac{\mathrm{d}P(r)}{\mathrm{d}r} = -\frac{GM(r)\rho(r)}{r^2}} \tag{3.69}$$

3. Energy generation

Nuclear fusion generates energy, which is transported outwards in a spherically symmetric shell of volume $4\pi r^2 \mathrm{d}r$. This produces a luminosity (i.e. a power) of $L(r)$ through the shell. If we have an energy generation rate per unit mass of $\varepsilon_m(r)$ (units of $\mathrm{erg\,s^{-1}\,g^{-1}}$) and we assume the star is in thermal equilibrium, then the change in luminosity across $\mathrm{d}r$ is $\mathrm{d}L = \varepsilon_m(r)\mathrm{d}M$. However, in general if we're not in thermal equilibrium, this will equal the heating rate:

$$\varepsilon_m(r)\mathrm{d}M - \mathrm{d}L = \frac{\mathrm{d}Q}{\mathrm{d}t}\mathrm{d}M = T\frac{\mathrm{d}S}{\mathrm{d}t}\mathrm{d}M \tag{3.70}$$

where Q is the heat and S is the entropy. Therefore, writing out $\mathrm{d}M$, our energy generation equation becomes

$$\boxed{\frac{\mathrm{d}L(r)}{\mathrm{d}r} = 4\pi r^2 \rho(r)\left[\varepsilon_m(r) - T(r)\frac{\mathrm{d}S(r)}{\mathrm{d}t}\right]} \tag{3.71}$$

where the $T\ \mathrm{d}S/\mathrm{d}t$ term goes to 0 in thermal equilibrium.

4. Energy Transport

The energy generated by nuclear fusion creates a luminosity gradient, as we saw in the last equation, but this difference in energy also creates a temperature gradient depending on how the energy is transported throughout the star. Energy transport can come in two forms: radiative and convective.

These each have a corresponding stellar structure equation that describes the temperature gradient in the star when each one is the dominant mechanism.

4.A. Radiative Energy Transport

We can start with the equation of radiative transfer in a plane-parallel stellar atmosphere

$$\mu \frac{\mathrm{d}I_\nu}{\mathrm{d}\tau_\nu} = I_\nu - S_\nu \tag{3.72}$$

where $\mu = \cos\theta$. Multiplying by μ/c and integrating over the solid angle $\mathrm{d}\Omega = \sin\theta \mathrm{d}\theta \mathrm{d}\phi = 2\pi \mathrm{d}\mu$, we get

$$2\pi \frac{\mathrm{d}}{\mathrm{d}\tau_\nu} \int_{-1}^{1} \mathrm{d}\mu\, \mu^2 \frac{I_\nu}{c} = \frac{2\pi}{c} \int_{-1}^{1} \mathrm{d}\mu\, \mu (I_\nu - S_\nu) \tag{3.73}$$

We identify the left integral as the radiation pressure (e.g. equation (12.56)). For the right integral, the source function S_ν is isotropic (blackbody), so it can be pulled out of the integral—this term then vanishes when integrating over the solid angle. Then we're just left with the integral of $I_\nu \mu$ over the solid angle, which is just the flux (equation (12.51)). Therefore

$$\frac{\mathrm{d}P_{\nu,\mathrm{rad}}}{\mathrm{d}\tau} = \frac{F_\nu}{c} \quad \longrightarrow \quad F_\nu = -\frac{c}{\alpha_\nu} \frac{\mathrm{d}P_{\nu,\mathrm{rad}}}{\mathrm{d}r} \tag{3.74}$$

The sign flipped when we replaced $\mathrm{d}\tau_\nu$ with $\alpha_\nu \mathrm{d}r$ since the optical depth gets larger as r gets smaller (we get closer to the center of the star). Now we integrate over the frequency. Typically we don't know the form of $\alpha_\nu = \rho\kappa_\nu$ precisely, so we define the **Rosseland mean opacity** κ_R such that

$$F = -\frac{c}{\rho\kappa_R} \frac{\mathrm{d}P_{\mathrm{rad}}}{\mathrm{d}r} \tag{3.75}$$

The radiation pressure is dominated by blackbody emission, $P_\nu = u_\nu/3 = (4\pi/3c)B_\nu$, so the pressure gradient is $\mathrm{d}P_\nu/\mathrm{d}r = (4\pi/3c)\, \partial B_\nu/\partial T \times \mathrm{d}T/\mathrm{d}r$. Then the Rosseland mean opacity is defined by

$$\frac{1}{\kappa_R} \equiv \frac{\int_0^\infty \frac{1}{\kappa_\nu} \frac{\partial B_\nu}{\partial T} \mathrm{d}\nu}{\int_0^\infty \frac{\partial B_\nu}{\partial T} \mathrm{d}\nu} \tag{3.76}$$

(the $\mathrm{d}T/\mathrm{d}r$ terms cancel out). Now we replace F with $L/4\pi r^2$ and we have

$$\frac{L}{4\pi r^2} = -\frac{c}{\rho\kappa_R} \frac{\mathrm{d}}{\mathrm{d}r} \left(\frac{a}{3} T^4 \right) = -\frac{4ca}{3\rho\kappa_R} T^3 \frac{\mathrm{d}T}{\mathrm{d}r} \tag{3.77}$$

Solving for the derivative of T gives us our stellar structure equation

$$\boxed{\frac{\mathrm{d}T(r)}{\mathrm{d}r} = -\frac{3\kappa_R(r) L(r) \rho(r)}{16\pi a c r^2 T(r)^3}} \tag{3.78}$$

4.B. Convective Energy Transport

Say a bubble with ρ_1 and P_1 expands adiabatically (i.e. with no exchange of heat) to a new ρ_2 and

P_2. The density and pressure of the *environment* at these two locations will be designated (ρ_1', P_1') and (ρ_2', P_2') respectively. We'll assume for simplicity that the bubble starts out in the same conditions as its environment, so $\rho_1 = \rho_1'$ and $P_1 = P_1'$. See Figure 3.24.

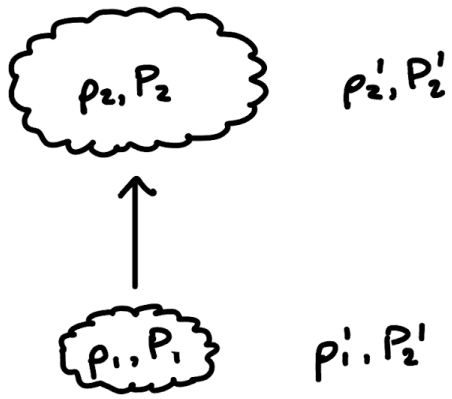


Figure 3.24: Adiabatic expansion of a bubble.

Since we're assuming adiabatic expansion, then we have an equation of state such that

$$P \propto \rho^{\gamma} \quad , \quad \rho \propto P^{1/\gamma} \tag{3.79}$$

Now we'll approximate the density of the bubble at the new location with a first-order Taylor series

$$\rho_2 = \rho_1 + \frac{\mathrm{d}\rho}{\mathrm{d}r}\Delta r = \rho_1 + \frac{\mathrm{d}\rho}{\mathrm{d}P}\frac{\mathrm{d}P}{\mathrm{d}r}\Delta r = \rho_1 + \frac{\rho_1}{\gamma P_1}\frac{\mathrm{d}P}{\mathrm{d}r}\Delta r \tag{3.80}$$

Whereas the density of the *environment* at the new position ρ_2', will be

$$\rho_2' = \rho_1' + \frac{\mathrm{d}\rho}{\mathrm{d}r}\Delta r \tag{3.81}$$

Here we can't do the same replacement that we did for the bubble since the environment is not in an adiabatic process. But, recall that $\rho_1' = \rho_1$. Therefore, the density contrast between the bubble and the environment at the new location is

$$\rho_2 - \rho_2' = \frac{\rho_1}{\gamma P_1}\frac{\mathrm{d}P}{\mathrm{d}r}\Delta r - \frac{\mathrm{d}\rho}{\mathrm{d}r}\Delta r \tag{3.82}$$

For the environment we can use the ideal gas law ($P \propto \rho T$) to obtain (dropping the "1" subscripts)

$$\rho_2 - \rho_2' = \left[\frac{\rho}{\gamma P}\frac{\mathrm{d}P}{\mathrm{d}r} - \frac{\rho}{P}\frac{\mathrm{d}P}{\mathrm{d}r} + \frac{\rho}{T}\frac{\mathrm{d}T}{\mathrm{d}r}\right]\Delta r = \left[-\left(1-\frac{1}{\gamma}\right)\frac{\rho}{P}\frac{\mathrm{d}P}{\mathrm{d}r} + \frac{\rho}{T}\frac{\mathrm{d}T}{\mathrm{d}r}\right]\Delta r \tag{3.83}$$

Thus, for convection to drive the energy transport the LHS must be positive. We can take the boundary for convection to be very close to the actual temperature gradient during convection, since it's an extremely efficient energy transport mechanism that only requires an energy gradient slightly steeper than the critical value. Thus, we just set the quantity in brackets to 0 and solve for the temperature

gradient

$$\boxed{\frac{\mathrm{d}T(r)}{\mathrm{d}r} = \left(1 - \frac{1}{\gamma}\right)\frac{T(r)}{P(r)}\frac{\mathrm{d}P(r)}{\mathrm{d}r}} \tag{3.84}$$

Review

The four equations of stellar structure are[35]

$$\frac{\mathrm{d}M(r)}{\mathrm{d}r} = 4\pi r^2 \rho(r) \qquad \text{(conservation of mass)}$$

$$\frac{\mathrm{d}P(r)}{\mathrm{d}r} = -\frac{GM(r)\rho(r)}{r^2} \qquad \text{(hydrostatic equilibrium)}$$

$$\frac{\mathrm{d}L(r)}{\mathrm{d}r} = 4\pi r^2 \rho(r)\left[\varepsilon_m(r) - T(r)\frac{\mathrm{d}S(r)}{\mathrm{d}t}\right] \qquad \text{(energy generation)}$$

$$\frac{\mathrm{d}T(r)}{\mathrm{d}r} = \begin{cases} -\dfrac{3\kappa_R(r)L(r)\rho(r)}{16\pi a c r^2 T(r)^3} & \text{(radiative energy transport)} \\ \left(1 - \dfrac{1}{\gamma}\right)\dfrac{T(r)}{P(r)}\dfrac{\mathrm{d}P(r)}{\mathrm{d}r} & \text{(convective energy transport)} \end{cases}$$

31. Make a dimensional analysis of the equation of hydrostatic equilibrium using a polytropic equation of state to find a general mass-radius relation for spherically-symmetric, self-gravitating bodies. For which two polytropic indices is the configuration unstable?

Using hydrostatic equilibrium (3.69) and an equation of state with a polytropic index γ, we have

$$\frac{\mathrm{d}P(r)}{\mathrm{d}r} \propto \frac{\mathrm{d}\rho(r)^\gamma}{\mathrm{d}r} \propto \frac{M(r)\rho(r)}{r^2} \tag{3.85}$$

$$\gamma\rho(r)^{\gamma-1}\frac{\mathrm{d}\rho(r)}{\mathrm{d}r} \propto \frac{M\rho(r)}{r^2} \tag{3.86}$$

When $r = R$, we have $M(r) = M$ and $\rho(r) \propto M/R^3$:

$$\left(\frac{M}{R^3}\right)^{\gamma-1}\frac{M}{R^4} \propto \frac{M^2}{R^5} \tag{3.87}$$

$$M^{\gamma-2} \propto R^{-4+3\gamma} \tag{3.88}$$

$$\therefore \quad \boxed{M \propto R^{(3\gamma-4)/(\gamma-2)}} \tag{3.89}$$

This relationship shows how M must change with R in order for hydrostatic equilibrium to hold in a polytropic gas. We can see that this is unstable in two cases:

$$\boxed{\gamma = \frac{4}{3} \longrightarrow M \propto R^0} \quad \text{(constant)} \tag{3.90}$$

$$\boxed{\gamma = 2 \longrightarrow M \propto R^\infty} \quad \text{(infinite)} \tag{3.91}$$

The first case is unstable because, as R decreases, the gas pressure $P_{\text{gas}} \propto \rho^\gamma \propto (M/R^3)^\gamma$ does not increase fast enough to counteract what is required by hydrostatic equilibrium $P_{\text{HSE}} \propto M^2/R^4$ (see (3.100)):

$$\mathrm{d}P_{\text{gas}} \propto -R^{-3\gamma-1}\mathrm{d}R \tag{3.92}$$

$$\mathrm{d}P_{\text{HSE}} \propto -R^{-5}\mathrm{d}R \tag{3.93}$$

$$-R^{-3\gamma} > -R^{-4} \longrightarrow R^{-3\gamma} < R^{-4} \longrightarrow \gamma > \frac{4}{3} \tag{3.94}$$

So any $\gamma \leqslant 4/3$ will be unstable[22].

Note: If instead of a polytropic equation of state we use the ideal gas law ($P \propto \rho T$) and ignore the scaling with T (the nuclear energy generation mechanisms all have a very steep dependence on T, so for all intents and purposes, all stars have roughly the same core temperature), we obtain the main sequence **mass-radius** scaling relation

$$R \propto M \tag{3.95}$$

Doing a more careful analysis to find how the temperature actually scales with the mass yields a slightly refined scaling relation

$$\boxed{R \propto M^{0.8}} \tag{3.96}$$

32. Make a dimensional analysis of the equation of radiative diffusion in stars to show that the luminosity of a star scales as its mass cubed, if the opacity is taken to be a constant and assuming that radiative diffusion is the dominant mechanism for energy transfer.

Looking at the equation of radiative diffusion (3.78), assuming a constant opacity:

$$\frac{\mathrm{d}T(r)}{\mathrm{d}r} \propto -\frac{L(r)\rho(r)}{r^2 T(r)^3} \tag{3.97}$$

$$L(r) \propto r^2 \frac{T(r)^3}{\rho(r)} \frac{\mathrm{d}T(r)}{\mathrm{d}r} \tag{3.98}$$

Evaluating everything at $r = R$ (so $\rho \propto M/R^3$) and approximating the derivative as T/R

$$L \propto R^2 \frac{R^3}{M} \frac{T^4}{R} \propto \frac{R^4 T^4}{M} \tag{3.99}$$

And using the ideal gas law ($P \propto \rho T$) and hydrostatic equilibrium:

$$\frac{\mathrm{d}P}{\mathrm{d}r} \sim \frac{P(0) - P(R)}{0 - R} \sim \frac{GM(M/R^3)}{R^2} \quad \longrightarrow \quad P(0) \sim \frac{M^2}{R^4} \tag{3.100}$$

Gives

$$L \propto \frac{R^4}{M}\left(\frac{P}{\rho}\right)^4 \propto \frac{R^4}{M}\left(\frac{M^2/R^4}{M/R^3}\right)^4 \propto \frac{R^4}{M}\left(\frac{M}{R}\right)^4 \quad \longrightarrow \quad \boxed{L \propto M^3} \tag{3.101}$$

Note: Considering the opacity more carefully leads to a few different scaling relations depending on the source of the opacity, since

$$L \propto \frac{M^3}{\kappa_R} \tag{3.102}$$

For example, Thomson scattering has $\kappa_R \sim$ const, so $L \propto M^3$ as before, but a Kramer's opacity law has

$$\kappa_R \propto \rho T^{-3.5} \propto \left(\frac{M}{R^3}\right)\left(\frac{M}{R}\right)^{-3.5} \propto M^{-2.5} R^{0.5} \propto M^{-2} \tag{3.103}$$

which gives $L \propto M^5$. The observed mass-luminosity relation for main sequence stars above $\sim 1\ M_\odot$ lies in between these two cases, with an exponent of ~ 3.5:

$$\boxed{L \propto M^{3.5}} \tag{3.104}$$

33. Use the known luminosity and mass of the Sun to estimate its nuclear lifetime. What is the current age of the sun, roughly?

The main sequence is by far the longest stage of a star's lifetime, so its total age can be estimated just from its lifetime on the main sequence when it's burning hydrogen. This can be estimated by taking the amount of energy the star could potentially produce by fusion, and dividing by its luminosity (energy released per unit time). Thus

$$\boxed{t_{\rm MS} \sim \frac{f X \eta M_* c^2}{L_*}} \tag{3.105}$$

Where

- $f \sim 0.1$ is the fraction of the star's Hydrogen that will ever get hot enough to be burned
- $X \sim 0.7$ is the mass fraction of Hydrogen—how much of the star's mass is actually Hydrogen
- $\eta = (m_{\rm He} - 4m_p)/4m_p \approx 0.00685$ is the efficiency of each nuclear reaction—how much of the mass-energy of each Hydrogen atom is actually converted into energy
- Thus, $f X \eta M_* c^2$ gives the total energy the star is expected to release from fusion over its main sequence lifetime

The mass of the Sun is $M_\odot \approx 1.989 \times 10^{33}$ g, and its luminosity is $L_\odot \approx 3.846 \times 10^{33}$ erg s^{-1}. So, doing an order of magnitude estimation,

$$t_{\rm MS} \sim \frac{10^{-1} \times 1 \times 10^{-2} \times 10^{33} \times (10^{10})^2}{10^{33}} \sim 10^{17}\,{\rm s} \tag{3.106}$$

Converting to Gyr

$$t_{\rm MS} \sim \frac{10^{17}\,{\rm s}}{10^7\,{\rm s\,yr^{-1}} \times 10^9\,{\rm yr\,Gyr^{-1}}} \sim \boxed{10\,{\rm Gyr}} \tag{3.107}$$

The Sun's current age, based on radioactive dating of meteorites in the Solar System and modeling of the solar spectrum, is estimated to be

$$\boxed{t \sim 4.5\,{\rm Gyr}} \tag{3.108}$$

So we've still got a good 5 billion years before we have to worry about being engulfed by the Sun as it enters the red giant phase[11,22].

For higher/lower mass stars, we can use the main sequence mass-luminosity scaling relation (§3.Q32) to see how the lifetime changes as a function of mass. Since $L \propto M^{3.5}$, we have

$$t_{\rm MS} \propto \frac{M}{M^{3.5}} \propto M^{-2.5} \tag{3.109}$$

However, for very low mass stars below $\lesssim 0.3\ M_\odot$, they are fully convective (see §3.Q37), which allows them to fuse closer to 100% of their hydrogen over their lifetimes, so $f \to 1$ and the lifetime increases by another factor of ~ 10.

34. Describe the prominent neutrino producing reactions in the sun, and the experiments designed to detect them. What was the solar neutrino problem and how was it solved?

Neutrinos are produced in the PP chain in the following reactions, all mediated by the weak force (see §3.Q35):

$$\begin{aligned} {}^1\mathrm{H} + {}^1\mathrm{H} &\longrightarrow {}^2\mathrm{H} + \mathrm{e}^+ + \nu_e \quad \text{(PP-I)} \\ {}^7\mathrm{Be} + \mathrm{e}^- &\longrightarrow {}^7\mathrm{Li} + \nu_e \quad \text{(PP-II)} \\ {}^8\mathrm{B} &\longrightarrow {}^8\mathrm{Be} + \mathrm{e}^+ + \nu_e \quad \text{(PP-III)} \end{aligned}$$

In the first one, two protons combine while one of them β-decays into a neutron, producing a deuterium nucleus along with a positron and an electron neutrino. This is the same logic for the third reaction except the proton that decays starts inside the boron nucleus. For the second one, we instead have a proton absorbing an electron to become a neutron, which also emits an electron neutrino. About 90% of the solar neutrinos come from the PP-I reaction, 10% from PP-II, and $\ll 1\%$ from PP-III.

The neutrinos are necessary in these processes to conserve energy, momentum, and angular momentum (spin). This is often grouped into a higher-order conservation law known as the "conservation of lepton number". In other words, in the first reaction we produce a positron, which we give a lepton number of -1. But the LHS has no leptons, so to conserve lepton number we need a particle on the RHS that has no charge but a lepton number of $+1$—hence, the neutrino.

In the 1960s, Ray Davis built a detector for solar neutrinos in the **Homestake** Mine in South Dakota (it was built underground in a mine to eliminate background signals as much as possible). The experiment consisted of a tank filled with cleaning fluid, which would detect neutrinos through the reaction

$$\nu_e + {}^{37}\mathrm{Cl} \longrightarrow {}^{37}\mathrm{Ar} + \mathrm{e}^-$$

The amount of Argon present could then be used as a proxy for the number of neutrinos. Despite their best efforts, the research team found that they were only getting about 1/3rd of the estimated number of neutrinos that they should be. This became known as the **solar neutrino problem**[36].

The solution to this problem—**neutrino oscillations**—was actually proposed before the problem itself was even found. The idea is that neutrinos can oscillate between their three flavors (ν_e, ν_μ, ν_τ), and since the experiment was only sensitive to electron neutrinos, the neutrino flux observed only accounted for about 1/3rd of the neutrinos that had actually originated from the Sun, while the others had changed flavors on their way to Earth[36].

Why do neutrinos oscillate between states?

Note: For this section, I use natural units to follow the conventions of the particle physics community ($c = \hbar = 1$). The reason neutrinos oscillate between flavors is because the **flavor eigenstates** that couple to the W/Z boson (ν_e, ν_μ, ν_τ) are different from the **mass eigenstates** that actually propagate (ν_1, ν_2, ν_3). The mixing between these eigenstates is characterized by the **PMNS matrix**

$$\begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix} = \begin{pmatrix} u_{1e} & u_{1\mu} & u_{1\tau} \\ u_{2e} & u_{2\mu} & u_{2\tau} \\ u_{3e} & u_{3\mu} & u_{3\tau} \end{pmatrix} \begin{pmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{pmatrix} \tag{3.110}$$

If we had a particle in a pure mass eigenstate that propagates with a 4-momentum $\mathbf{p}_j$, then it would evolve as

$$\left|\nu_j\right\rangle = e^{-i\mathbf{p}_j\cdot\mathbf{x}}\left|j\right\rangle = e^{-i(E_j t - p_j \ell)}\left|j\right\rangle \tag{3.111}$$

where t is the time it has propagated for and ℓ is the distance it has propagated. Therefore, we can write the eigenstate for the electron neutrino as a superposition of the mass eigenstates

$$\begin{aligned}\left|\nu_e\right\rangle &= u_{1e}\left|\nu_1\right\rangle + u_{2e}\left|\nu_2\right\rangle + u_{3e}\left|\nu_3\right\rangle &(3.112)\\ &= u_{1e}e^{-i(E_1 t - p_1\ell)}\left|1\right\rangle + u_{2e}e^{-i(E_2 t - p_2\ell)}\left|2\right\rangle + u_{3e}e^{-i(E_3 t - p_3\ell)}\left|3\right\rangle &(3.113)\end{aligned}$$

If the mass states' phases rotate at different frequencies, then they will interfere with each other, and the flavor of neutrino that's measured will depend on the distance and time that it has propagated for. We can find the phase shift between two states with

$$\delta\phi_{ij} = -(E_i t - p_i\ell) + (E_j t - p_j\ell) = -(E_i - E_j)t + (p_i - p_j)\ell \tag{3.114}$$

And if $p_i \gg m_i$

$$E_i = \sqrt{p_i^2 + m_i^2} = p_i\sqrt{1 + \left(\frac{m_i}{p_i}\right)^2} \simeq p_i\left[1 + \frac{1}{2}\left(\frac{m_i}{p_i}\right)^2\right] \simeq p_i + \frac{m_i^2}{2p_i} \tag{3.115}$$

Plugging in, and using $\ell \simeq ct$, we have

$$\delta\phi_{ij} = -\left(\frac{m_i^2}{2p_i} - \frac{m_j^2}{2p_j}\right)t \simeq \frac{\Delta m_{ij}^2}{2E}\ell \tag{3.116}$$

This tells us that the phase difference between mass eigenstates is proportional to the difference in their masses Δm_{ij}^2 relative to the energy they propagate with E, and it grows linearly with the distance ℓ. This allows us to measure the probability of a neutrino of flavor α turning into a different flavor β as a function of the distance ℓ:

$$P(\nu_\alpha \to \nu_\beta) = \left|\left\langle \nu_\alpha \middle| \nu_\beta \right\rangle\right|^2 = \left|\sum_j u_{j\alpha}^* u_{j\beta} e^{-im_j^2\ell/2E}\right|^2 \tag{3.117}$$

Which is often written as[37]

$$\begin{aligned}P(\nu_\alpha \to \nu_\beta) = \delta_{\alpha\beta} &- 4\sum_{j>k}\mathrm{Re}[u_{j\alpha}u_{j\beta}^* u_{k\alpha}^* u_{k\beta}]\sin^2\left[\frac{\Delta m_{jk}^2\ell}{4E}\right]\\ &+ 2\sum_{j>k}\mathrm{Im}[u_{j\alpha}u_{j\beta}^* u_{k\alpha}^* u_{k\beta}]\sin\left[\frac{\Delta m_{jk}^2\ell}{2E}\right]\end{aligned} \tag{3.118}$$

The first evidence of neutrino oscillations was found by the **Super-Kamiokande** experiment in 1998, which detected atmospheric neutrinos from cosmic ray events. Later in 2001, the **Sudbury Neutrino Observatory (SNO)** looked for reactions that were sensitive to all types of neutrinos as well as reactions only specific to electron neutrinos. They were able to measure that the fraction of solar neutrinos that were electron neutrinos was $\sim 34\%$, exactly matching the predictions of the neutrino oscillation theory. The 2015 Nobel Prize was jointly awarded to Arthur McDonald of the SNO and

Takaaki Kajita of Super-Kamiokande for the discovery of neutrino oscillations[37].

Neutrinos are also produced in the CNO cycle, but since the dominant source of the Sun's energy is the PP chain ($\sim 99\%$), neutrinos produced from the CNO cycle were only experimentally detected recently in 2020 by the **Borexino experiment**[38] at the Laboratori Nazionali del Gran Sasso in Italy. The relevant neutrino-producing reactions are

$$^{13}\mathrm{N} \longrightarrow {}^{13}\mathrm{C} + \mathrm{e}^{+} + \nu_e \tag{3.119}$$

$$^{15}\mathrm{O} \longrightarrow {}^{15}\mathrm{N} + \mathrm{e}^{+} + \nu_e \tag{3.120}$$

$$^{17}\mathrm{F} \longrightarrow {}^{17}\mathrm{O} + \mathrm{e}^{+} + \nu_e \tag{3.121}$$

These neutrinos are detected by elastic scatterings off of electrons—the recoiling electrons then produce a Cherenkov radiation spectrum that can be used to measure the energy of the neutrinos.

35. Write down the basic equations of the p-p chain that provides the Sun's nuclear power

The **proton-proton chain**, also called the **p-p chain** (hehe) is the set of nuclear reactions in the cores of stars that ultimately fuses Hydrogen into Helium. See this summarized in Figure 3.25.

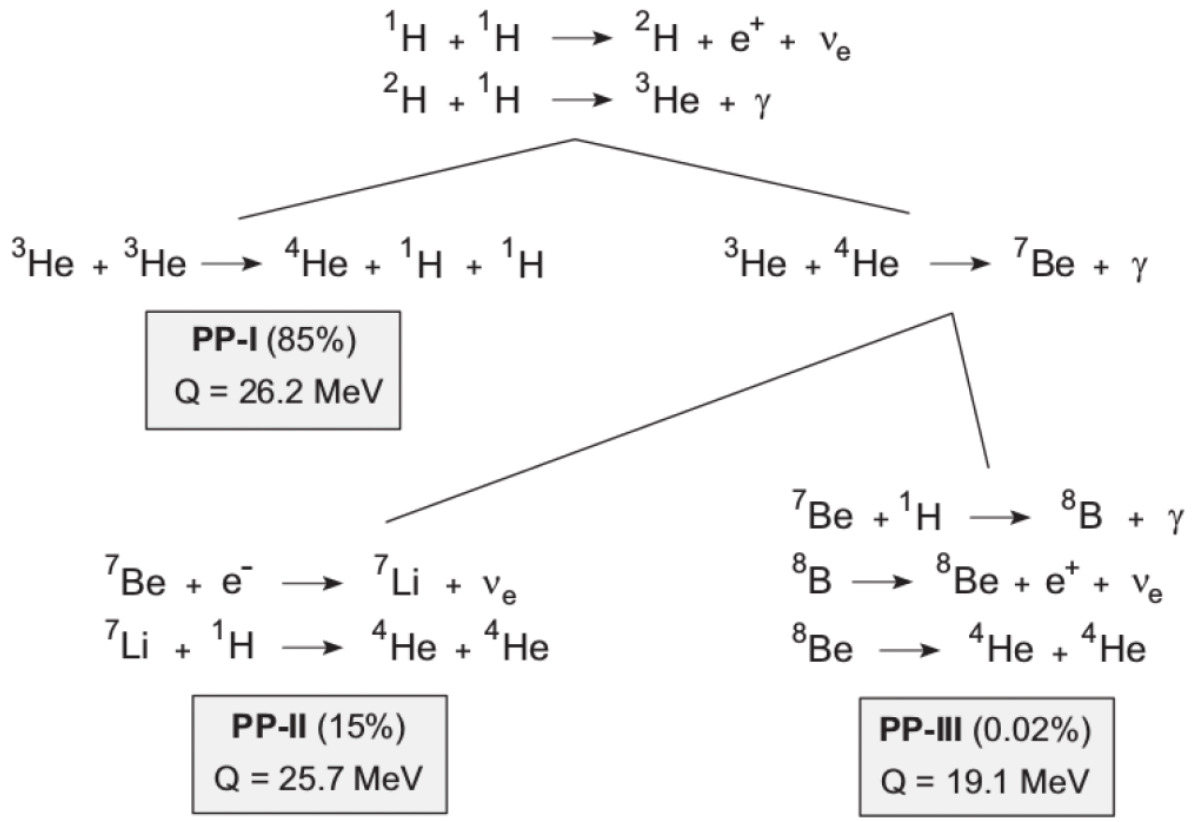


Figure 3.25: The reactions of the p-p chain[39].

The net output is

$$4\,{}^1_1\mathrm{H} \longrightarrow {}^4_2\mathrm{He} + 2\,\gamma + 2\,\mathrm{e}^+ + 2\,\nu_e \tag{3.122}$$

In the figure, the percentages indicate the amount of ^{4}He produced in each branch of the p-p chain. These are dependent on the temperature of the environment, with PP-II and PP-III becoming more dominant at hotter temperatures, but the values shown are for the Sun. Note that PP-I requires the first 2 reactions to happen twice, whereas PP-II and PP-III only require them to happen once (as it is assumed there is enough ^{4}He in the core already).

Any steps involving the weak force will be much slower than the others, and thus they will be the rate limiting steps. This includes the very first reaction where a proton has to β-decay into a neutron to form deuterium, the first step in PP-II where a proton in ^{7}Be absorbs an electron and becomes a neutron, and the second step in PP-III where a proton in ^{8}B β-decays a neutron, converting it to ^{8}Be.

The first step has a reaction rate of roughly 1 every Gyr, whereas the second step happens almost immediately (1 second).

The Q values indicate the net radiative energy that is released in each branch. In PP-I, in addition to the energy released from fusing 4 protons into ^{4}He, we also have 2 positrons that will annihilate with electrons and produce additional energy, and 2 neutrinos that will take some energy away. The net result is

$$Q_{\mathrm{PP-I}} = 4m_p c^2 \eta + 2m_e c^2 - 2E_\nu \tag{3.123}$$

The first two terms add up to 26.7 MeV, and on average $E_\nu = 0.262$ MeV, so $Q_{\mathrm{PP-I}} = 26.2$ MeV. For PP-II, since the first two reactions only happen once, we only produce one positron (but still 2 neutrinos since another one is produced in the PP-II reactions). Therefore

$$Q_{\mathrm{PP-II}} = 4m_p c^2 \eta + m_e c^2 - 2E_\nu \tag{3.124}$$

which gives $Q_{\mathrm{PP-II}} = 25.7$ MeV. Finally, in PP-III, we should normally expect Q to be the same as PP-II, but in fact the average energy of the neutrino produced in the PP-III branch is much higher, so we get a $Q_{\mathrm{PP-III}} = 19.1$ MeV[39].

We can convert the Q values into an energy generation rate per unit volume, ε_V, like so

$$\varepsilon_V = QR_{AB} \ , \ R_{AB} = n_A n_B \langle \sigma v \rangle \left(\frac{1}{1+\delta_{AB}} \right) \tag{3.125}$$

Here, n_A and n_B are the number densities of the two reactants (both protons in this case), $\langle \sigma v \rangle$ is the velocity-averaged cross section, and the factor with δ_{AB} is 1 if A and B are the same and 0 if they are not (this prevents us from double counting). We typically write the cross section as

$$\sigma(E) = \frac{S(E)}{E} \exp\left(-\sqrt{\frac{E_G}{E}} \right) \ , \ E_G = (\pi Z_1 Z_2 \alpha)^2 2\mu c^2 \tag{3.126}$$

where $S(E)$ is constrained by nuclear physics, E_G is the **Gamow energy**, and α is the fine structure constant. The velocity is given by a thermal distribution, so we can use Maxwell-Boltzmann:

$$f(E)\mathrm{d}E = \frac{2}{\sqrt{\pi}} \frac{E^{1/2}}{(kT)^{3/2}} e^{-E/kT} \mathrm{d}E \tag{3.127}$$

Therefore the average of σv is

$$\langle \sigma v \rangle = \int_0^\infty \mathrm{d}E \, f(E) \sigma(E) v(E) \tag{3.128}$$

$$= \frac{2}{\sqrt{\pi}} \frac{1}{(kT)^{3/2}} \int_0^\infty \mathrm{d}E \, E^{1/2} \times \frac{S(E)}{E} \times \left(\frac{2E}{\mu} \right)^{1/2} \exp\left(-\frac{E}{kT} \right) \exp\left(-\sqrt{\frac{E_G}{E}} \right) \tag{3.129}$$

$$= \sqrt{\frac{8}{\mu\pi}} \frac{1}{(kT)^{3/2}} \int_0^\infty \mathrm{d}E \, S(E) \exp\left(-\frac{E}{kT} - \sqrt{\frac{E_G}{E}} \right) \tag{3.130}$$

The $e^{-E/kT}$ term goes to 0 at high E whereas the $e^{-\sqrt{E_G/E}}$ term goes to 0 at low E, so the integral is only significant around an intermediate energy of $E_0 = (E_G k^2 T^2/4)^{1/3}$, called the **Gamow peak** (see Figure 3.26.

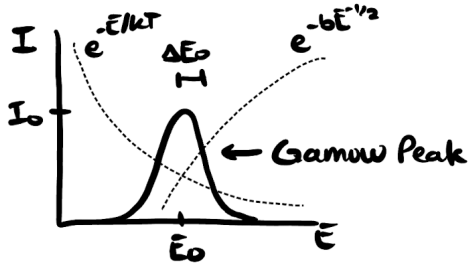


Figure 3.26: Illustration of the Gamow peak.

This allows us to rewrite the exponential argument as a second-order Taylor expansion, which turns it into a Gaussian

$$\exp\left(-\frac{E}{kT}-\sqrt{\frac{E_G}{E}}\right)\approx C\exp\left[-\left(\frac{E-E_0}{\Delta E/2}\right)^2\right] \tag{3.131}$$

for which the integral is easy

$$\int_0^\infty \mathrm{d}E\, C\exp\left[-\left(\frac{E-E_0}{\Delta E/2}\right)^2\right]=C\sqrt{\pi}\frac{\Delta E}{2} \tag{3.132}$$

Skipping the algebra, this results in a reaction rate per unit volume of approximately:

$$R_{AB}\approx\frac{8}{\sqrt{3}\pi\alpha m_p c}\frac{(1+\delta_{AB})^{-1}}{Z_A Z_B A_r}n_A n_B S(E_0)\left(\frac{E_G}{4kT}\right)^{2/3}\exp\left[-3\left(\frac{E_G}{4kT}\right)^{1/3}\right] \tag{3.133}$$

where $A_r=\mu/m_p$ is the reduced mass number[39]. This can be further simplified to give the energy rate per unit mass for p-p chain reactions. This becomes complicated when considering the relative contributions from the different branches, but an approximate solution looks something like[40]

$$\boxed{\varepsilon_m(\mathrm{PP})\approx 2.57\times10^4\psi f_{11}g_{11}X^2\left(\frac{\rho}{\mathrm{g\,cm^{-3}}}\right)T_9^{-2/3}\exp\left(-3.381T_9^{-1/3}\right)\,\mathrm{erg\,s^{-1}\,g^{-1}}} \tag{3.134}$$

where ψ accounts for the relative contributions of PP-I, PP-II, and PP-III, and the factor g_{11} does not have any physical significance, it's just a fitted polynomial correction factor that does a good job at reproducing observed results:

$$g_{11}=1+3.82T_9+1.51T_9^2+0.144T_9^3-0.0114T_9^4 \tag{3.135}$$

This has the rough proportionality

$$\boxed{\varepsilon_m(\mathrm{PP})\propto\rho X^2T^4} \tag{3.136}$$

There is an additional possible avenue for the first reaction, called the PEP branch, which looks like

$$^1\mathrm{H}+\mathrm{e}^-+{}^1\mathrm{H}\longrightarrow{}^2\mathrm{H}+\nu_e \tag{3.137}$$

but this only happens in the Sun about 1/400 as often as the canonical first reaction in the PP-chain.

36. How does the CNO cycle work?

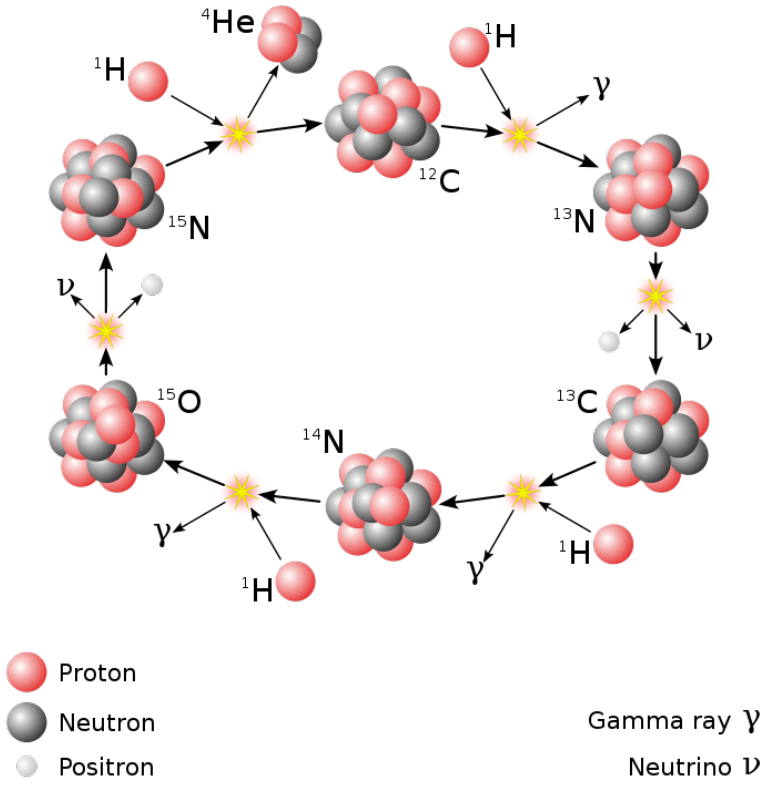


Figure 3.27: The CNO cycle

The reactions of the CNO cycle go as follows

$$
\begin{aligned}
{}^{12}_{6}\mathrm{C} + {}^{1}_{1}\mathrm{H} &\longrightarrow {}^{13}_{7}\mathrm{N} + \gamma \\
{}^{13}_{7}\mathrm{N} &\longrightarrow {}^{13}_{6}\mathrm{C} + \mathrm{e}^{+} + \nu_e \\
{}^{13}_{6}\mathrm{C} + {}^{1}_{1}\mathrm{H} &\longrightarrow {}^{14}_{7}\mathrm{N} + \gamma \\
{}^{14}_{7}\mathrm{N} + {}^{1}_{1}\mathrm{H} &\longrightarrow {}^{15}_{8}\mathrm{O} + \gamma \quad *** \\
{}^{15}_{8}\mathrm{O} &\longrightarrow {}^{15}_{7}\mathrm{N} + \mathrm{e}^{+} + \nu_e \\
{}^{15}_{7}\mathrm{N} + {}^{1}_{1}\mathrm{H} &\longrightarrow {}^{12}_{6}\mathrm{C} + {}^{4}_{2}\mathrm{He}
\end{aligned}
$$

The general pattern is that we add a proton to go up 1 element, convert the proton into a neutron to become an isotope of the previous element, and then add another proton to become a more stable isotope of the next element. Repeat until we have enough extra protons to make a ^{4}He nucleus. The step annotated with the *** is the rate-limiting step here, since at temperatures found in the cores of stars, the Coulomb barrier between the protons is very large. The net reaction is

$$4\,{}^{1}_{1}\mathrm{H} \longrightarrow {}^{4}_{2}\mathrm{He} + 3\,\gamma + 2\,\mathrm{e}^{+} + 2\,\nu_e \tag{3.138}$$

There is also a second branch of the CNO cycle where in the last step we produce an ${}^{16}_{8}$O instead of

a helium nucleus:

$$
\begin{aligned}
{}^{15}_{7}\mathrm{N} + {}^{1}_{1}\mathrm{H} &\longrightarrow {}^{16}_{8}\mathrm{O} + \gamma \\
{}^{16}_{8}\mathrm{O} + {}^{1}_{1}\mathrm{H} &\longrightarrow {}^{17}_{9}\mathrm{F} + \gamma \\
{}^{17}_{9}\mathrm{F} &\longrightarrow {}^{17}_{8}\mathrm{O} + \mathrm{e}^{+} + \nu_e \\
{}^{17}_{8}\mathrm{O} + {}^{1}_{1}\mathrm{H} &\longrightarrow {}^{14}_{7}\mathrm{N} + {}^{4}_{2}\mathrm{He}
\end{aligned}
$$

The relative rates of these two processes are determined by the temperature of the environment, and the second cycle requires the first two proton captures to happen in very rapid succession to prevent the ^{16}O from decaying first. Typically, the first cycle happens 99.96% of the time and the second cycle happens the other 0.04%.

The reason this is a "cycle" is because we start with a carbon and end with a carbon, so the carbon has no net gain or loss (we just use up a few protons along the way). The second branch, similarly, produces a nitrogen which we use to restart the cycle at an earlier step. See this visually in Figure 3.27.

Why is it preferred to decay into an α particle (^{4}He nucleus)? This has to do with the binding energy per nucleon. ^{4}He is one of the most stable nuclei that exists at low atomic numbers, and has one of the highest binding energies per nucleon relative to those around it. It produces a noticeable spike in the chart of binding energy per nucleon vs. mass number (Figure 3.28).

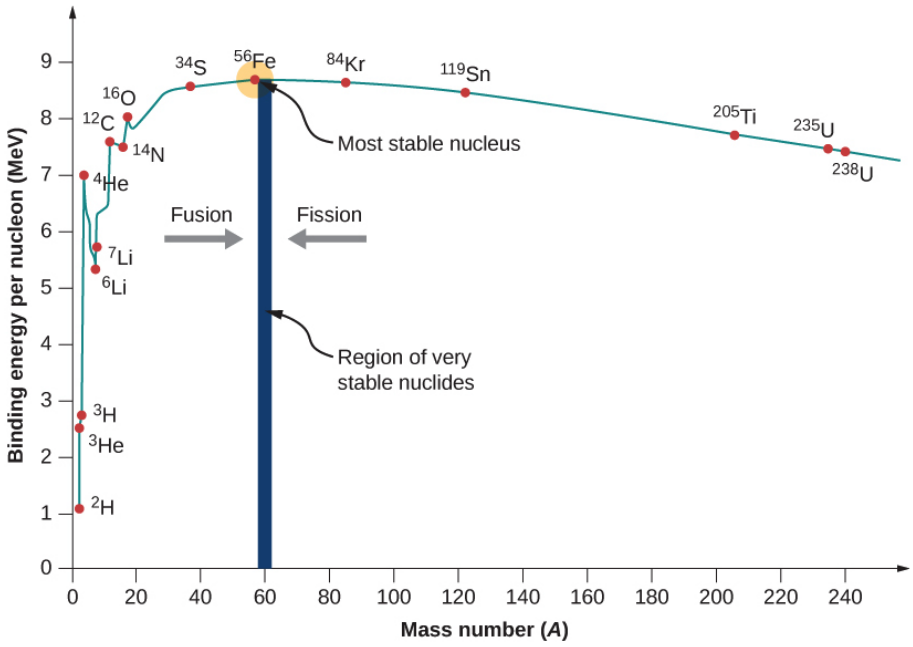


Figure 3.28: A plot of the binding energy per nucleon vs. the mass number. Notice the spike at ^{4}He, and the peak of all elements at ^{56}Fe. For isotopes below ^{56}Fe, fusion is exothermic, whereas for isotopes above ^{56}Fe, fission is exothermic[41].

The energy generation rate per unit mass for the CNO cycle is approximately[40]

$$
\boxed{\varepsilon_m(\mathrm{CNO}) \approx 8.24 \times 10^{25} g_{14,1} X_{\mathrm{CNO}} X \left(\frac{\rho}{\mathrm{g\,cm^{-3}}}\right) T_9^{-2/3} \exp\left[-15.231 T_9^{-1/3} - (T_9/0.8)^2\right] \mathrm{erg\,s^{-1}\,g^{-1}}} \tag{3.139}
$$

where

$$g_{14,1} = 1 - 2.00T_9 + 3.41T_9^2 - 2.43T_9^3 \tag{3.140}$$

With the rough proportionality

$$\boxed{\varepsilon_m(\text{CNO}) \propto \rho X_{\text{CNO}} X T^{20}} \tag{3.141}$$

Comparing this to the pp chain, one sees that the CNO cycle does not start becoming important until much higher temperatures, but it rises more quickly than the pp chain (see Figure 3.29). The turnover point is at around $T \sim 2 \times 10^7$ K. The Sun, with a core temperature of 1.5×10^7 K, generates about 1% of its energy from the CNO cycle.

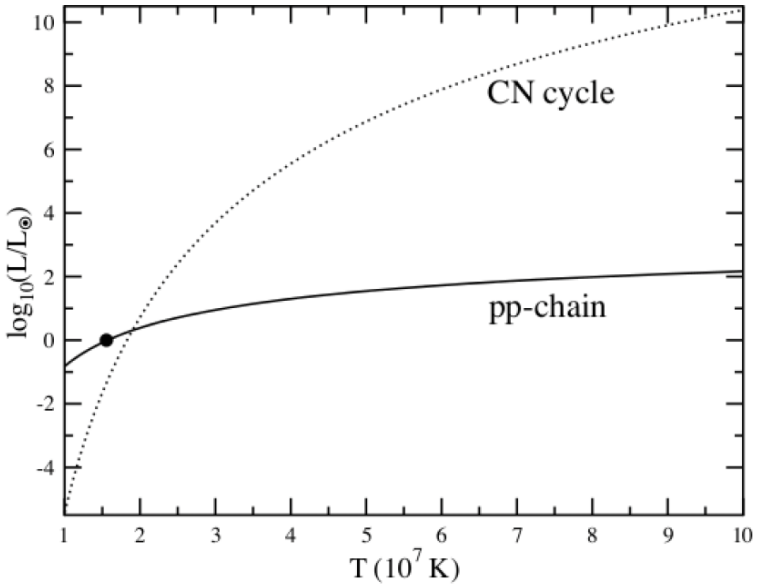


Figure 3.29: The energy generation rates of the pp-chain and CNO cycle as a function of temperature[39,42]. The Sun's core temperature is labeled by the dot.

Note: It is not mentioned in either of the last two questions, but there is also a third thermonuclear energy generation process in stars for *Helium fusion* that is at least worth knowing about: the **triple-alpha process**. This is explained in Appendix §12.8.1, along with a few other nuclear processes not explained in the questions.

37. Describe the internal structure of the sun. What is the Schwarzschild criterion for convective instability, and how does it delineate the convective and radiative zones in the sun?

The interior of the Sun is primarily composed of 3 layers[39]

1. **The Core**: This is the innermost layer where nuclear fusion happens.
   - $R \sim 0 - 0.3\, R_\odot$
   - $T \sim (7-15) \times 10^6$ K
   - $\rho \sim 10 - 150\ \mathrm{g\,cm^{-3}}$
2. **The Radiative Zone**: Surrounding the core; energy transport is dominated by radiation.
   - $R \sim 0.3 - 0.7\, R_\odot$
   - $T \sim (3-7) \times 10^6$ K
   - $\rho \sim 1 - 10\ \mathrm{g\,cm^{-3}}$
3. **The Convective Zone**: The outermost layer; energy transport is dominated by convection.
   - $R \sim 0.7 - 1\, R_\odot$
   - $T \sim 5.7 \times 10^3 - 3 \times 10^6$ K
   - $\rho \sim 1\ \mathrm{g\,cm^{-3}}$

See Figure 3.30.

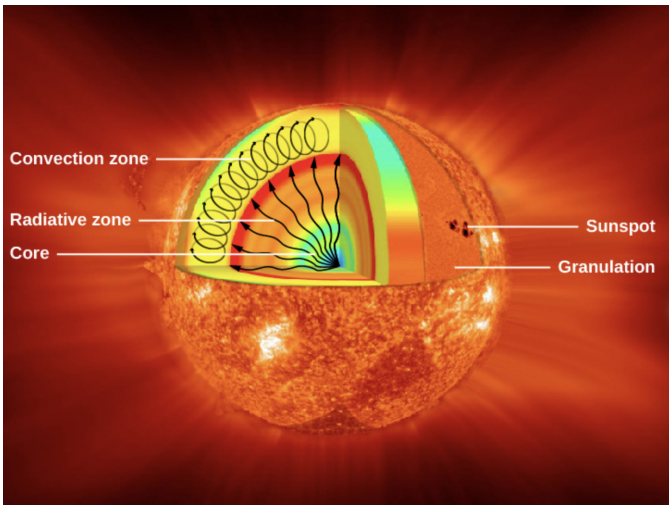


Figure 3.30: An illustration of the interior layers of the Sun[43].

The **Schwarzschild criterion** for convective instability is

$$\boxed{\left|\frac{\mathrm{d}T}{\mathrm{d}r}\right| < \left(1 - \frac{1}{\gamma}\right)\frac{T}{P}\frac{\mathrm{d}P}{\mathrm{d}r}} \tag{3.142}$$

For the derivation of this, see §3.Q30, particularly the derivation of equation (3.84). The RHS of this equation is the critical value of the temperature gradient where convection starts to become dominant. So if the actual temperature gradient is less than this critical value, convection is **unstable**

and radiation will be the dominant energy transport mechanism. And when the actual temperature gradient is larger, convection is **stable** so it will dominate.

If we assume the energy transport is dominated by radiation and plug this into (3.142), we get the criterion for convection instability to be

$$\frac{\gamma}{\gamma-1}\frac{3}{64\pi\sigma_{\text{SB}}Gm_p}\frac{\rho\kappa_R}{\mu T^3}\frac{L(r)}{M(r)} < 1 \tag{3.143}$$

This reveals a few different reasons why convection might become stable (causing this instability criterion to become false)[39]:

1. The opacity κ_R is large, which tends to happen in the cooler, outer layers of stars where bound-free absorption occurs.
2. The adiabatic index γ approaches 1, which may occur where H or He is partially ionized, also in the outer layers of stars. This happens because rising parcels of gas cannot cool as much as they normally would since electrons recombine with ions and release energy. Therefore the density of parcels does not decrease as quickly, and they remain buoyant.
3. The ratio $L(r)/M(r)$ is large, which can happen near the cores of stars that fuse using the CNO cycle, which is extremely sensitive to the temperature and thus only occurs in a very small core region close to the center. The steep temperature gradient forces convection to occur.

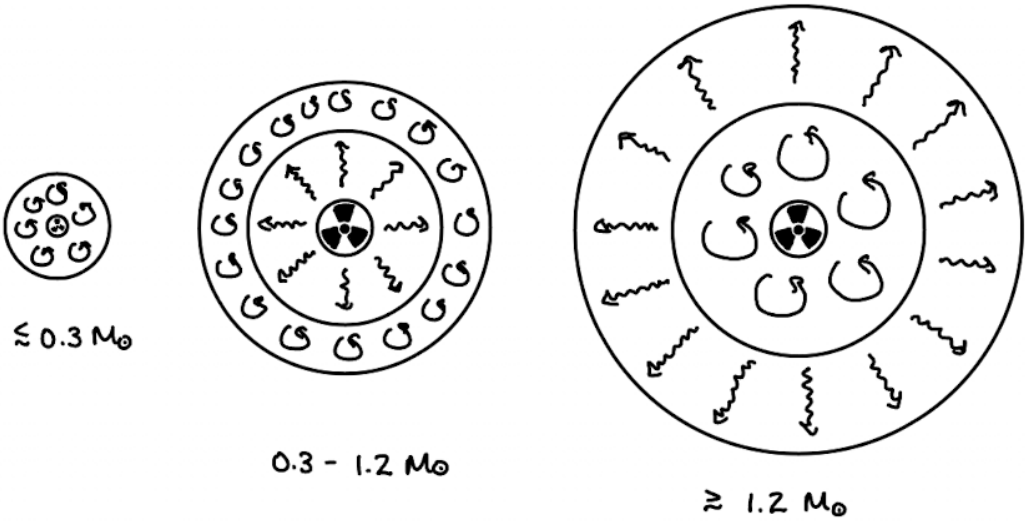


Figure 3.31: The different interior structures of stars in three different mass regimes. Very low mass stars have only convective zones, intermediate mass stars have inner radiative and outer convective zones, and high mass stars have inner convective zones and outer radiative zones.

We can see here that the first two will be dominant in stars that fuse with the pp chain, meaning they will only have convection in their outer layers. In contrast, the third will be dominant in stars that fuse with the CNO cycle, meaning they will only have convection close to their cores. Recall that the CNO cycle requires higher core temperatures than the pp chain (§3.Q36), which translates to higher stellar masses. Thus, more massive stars have their convective and radiative zones swapped relative to the Sun[39]. Very low mass stars, on the other hand, *only* have a convective zone, due to having

a relatively cooler temperature and higher opacity. See the relative configurations of the convective and radiative zones as a function of stellar mass schematically in Figure 3.31, and mathematically in Figure 3.32.

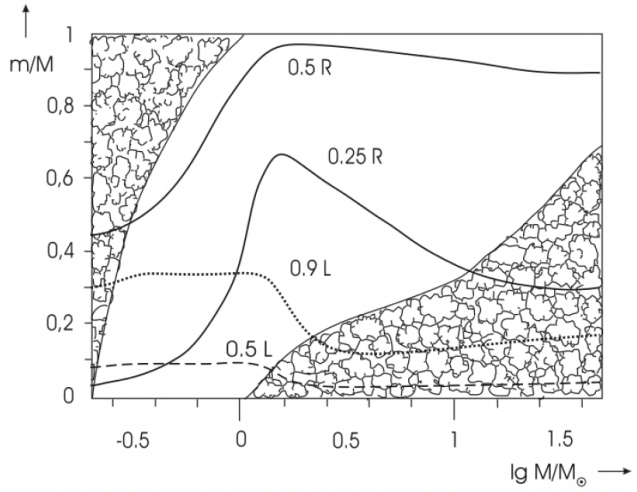


Figure 3.32: The extent of the convective zone as a function of stellar mass. The x axis is $\log_{10}(M/M_\odot)$ and the y axis is the mass fraction $M(r)/M(R)$ which is effectively a measure of the radius of the star. The cloudy shaded regions show where convection is active[40].

One of the factors described above, the opacity, has a few different sources contributing to it. There is bound-bound absorption (i.e. absorption lines that can be pressure broadened), bound-free absorption, free-free absorption, and electron scattering. For a detailed summary of each of these processes, see §5.Q83. The opacity as a function of temperature (which is a proxy for radius) is shown in Figure 3.33. Moving in from the surface, we see that the opacity has a strong peak due to H^- opacity (bound-free absorption) as more free electrons become available from ionized metals to make the H^- ion. But once we keep getting hotter, H^- ions can no longer be created, and the opacity becomes dominated by free-free and bound-free absorption, which both follow Kramer's opacity laws ($\kappa \propto \rho T^{-3.5}$). At the highest temperatures, electron scattering dominates and the curve flattens out, since the opacity for electron scattering is independent of temperature. The tiny dip at the very high temperatures ($T \sim 10^8$) is when we transition from Thomson scattering to Compton scattering, which has a marginally reduced opacity[39].

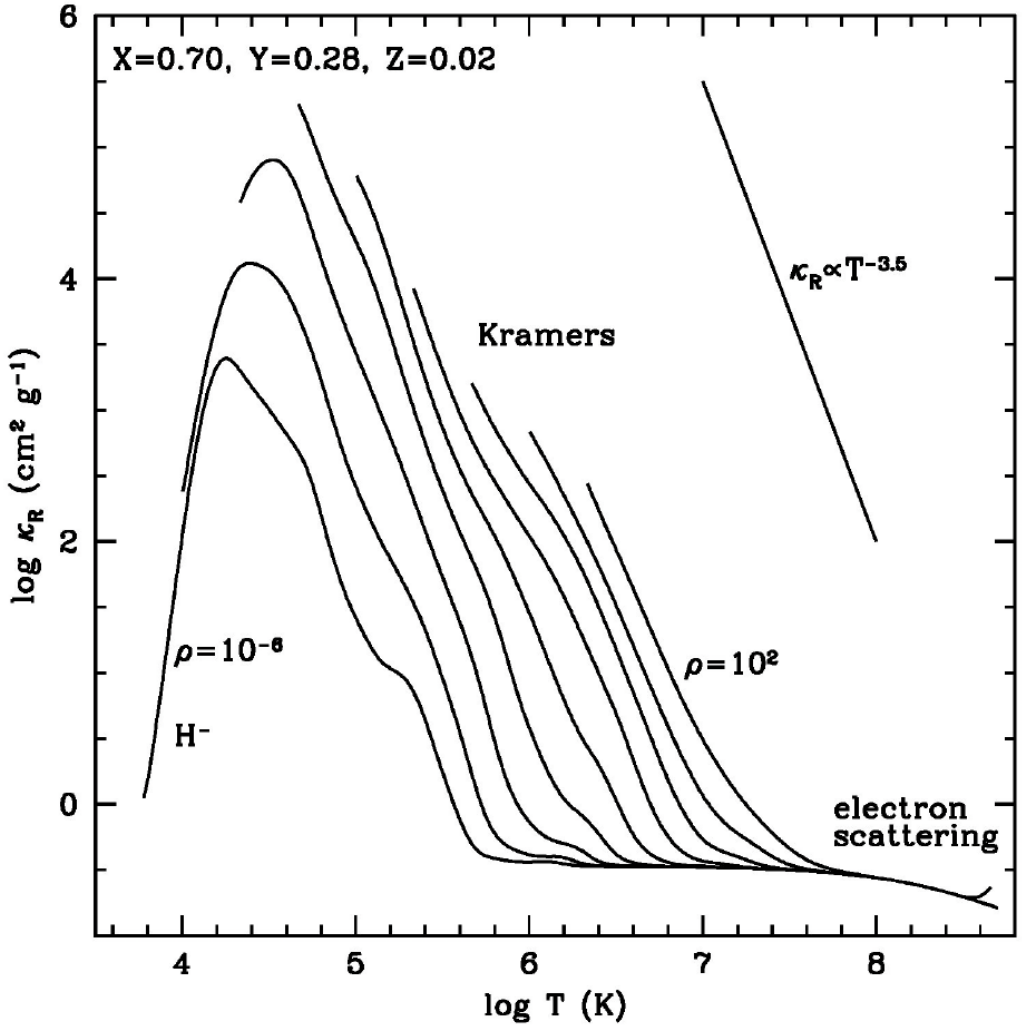
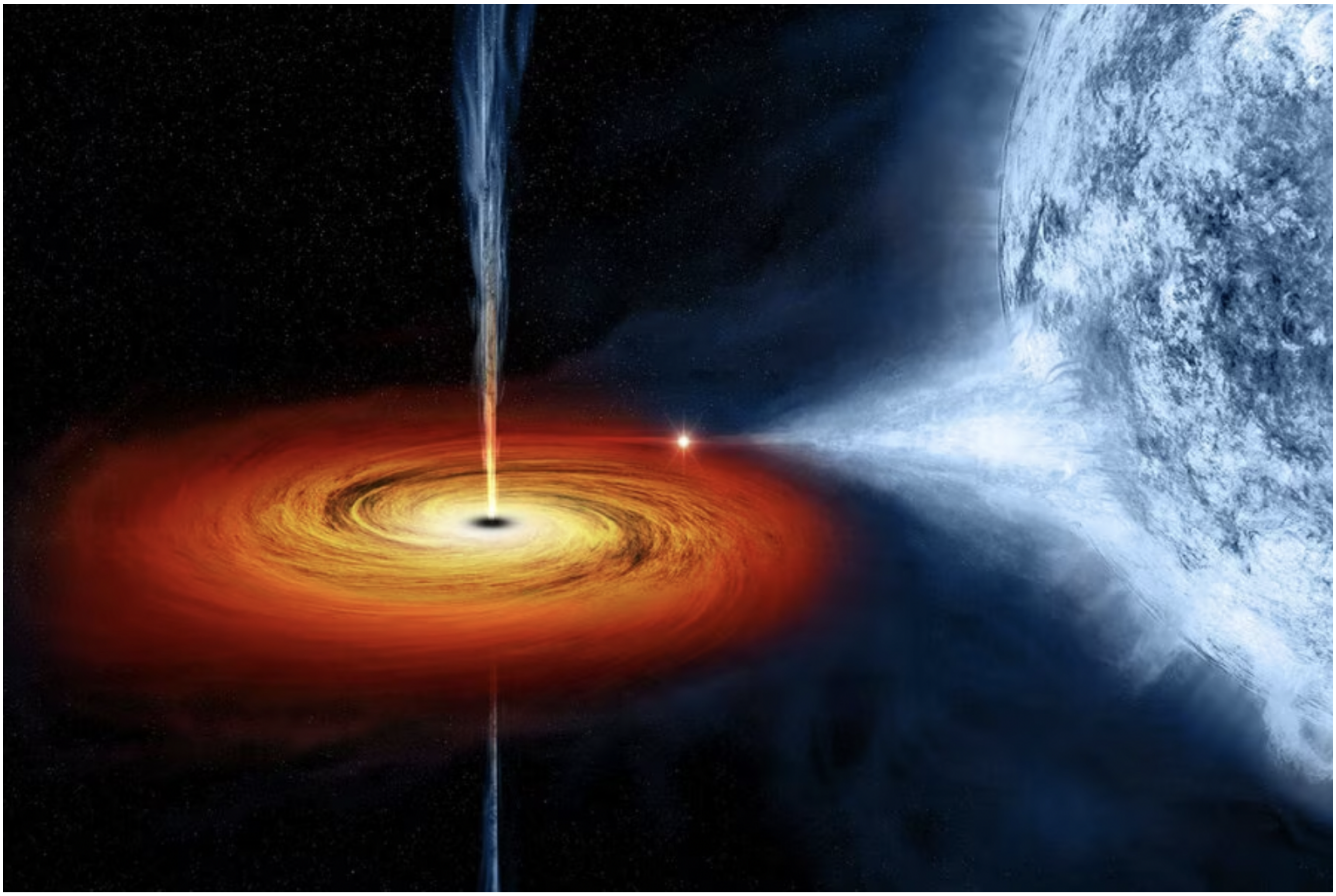


Figure 3.33: Opacity as a function of temperature in a Sun-like star[39]. Courtesy of the OPAL project[44,45].

4 Compact Objects and Gravitational Waves



38. What do we know about the masses and spins of (i) stellar mass black holes and (ii) supermassive black holes, and how do we know these things?

(i) Stellar mass black holes

- From XRBs: $M \sim 10\ M_\odot$, rapidly spinning
- From LIGO: $M \sim 30\ M_\odot$, slowly spinning

Masses

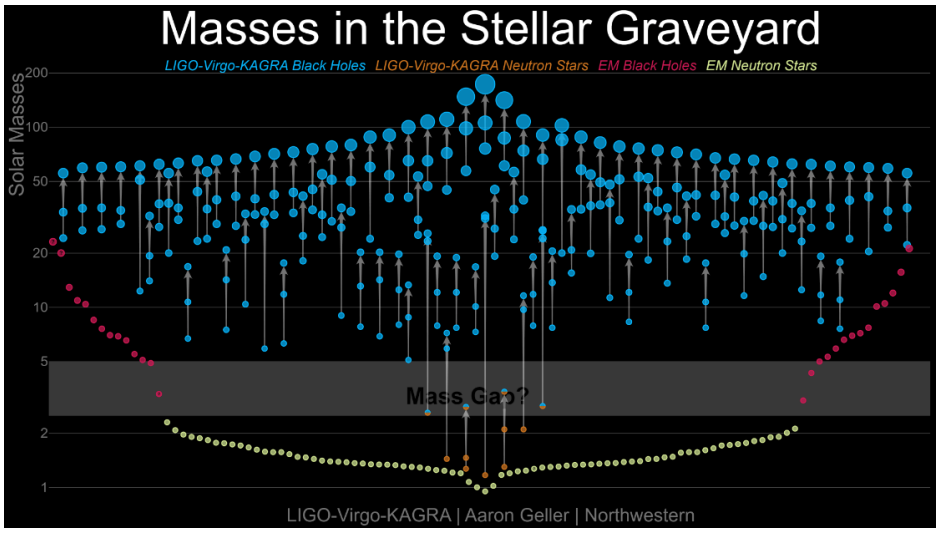


Figure 4.1: The masses of black holes detected with EM observations and with LIGO. Credit: LIGO-Virgo-KAGRA / Aaron Geller / Northwestern

The distribution of masses from EM observations and from LIGO are shown in Figure 4.1. Notice there are two gaps: the lower mass gap from 2–5 $M_\odot$, and the upper mass gap from 60–150 $M_\odot$. These are gaps that are constrained by the theory. Based on late-stage stellar evolution models, we think black holes cannot form with masses $\lesssim 2-3\ M_\odot$ or $\gtrsim 45-60\ M_\odot$, but the precise values have a lot of uncertainty. The observations mostly back up the lower mass gap, with almost no black holes being found under 2 solar masses. However, the upper mass gap *does* have observed black holes in it, so there is active research going on to discover ways of populating this mass gap. The main ways these black holes are thought to be formed is through hierarchical mergers, accretion from a companion, stellar collisions before BH collapse, etc.

The masses of black holes in XRBs (X-Ray Binaries) can be measured with radial velocities from a luminous companion using the binary mass function (see §3.Q20). In this case M_2 is the black hole.

$$K = \left(\frac{2\pi G}{P}\right)^{1/3} \frac{M_2 \sin i}{(M_1 + M_2)^{2/3}} \frac{1}{\sqrt{1-e^2}} \tag{4.1}$$

For gravitational wave sources with LIGO, the evolution of the frequency as two compact objects inspiral is described by the differential equation (4.43) (see §4.Q43). This depends on a quantity called the **chirp mass** (4.41), which can be measured with ν and $\dot{\nu}$, both of which are direct observables

from the waveform

$$\mathscr{M} = \frac{c^3}{G}\left[\frac{5}{96}\pi^{-8/3}\nu^{-11/3}\dot{\nu}\right]^{3/5} \tag{4.2}$$

We can also use Kepler's 3rd law to find the frequency at some instantaneous point in time before the merger, as long as we choose some point long enough before the merger that Newtonian gravity still applies. In that case, say we choose a point where the objects are at each other's ISCO (§4.Q46), then we have

$$\omega = \sqrt{\frac{GM}{r_{\rm ISCO}^3}} = \frac{c^3}{6^{3/2}GM} \quad \longrightarrow \quad \omega_{\rm gw} = 2\omega = \frac{2c^3}{6^{3/2}GM} \tag{4.3}$$

This gives a second independent constraint on the masses, so we can use (4.2) and (4.3) to solve for $\mathscr{M}$ and M, which can then be converted into M_1 and M_2.

Spins

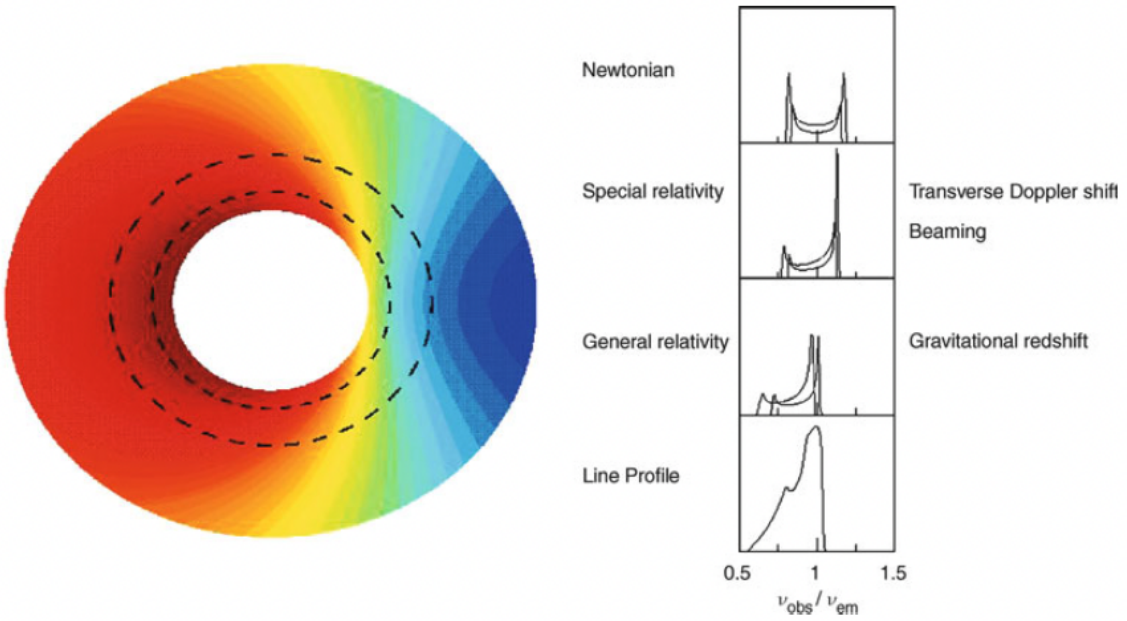


Figure 4.2: The different effects that contribute to the observed line profile in the X-ray reflection spectrum from an accretion disk. The observer is looking up from the bottom of the image. The different effects are explained in the text[46].

Spins can be measured a few different ways. First, some definitions. We define the **spin parameter** a as the dimensionless ratio

$$a \equiv \frac{Jc}{GM^2} \tag{4.4}$$

where J is the angular momentum.

For XRBs, if they are accreting enough ($\dot{M} \sim 0.01 - 0.3\,\dot{M}_{\rm Edd}$), we can use **X-ray reflection spectroscopy**. This technique takes advantage of the fact that there is a hot corona surrounding the accretion disk, which photoionizes the outer layers of the optically thick disk and creates a non-

thermal reflection spectrum. This spectrum is then distorted by the effects of the disk's rotation, which can be directly linked to the spin of the black hole. See Figure 4.2[47].

In a purely Newtonian world, we would see a double-peaked emission line due to the portion of the disk rotating away from us and the portion rotating towards us at the same speed (Doppler shift). Adding in the effects of special relativity, we also get a relativistic beaming effect for the gas closest to the black hole (see §5.Q75) which enhances the blue peak, and a transverse Doppler shift (i.e. time dilation) that shifts everything to the red side. Finally, in general relativity we also have a gravitational redshift which further shifts everything to the red. If the black hole itself is rotating, this changes the distance of the innermost stable circular orbit (ISCO) and smears out the line profiles (see Figure 4.3).

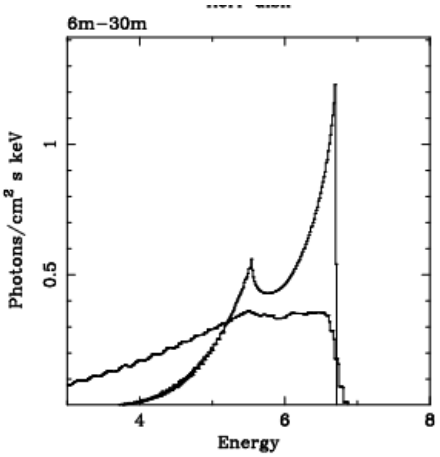


Figure 4.3: A comparison between the line profile shape for a non-rotating vs. a rotating black hole[47].

Another method commonly used is **thermal continuum fitting**, which takes advantage of the same logic that the spin of the black hole affects the location of the ISCO. This also affects the temperature of the disk, which affects its radiated power. The power per unit area of the disk (i.e. intensity) can be found in the Newtonian limit to be

$$I(r) = \frac{3GM\dot{M}}{4\pi r^3}\left(1 - \sqrt{\frac{r_{\text{isco}}(a)}{r}}\right) \tag{4.5}$$

The $r_{\text{isco}}(a)$ gets smaller as a gets larger (i.e. as the black hole spins faster), so the inner disk gets hotter and radiates more energy. The temperature can be obtained by $I(r) = 2\sigma_{\text{SB}} T_{\text{eff}}^4$, where the factor of 2 accounts for the 2 sides of the disk[47].

Finally, for LIGO sources, we usually measure the mass-weighted **effective spin parameter**

$$\chi_{\text{eff}} = \frac{M_1 \mathbf{a}_1 + M_2 \mathbf{a}_2}{M_1 + M_2} \cdot \hat{\mathbf{L}} \tag{4.6}$$

This alone does not characterize the spins of both individual black holes, since a $|\chi_{\text{eff}}| \ll 1$ could be due to both black holes having small spins, both having significant spins in the $\hat{\mathbf{L}}$ direction but misaligned from each other, or significant spins not in the $\hat{\mathbf{L}}$ direction. However, if $\chi_{\text{eff}} = \pm 1$, we know that both black holes have maximal spins in the same direction, and they are aligned $(+1)$ or antialigned (-1) with the orbital angular momentum. LIGO results have shown mostly low effective spin parameters. We can also measure the spin precession χ_p, but this is mostly subdominant[47].

(ii) Supermassive black holes

- $M \sim 10^6 - 10^{10}\ M_\odot$, $a \sim$ maximal

Masses

There are a number of ways to measure SMBH masses. In the unique case of the Milky Way, we can directly resolve the orbits of stars around the central SMBH and infer a mass from the dynamics, obtaining $M \sim 4 \times 10^6\ M_\odot$.

More commonly, we use the ***M*–σ relation** (see §7.Q124), which relates the mass of the black hole to the velocity dispersion in the bulge: $M \propto \sigma^\gamma$ where γ is somewhere between 4–5.

There is also **reverberation mapping**. This involves measuring the time delay between when the continuum varies in brightness and when optical broad lines vary in response (the lines are excited in response to emission from the accretion disk). The time delay τ can be used to find the radius of the broad line region by $R_{\rm BLR} = c\tau$. Then, we can use $R_{\rm BLR}$ and the width of the broad lines to esimate the black hole mass:

$$M_{\rm BH} = \frac{f(\Delta v)^2}{G} R_{\rm BLR} \tag{4.7}$$

The factor of f here is a dimensionless factor that depends on the geometry and is typically ~ 5. This technique requires high time resolution (days) over the course of long monitoring durations (months), with a moderate spectral resolution (600 $\mathrm{km\,s^{-1}}$), and a high SNR. For these reasons, it has only been used on 10s of sources[48].

Reveberation mapping can also be done in the X-ray, measuring the time delay between light emitted from the corona and reflecting off of the disk. The earliest light can reflect off the disk is if it hits right at the ISCO, so this technique probes $R_{\rm ISCO}$, which is related to the black hole mass[49].

Finally, for a couple of nearby sources, we've measured black hole masses by probing their shadows with the Event Horizon Telescope (EHT) and Very Long Baseline Interferometry (VLBI). The size of a black hole's shadow is related to its mass[50].

Spins

These can be measured with **X-ray reflection spectroscopy**, just like for stellar mass black holes. However, we cannot use thermal continuum fitting since the temperatures will usually be too low ($kT \lesssim 0.05$ keV). These spins are usually found to be very close to maximal (see Figure 4.4). Note: why do we care about spins? They can give us information on the formation channels of black holes and how much angular momentum their progenitors had.

BONUS: Intermediate Mass Black Holes (IMBHs)

Notice the gap between stellar mass black holes of up to $10^2\ M_\odot$, and SMBHs at greater than $10^6\ M_\odot$. So, what about the 4 orders of magnitude in between? This is where IMBHs should live, but to date no black holes in this mass range have been confirmed. It's a huge mystery since, theoretically, SMBHs should form through mergers from stellar mass black holes, so we should be able to find black holes in the intermediate stages of evolution, but we don't. As my undergraduate advisor once said, it's like we have a population of just babies and old people. The current hypothesis seems to be that SMBHs had a different, more efficient formation channel in the early universe that allowed them to grow rapidly (Pop III stars?), but there is still a lot that we don't know about this.

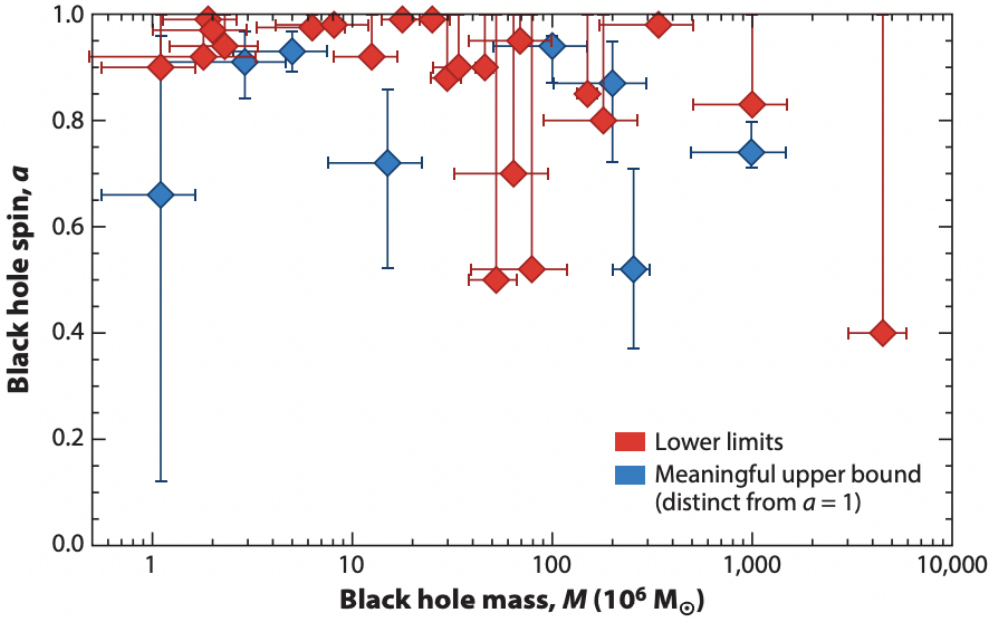


Figure 4.4: The observed spins of black holes as a function of mass[47].

39. What properties of supermassive black holes correlate with properties of the host galaxy?

The black hole mass correlates with the velocity dispersion in the bulge. This is known as the ***M*–*σ* relation**. See §7.Q124.

The **feedback** from the SMBH also correlates with the **star formation rate** in the host galaxy, either suppressing it (negative feedback) or enhancing it (positive feedback). In this case there isn't a simple numerical relationship since feedback processes are complex and dynamic. When feedback turns off star formation, this is called **"quenching"**. For more info on AGN feedback, see §7.Q123.

40. It is believed that most stars leave a collapsed remnant at the end of their evolution. What stars leave (i) white dwarfs? (ii) neutron stars? (iii) black holes?

The exact mass cutoffs are not known. Sources can be found within ±25% of the thresholds shown below.

- $M \lesssim 0.08\ M_\odot \implies$ brown dwarf
- $0.08 \lesssim M \lesssim 0.8\ M_\odot \implies$ He white dwarf (theoretical)
- $0.8 \lesssim M \lesssim 8\ M_\odot \implies$ CO white dwarf ($\lesssim 1.4\ M_\odot$)
- $8 \lesssim M \lesssim 20\ M_\odot \implies$ neutron star (1.4–3 $M_\odot$)
- $M \gtrsim 20\ M_\odot \implies$ black hole ($\gtrsim 3\ M_\odot$)

What causes the variance/uncertainty in the exact cutoffs between these regimes? Two main factors:

1. **Metallicity** variations, and the abundances of different elements. This can cause a difference in the mass loss rates of stars over their lifetimes.
2. **Multiplicity**, i.e. stars in binary, triple, or higher-order systems. These can exchange mass through stellar winds and/or Roche lobe overflow.

The significance of these limits is explained below[22,11]:

- The core of anything below $\lesssim 0.08\ M_\odot$ will never get hot enough to fuse Hydrogen, so it will never become a true star.
- Very low mass stars' cores will never get hot enough to fuse Helium after their main sequence lifetime ends. Thus, it is thought that they will form pure **He white dwarfs** when they die, but their lifetimes are longer than the current age of the universe, so this has not been confirmed.
- In the next mass regime, the cores will get hot enough to fuse He into C and O, but will never get hot enough to fuse anything higher. Thus, they form **CO white dwarfs**. In this regime, mass loss from stellar winds during the AGB phase starts becoming important, and we get planetary nebulae. Combined with the fact that the remnant is composed of only the core of the original star that undergoes fusion, this results in a final mass much smaller than the initial stellar mass. These remnants cannot exist above a mass of $\sim 1.4\ M_\odot$ (the Chandrasekhar limit) since they are degenerate (electron degeneracy pressure), otherwise they will collapse on themselves.
- More massive stars will continue fusing higher and higher elements until they reach ^{56}Fe, which has the highest binding energy per nucleon. At this point they can't fuse anything else since it would result in a net loss of energy. As a result, when their cores collapse they experience extreme contraction that causes neutron degeneracy pressure to kick in, forming **neutron stars**. Very massive stars like this experience significant mass loss due to winds throughout their entire lifetime, not just the AGB phase. Like white dwarfs, neutron stars are much less massive than the original star.
- Even more massive stars' cores will not be able to resist gravitational collapse, even by neutron degeneracy pressure, so they form **black holes**.

For details on stellar evolution, see §3.Q22.

41. What is a neutron star? What assumptions and inputs go into determining the upper mass limit for a neutron star? What is the approximate ratio of neutrons to protons (and electrons) in the interior of a neutron star?

A **neutron star** is a type of stellar remnant that is produced in the end stages of stellar evolution for a massive star ($8 \lesssim M \lesssim 20\ M_\odot$; see §4.Q40). It is the extremely dense core of the star that has been left behind after its outer layers are blown away in the AGB and supernova phases (see §3.Q22). It is composed almost entirely of neutrons and is held up against its own gravity by **neutron degeneracy pressure**. This pressure arises because neutrons (like electrons and protons) are fermions, so by the **Pauli exclusion principle**, they cannot occupy the same quantum state as other neutrons in the same interacting system. As the core of a massive star collapses and the density increases, neutrons get packed together and all of the lowest energy states get filled up first. Thus, the only available energy states left are high energies, which produce an inherent pressure.

Typical properties of a neutron star

- $M \sim 1$–$3\ M_\odot$
- $R \sim 10$ km
- $B \sim 10^{12}$ G

The Interior of a Neutron Star[11]

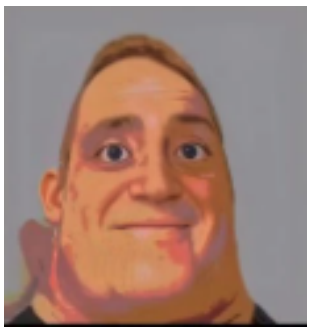
The neutron star consists of a few layers of increasing density as one gets closer to the center.

1. **Outer Crust** ($\rho \sim 10^6 - 4 \times 10^{11}\ \mathrm{g\,cm^{-3}}$)

   - $\rho \sim 10^6\ \mathrm{g\,cm^{-3}}$: Heavy nuclei, mostly in the form ^{56}Fe, in a solid lattice or fluid "ocean", along with nonrelativistic free electrons.
   - $\rho \sim 10^9\ \mathrm{g\,cm^{-3}}$: Electrons become relativistic.
   - $\rho \sim 4 \times 10^{11}\ \mathrm{g\,cm^{-3}}$: **Neutron drip**—some neutrons start being found outside of nuclei

2. **Inner Crust** ($\rho \sim 4 \times 10^{11} - 2 \times 10^{14}\ \mathrm{g\,cm^{-3}}$)

   - $\rho \sim 4 \times 10^{12}\ \mathrm{g\,cm^{-3}}$: Free neutrons start pairing up with each other, creating bosons which are not subject to Pauli exclusion. They form a **superfluid** that flows without resistance. This neutron superfluid exists alongside relativistic free electrons and lattice nuclei.
   - $\rho \sim 4 \times 10^{12}\ \mathrm{g\,cm^{-3}}$: At the same time, neutron degeneracy pressure starts to dominate.



3. **Interior** ($\rho \sim 2\times10^{14} - 4\times10^{14}\ \mathrm{g\,cm^{-3}}$)

   - $\rho \sim 2\times10^{14}\ \mathrm{g\,cm^{-3}}$: The density reaches the density of the nucleus, and atoms dissolve into neutrons. We now have a superfluid of free neutrons alongside superfluid, superconducting protons and relativistic free electrons.

4. **Core** ($\rho \sim 4\times10^{14} - 10^{15}\ \mathrm{g\,cm^{-3}}$)

   - $\rho \sim 4\times10^{14}\ \mathrm{g\,cm^{-3}}$: There may or may not be a solid core where pions and other sub-nuclear particles can be created through neutron decays:

$$\mathrm{n} \longrightarrow \mathrm{p}^+ + \pi^- \tag{4.8}$$

Degeneracy Pressure

What does neutron degeneracy pressure look like? We can get an approximate equation of state for degeneracy pressure in a Fermi gas by assuming some probability distribution $f(p)$, which gives the number of particles per unit volume per unit momentum. In degenerate conditions, there should be some cutoff momentum p_F below which *all* states are filled, and above which *all* states are empty:

$$f(p) = \begin{cases} 1 & p \leqslant p_F \\ 0 & p > p_F \end{cases} \tag{4.9}$$

This is, of course, a simplification, but it illustrates the effects well. p_F is called the **Fermi momentum**.

The volume of a unit cell in phase space is $g\mathrm{d}^3\mathbf{p}/h^3$, where in our case $g = 2$ for 2 spin states. Therefore, the number density of particles with a specific momentum is $n(p)\mathrm{d}^3\mathbf{p} = 2f(p)\mathrm{d}^3\mathbf{p}/h^3$, which we can integrate over $\mathrm{d}^3\mathbf{p} = 4\pi p^2 \mathrm{d}p$ to obtain the total number density

$$n = \frac{2}{h^3}\int_0^\infty 4\pi p^2 f(p)\mathrm{d}p = \frac{8\pi}{h^3}\int_0^{p_F} \mathrm{d}p\, p^2 = \frac{8\pi}{3h^3}p_F^3 \quad\longrightarrow\quad p_F = \left(\frac{3nh^3}{8\pi}\right)^{1/3} \tag{4.10}$$

Each particle, when hitting a wall, imparts a change in momentum of $\Delta p = 2p\cos\theta$, and the flux of particles onto the wall is $n(p)v\cos\theta = n(p)p\cos\theta/m$ (number per unit area per unit time). Therefore, the pressure is just the flux times the change in momentum per particle: $P(p,\theta) = 2n(p)p^2\cos^2\theta/m$. Notice this is a function of p and θ. First we need to average over the solid

angle, which gives

$$P(p) = \langle P(p,\theta)\rangle = \frac{1}{4\pi}\int \mathrm{d}\Omega \cos^2\theta 2n(p)\frac{p^2}{m} = \frac{1}{3}\frac{p^2}{m}n(p) \tag{4.11}$$

We then need to integrate over p to get the total pressure

$$P = \frac{1}{3}\frac{2}{h^3}\int_0^\infty f(p)\frac{p^2}{m}4\pi p^2 \mathrm{d}p = \frac{8\pi}{3h^3}\int_0^{p_F}\frac{p^4}{m}\mathrm{d}p \tag{4.12}$$

In the non-relativistic limit, this becomes

$$P_{\mathrm{NR}} = \frac{8\pi}{3mh^3}\frac{p_F^5}{5} \quad\longrightarrow\quad \boxed{P_{\mathrm{NR}} = K_{\mathrm{NR}}\rho^{5/3}} \tag{4.13}$$

where K_{NR} is a proportionality constant:

$$K_{\mathrm{NR}} = \frac{1}{20}\left(\frac{3}{\pi}\right)^{2/3}\frac{h^2}{m(\mu m_p)^{5/3}} \tag{4.14}$$

In the ultra-relativistic limit, out replacement of $v = p/m$ that we did in deriving P no longer holds (instead we have $v = p/\gamma m$), so our denominator in the integral becomes $\gamma m = E/c^2 = \sqrt{(pc)^2 + (mc^2)^2}/c^2$. If we assume $pc \gg mc^2$ then we can simplify to

$$P_{\mathrm{UR}} = \frac{8\pi}{3h^3}\int_0^{p_F} cp^3\mathrm{d}p = \frac{2\pi c}{3h^3}p_F^4 \quad\longrightarrow\quad \boxed{P_{\mathrm{UR}} = K_{\mathrm{UR}}\rho^{4/3}} \tag{4.15}$$

Note the constant K_{UR} here is different than in the non-relativistic case[24]:

$$K_{\mathrm{UR}} = \frac{1}{8}\left(\frac{3}{\pi}\right)^{1/3}\frac{hc}{(\mu m_p)^{4/3}} \tag{4.16}$$

Once again, I want to emphasize that these equations of state are pretty gross simplifications for neutron degeneracy pressure, but they work pretty well for *electron* degeneracy pressure (important for white dwarfs). This is because we derived them using Fermi-Dirac statistics, which assumes the particles are non-interacting. For an electron gas inside a white dwarf, this is not a bad assumption, but for a neutron star where the particles are so close that they interact via the strong nuclear force, this assumption breaks down.

A rough approximation for the equation of state of a neutron star that people have used in the literature in the past is

$$P = K_{\mathrm{NS}}\rho^2 \tag{4.17}$$

where the adiabatic index $\gamma \sim 2$, and K_{NS} is another constant.

Mass limits

The effects of degeneracy pressure also require a modification of the equation of hydrostatic equilibrium:

$$\boxed{\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{G}{r^2}\left(\frac{M(r) + 4\pi r^3 P(r)/c^2}{1 - 2GM(r)/rc^2}\right)\left(\rho(r) + \frac{P(r)}{c^2}\right)} \tag{4.18}$$

This is the **Tolman-Oppenheimer-Volkoff Equation**. It includes relativistic effects, and thus the mass $M(r)$ now also includes the rest masses of all the particles[22]. This can be used to estimate an upper mass limit for neutron stars, called the **Tolman-Oppenheimer-Volkoff limit**. We can get a simplified version of this by treating the neutron star as an incompressible fluid with constant density ρ. Integrating the TOV equation for a constant ρ gives a pressure of

$$P(r) = \rho c^2 \frac{\sqrt{1-(R_S r^2/R^3)} - \sqrt{1-(R_S/R)}}{3\sqrt{1-(R_S/R)} - \sqrt{1-(R_S r^2/R^3)}} \tag{4.19}$$

where $R_S = 2GM/c^2$ is the Schwarzschild radius. In order for the pressure to remain finite, the denominator cannot blow up, which gives us the condition that

$$R > \frac{9}{8} R_S \tag{4.20}$$

We can convert this to an upper limit on the mass

$$\boxed{M < \frac{4}{9\sqrt{3\pi}} \frac{c^3}{\sqrt{\rho G^3}}} \tag{4.21}$$

For a neutron star with a constant density of $\rho = 10^{15}\ \mathrm{g\,cm^{-3}}$, this leads to an upper mass limit of $M \lesssim 4\ M_\odot$. The original calculations by TOV were more sophisticated and used the equation of state obtained above, which yields a limit of $0.7\ M_\odot$, but this comes with the same assumptions and caveats of this equation of state (a Fermi gas of non-interacting particles)[51,52]. They also assumed no rotation.

More sophisticated modern estimates place the limit higher, at around $M_{\max} \sim 2.2 - 2.9\ M_\odot$. These estimates include the effects of the strong nuclear force and the rotation rate of a neutron star, which both increase the mass limit (faster rotation $\longrightarrow$ higher centripetal force $\longrightarrow$ can be more massive). However, these limits are still highly uncertain because the equation of state for neutron stars is still not well understood. The most massive observed neutron star has $M \sim 2.1\ M_\odot$ (PSR J0740+6620, discovered in 2019)[53].

Approximate ratio of neutrons to protons (and electrons)

This will be dictated by the relative reaction rates for β decays, which turn neutrons into protons, and electron captures, which turn protons into neutrons:

$$\begin{aligned} &\mathrm{n} \longrightarrow \mathrm{p}^+ + \mathrm{e}^- + \bar{\nu}_e \quad (\beta \text{ decay}) \\ &\mathrm{p}^+ + \mathrm{e}^- \longrightarrow \mathrm{n} + \nu_e \quad (\text{electron capture}) \end{aligned}$$

Because of neutron and electron degeneracy, β decays will only be able to produce electrons and protons close to their respective Fermi energies $E_{p,F}$ and $E_{e,F}$. On the other hand, in order for electron captures to happen, the electron must be energetic enough to account for the difference in masses between the proton and neutron (the neutron is slightly heavier). Thus, the electron must have a kinetic energy of at least $m_n c^2 - m_p c^2 - m_e c^2 = 0.78\,\mathrm{MeV}$. These are extremely high energies, so only the most energetic electrons and protons will be used up in electron captures. We notice then that both of these processes will be exchanging particles around their Fermi energies. We will then reach an equilibrium between these two reactions: if too many protons and electrons are produced by β decays, their Fermi energies will climb up until they become energetic enough to recombine through

electron captures. However, if too few protons and electrons are produced, their Fermi energies will be small and leave holes at the top of their distributions that can be easily filled by more β decays[2].

When we reach an equilibrium state between these two processes, the Fermi energies must satisfy

$$\boxed{E_{n,F} = E_{p,F} + E_{e,F}} \tag{4.22}$$

Since $E_F = \sqrt{(p_F c^2) + (mc^2)^2}$, we can plug in the Fermi momentum, which depends on the number densities of each particle, and solve for n_n/n_p. At typical neutron star densities, $\boldsymbol{n_n/n_p \sim 100}$. At even higher densities near the core, this will drop to $\boldsymbol{n_n/n_p \sim 8}$.

[2] See this StackExchange post

42. Why can we not have monopole and dipole terms in gravitational radiation (within Einstein's GR)?

Let's use electromagnetic radiation as an analogy. **Radiation**, by definition, involves energy transported to infinity. As energy radiates outwards spherically from a point source, it covers an area that increases like $4\pi r^2$, so the flux of radiation must fall off like $F(r) \propto 1/r^2$ (or more slowly) to carry any energy to arbitrary distances. If it falls off any faster than this, then there can be no energy carried to infinity since $L(r) = 4\pi r^2 F(r)$ will go to 0 for large r.

In the case of electromagnetism, energy is carried by the **Poynting vector** $\mathbf{S} = (c/4\pi)\mathbf{E} \times \mathbf{B}$. But we know that electrostatic fields go like $E \propto 1/r^2$, and magnetostatic fields go like $B \propto 1/r^2$. In this case $S \propto 1/r^4$ falls off *faster* than $1/r^2$, so there is **no radiation** from static fields. In contrast, accelerating charges produce electric and magnetic fields that have terms $\propto 1/r$, so $S \propto 1/r^2$ and they **do** have radiation. Thus, charges must be able to move to produce radiation.

We can generalize this concept of "accelerating charges produce radiation" by looking at different **moments** of some electric charge density $\rho_e(\mathbf{r})$.

1. **Electric Monopole**: $\int d^3\mathbf{r}\, \rho_e(\mathbf{r})$ — This is just the total charge Q, which is conserved. Since this can't vary over time, we cannot have any radiation.
2. **Electric Dipole**: $\int d^3\mathbf{r}\, \rho_e(\mathbf{r})\, \mathbf{r}$ — There is no conservation law relating to the electric dipole moment, so this can vary, and we can have electric dipole radiation.
3. **Magnetic Dipole**: $\int d^3\mathbf{r}\, \rho_e(\mathbf{r})\, \mathbf{r} \times \mathbf{v}(\mathbf{r})$ — Again, there is no conservation law involving the magnetic dipole moment, so we can have magnetic dipole radiation too.

Now let's compare these directly with the corresponding moments for some mass density $\rho_m(\mathbf{r})$.

1. **Monopole**: $\int d^3\mathbf{r}\, \rho_m(\mathbf{r})$ — This is just the total mass M, which, like Q, is conserved. Therefore, we can't have monopolar gravitational radiation.
2. **Dipole 1**: $\int d^3\mathbf{r}\, \rho_m(\mathbf{r})\, \mathbf{r}$ — This, you may recognize, is the center of mass of the system. If we look in the center of mass frame, then, this moment does not change (due to conservation of momentum), and there is no dipolar radiation. The existence of radiation is frame-independent, so we cannot have dipolar radiation in any frame.
3. **Dipole 2**: $\int d^3\mathbf{r}\, \rho_m(\mathbf{r})\, \mathbf{r} \times \mathbf{v}(\mathbf{r})$ — This is the total angular momentum of the system, which is conserved. Thus, we can't have this form of dipole radiation either.
4. **Quadrupole**: $\int d^3\mathbf{r}\, \rho_m(\mathbf{r})\, r_i r_j$ — This is the moment of inertia tensor, which is not conserved, so we can have quadrupolar gravitational radiation.

Note that the underlying reason for the difference in the dipole moments between EM and gravitational radiation is that **gravity is purely attractive**, or put another way, **there is no negative mass**. In the case of point particles, our gravitational dipole moment is $\sum_i m_i \mathbf{r}_i$, which has a derivative $\sum_i \mathbf{p}_i$ and second derivative $\sum_i \dot{\mathbf{p}}_i$. We can see immediately from the conservation of momentum that this must be 0.

Thus, our conclusions are that gravitational waves **cannot** be produced by any spherically symmetric variations, or by any axisymmetric rotations[54].

43. Give a qualitative description of the frequency and amplitude evolution of the gravitational wave signal from a binary compact star system. Compute actual numbers for a 30–30 $M_\odot$ binary black hole system.

We start by assuming gravitational waves are a small perturbation in locally flat spacetime, such that the metric can be written

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \tag{4.23}$$

where $\eta_{\mu\nu}$ is the Mike Wazowski metric and $h_{\mu\nu}$ is a small perturbation. With this metric, the Einstein equations become a wave equation for $h_{\mu\nu}$

$$\Box^2 h_{\mu\nu} = 0 \ , \ \Box^2 \equiv \nabla^2 - \frac{\partial^2}{\partial t^2} \tag{4.24}$$

Thus, we can assume an ansatz solution of the form

$$h_{\mu\nu} = A_{\mu\nu} \exp(ik_\alpha x^\alpha) \tag{4.25}$$

where $k^\alpha = (\omega, \mathbf{k})$ is the wave (4-)vector and $A^{\mu\nu}$ is a tensor with constant components. We have gauge freedom here (i.e. the freedom to choose our coordinate system), which we can use up to enforce that the coordinates do not oscillate with the waves (which is obviously not a physical effect since coordinates are not physical objects). This leaves the waving of spacetime as a physical effect. This is called the **transverse-traceless (TT) gauge**, and it makes the amplitude tensor look like

$$A_{\mu\nu} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & A_{xx} & A_{xy} & 0 \\ 0 & A_{xy} & -A_{xx} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{4.26}$$

The components can be interpreted as the different *polarizations* of the gravitational wave. The A_{xx} component corresponds to "+" polarization, and the A_{xy} component corresponds to "×" polarization. Look at Figure 4.5 and imagine a wave propagating in the z direction (into the page). The bottom left panel shows the "+" polarization where the wave compresses along the x direction while stretching along the y direction, and vice versa. The bottom right panel shows the "×" polarization where the wave stretches and compresses along the diagonals.

Now we're going to make an analogy with EM. Remember that the vector potential A^μ may be written as an integral over the current density J^μ (i.e. the *source*)

$$A^\mu(t, \mathbf{r}) = \frac{1}{c^2} \int \mathrm{d}^3\mathbf{r}' \, \frac{\mathbf{J}(t - |\mathbf{r} - \mathbf{r}'|/c, \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \tag{4.27}$$

We can do the same thing for $h_{\mu\nu}$, where the source is the stress-energy tensor $T_{\mu\nu}$

$$h_{\mu\nu}(t, \mathbf{r}) = \frac{4G}{c^4} \int \mathrm{d}^3\mathbf{r}' \, \frac{T_{\mu\nu}(t - |\mathbf{r} - \mathbf{r}'|/c, \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \tag{4.28}$$

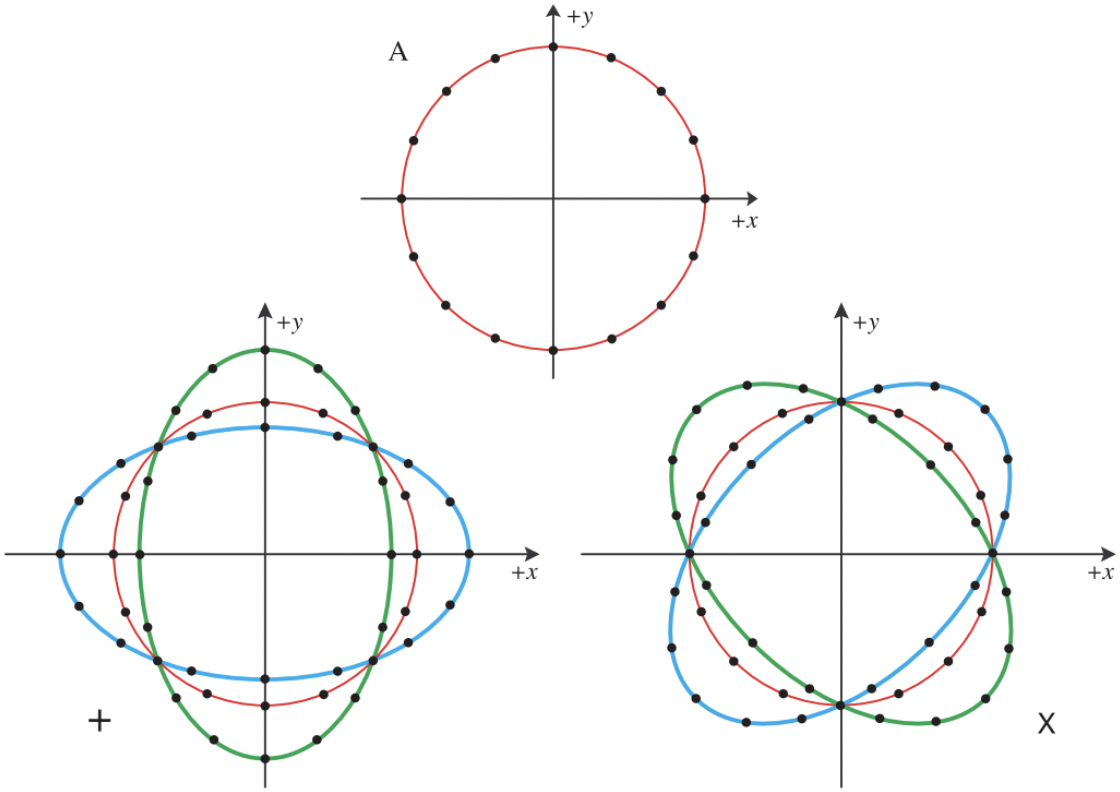


Figure 4.5: The different types of gravitational wave polarizations.

In the non-relativistic, weak-field limit, this simplifies to the **quadrupole formula**

$$h_{jk} = \frac{2G}{c^4}\frac{1}{r}\ddot{\mathscr{I}}_{jk}(t - r/c) \tag{4.29}$$

Here, $\mathscr{I}_{jk}$ is the traceless moment of intertia tensor

$$\mathscr{I}^{jk} = I^{jk} - \frac{1}{3}\delta^{jk}\delta_{\ell m}I^{\ell m} \quad \longrightarrow \quad I^{jk} = \int \mathrm{d}^3\mathbf{r}\,\rho(t,\mathbf{r})r^j r^k \tag{4.30}$$

Take a moment to stare at (4.29) for a second. Notice the similarities and differences to the **Larmor formula** for EM radiation. $\ddot{\mathscr{I}}_{jk}$ is just like the "acceleration" of the mass distribution that produces the gravitational radiation, and the factor of $1/r$ is the same as the scaling of the electromagnetic fields for dipole radiation. Since the power is proportional to $4\pi r^2 h^2 \sim$ const, energy is conserved. From this, we can find the "luminosity" (power) to be

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{G}{c^5}\frac{1}{5}\langle \dddot{\mathscr{I}}_{jk}\dddot{\mathscr{I}}^{jk}\rangle \tag{4.31}$$

If we assume a binary system with circular orbits, we have

$$\mathscr{I}^{jk} = \frac{1}{2}\mu a^2 \begin{pmatrix} \cos(2\omega t) + 1/3 & \sin(2\omega t) & 0 \\ \sin(2\omega t) & -\cos(2\omega t) + 1/3 & 0 \\ 0 & 0 & -2/3 \end{pmatrix} \tag{4.32}$$

Taking 3 derivatives

$$\dddot{\mathscr{I}}^{jk} = 4\mu a^2\omega^3 \begin{pmatrix} \sin(2\omega t) & -\cos(2\omega t) & 0 \\ -\cos(2\omega t) & -\sin(2\omega t) & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{4.33}$$

And then we sum and average

$$\dddot{\mathscr{I}}_{jk}\dddot{\mathscr{I}}^{jk} = 32\mu^2 a^4\omega^6(\sin^2(2\omega t) + \cos^2(2\omega t)) \tag{4.34}$$

$$\langle \dddot{\mathscr{I}}_{jk}\dddot{\mathscr{I}}^{jk} \rangle = 32\mu^2 a^4\omega^6 \tag{4.35}$$

So our power is

$$\boxed{L = \frac{32G}{5c^5}\mu^2 a^4\omega^6} \tag{4.36}$$

We know from orbital mechanics that the energy of an orbit (§1.Q2) changes by

$$E = \frac{GM\mu}{2a} \quad \longrightarrow \quad \frac{\mathrm{d}E}{\mathrm{d}t} - = \frac{GM\mu}{2a^2}\dot{a} \tag{4.37}$$

And, using Kepler's 3rd law, with $P = 1/f = 2\pi/\omega$:

$$P^2 = \frac{4\pi^2}{GM}a^3 \quad \longrightarrow \quad a = \left(\frac{GM}{\omega^2}\right)^{1/3} \quad , \quad \dot{a} = -\frac{2}{3}\left(\frac{GM}{\omega^5}\right)^{1/3}\dot{\omega} \tag{4.38}$$

Equating the two energy derivatives, we have

$$\frac{32G}{5c^5}\mu^2\left(\frac{GM}{\omega^2}\right)^{4/3}\omega^6 = \frac{GM\mu}{3}(GM\omega)^{-1/3}\omega^{11/3} \tag{4.39}$$

$$\dot{\omega} = \frac{96}{5c^5}G^{5/3}M^{2/3}\mu\omega^{11/3} \tag{4.40}$$

We now define the **chirp mass**

$$\mathscr{M} \equiv \mu^{3/5}M^{2/5} = \frac{(M_1M_2)^{3/5}}{(M_1+M_2)^{1/5}} \tag{4.41}$$

Such that

$$\dot{\omega} = \frac{96}{5}\left(\frac{G\mathscr{M}}{c^3}\right)^{5/3}\omega^{11/3} \tag{4.42}$$

Now, because gravitational waves are quadrupolar (§4.Q42), the wave frequency ν will be twice the frequency of the orbit (i.e. the wave has two maxima and two minima for each orbit), so we make the replacement $\omega = \omega_{gw}/2 = \pi\nu$ and find

$$\boxed{\dot{\nu} = \frac{96}{5}\pi^{8/3}\left(\frac{G\mathscr{M}}{c^3}\right)^{5/3}\nu^{11/3}} \tag{4.43}$$

We can then integrate this over time to find an expression for the frequency

$$\boxed{\nu(t) = \frac{1}{\pi}\left(\frac{5}{256}\right)^{3/8}\left(\frac{G\mathscr{M}}{c^3}\right)^{-5/8}(t_c - t)^{-3/8}} \tag{4.44}$$

where t_c is the time at which the binary coalesces. Notice that $\nu \propto \Delta t^{-3/8}$, so as the time gets closer to coalescence, Δt gets smaller and ν gets larger, until it diverges at the time of coalescence. This is called a **chirp waveform**, and it looks something like Figure 4.6.

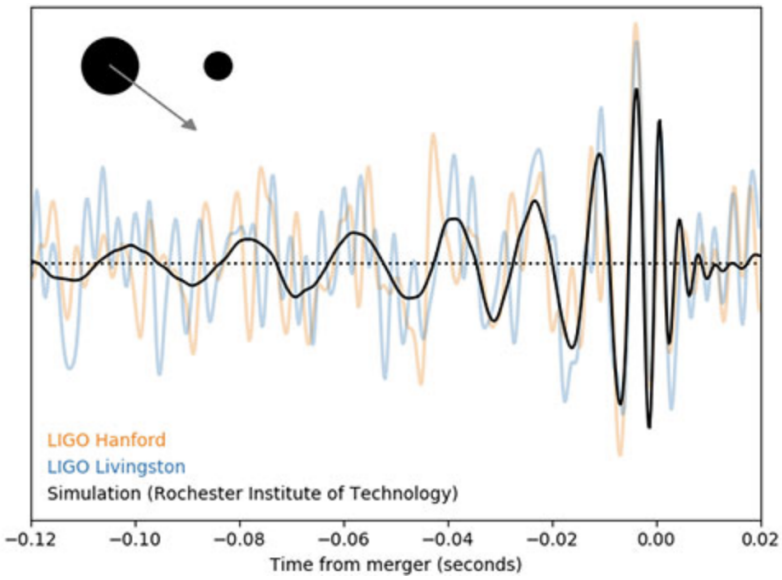


Figure 4.6: A typical gravitational wave signal from a merger event. First there is a low-amplitude, low-frequency signal. This increases gradually in both amplitude and frequency during the inspiral, until it formally diverges at the merger/coalescense. Then it settles back down during the ring-down.

The amplitude h_0 can be obtained by using equation (4.29) and averaging over the polarizations:

$$h_0 = \sqrt{\langle h_+^2 \rangle + \langle h_\times^2 \rangle} \tag{4.45}$$

But we can also use dimensional analysis to argue that the observed flux must be

$$F \propto \frac{h_0^2 \nu^2 c^3}{G} \quad , \quad [F] = \left[\frac{\mathrm{s}^{-2}\,\mathrm{cm}^3\,\mathrm{s}^{-3}}{\mathrm{g}^{-1}\,\mathrm{cm}^3\,\mathrm{s}^{-2}}\right] = [\mathrm{g\,s}^{-3}] = [\mathrm{erg\,s}^{-1}\,\mathrm{cm}^{-2}] \tag{4.46}$$

And we can replace the flux with a luminosity ($F = L/4\pi r^2$), using (4.36), to find

$$L = \frac{32G^{7/3}}{5c^5}(\mathscr{M}\pi\nu)^{10/3} \quad \longrightarrow \quad \frac{1}{4\pi r^2}\frac{32G^{7/3}}{5c^5}(\mathscr{M}\pi\nu)^{10/3} \propto \frac{h_0^2\nu^2c^3}{G} \tag{4.47}$$

Then, we can rearrange for h_0. We see that it's proportional to $\nu^{2/3}$, $\mathscr{M}^{5/3}$, and r^{-1}. So, more massive, closer, and more rapidly spinning binaries produce larger gravitational waves. The actual constants turn out such that the full expression is

$$\boxed{h_0 = \frac{4}{r}\left(\frac{G\mathscr{M}}{c^3}\right)^{5/3} c(\pi\nu)^{2/3}} \tag{4.48}$$

See the rough behavior of ν and h over time plotted in Figure 4.7. We can also plug in equation (4.44) to find the proportionality with time, i.e. $h_0 \propto (t_c - t)^{-2/8} \propto (t_c - t)^{-1/4}$.

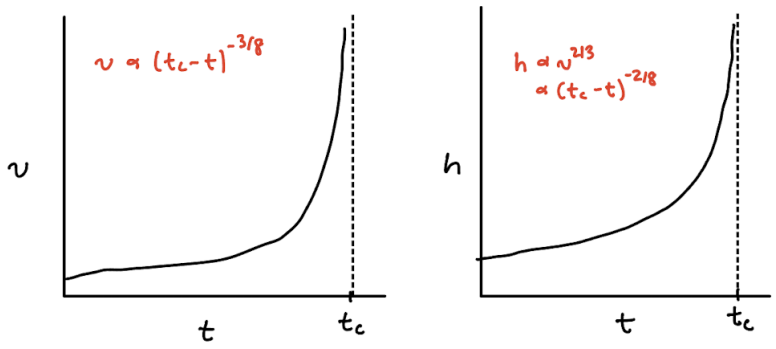


Figure 4.7: The frequency (left) and amplitude (right) of a gravitational wave from a binary system as it inspirals.

The behavior of the waveform can then be qualitatively modeled by an expression such as

$$h(t) = h_0(t)\cos\left(2\pi\nu t + \pi\dot{\nu}t^2 + \phi_0\right) \tag{4.49}$$

where $h_0(t)$ captures the evolution of the amplitude over time, as in (4.48).

For a 30 $M_\odot$–30 $M_\odot$ Binary

The chirp mass in this case is

$$\mathscr{M} = \frac{(M_1 M_2)^{3/5}}{(M_1 + M_2)^{1/5}} = \frac{30^{6/5}}{60^{1/5}} = \frac{1}{2^{1/5}} 30 \approx \boxed{26\,M_\odot} \tag{4.50}$$

Now what is the typical frequency near the chirp? Of course, the frequency diverges at the chirp itself, but before the chirp, during the period between the inspiral and the coalescence, we can estimate the "intermediate" frequency by calculating the frequency when the two black holes are at each other's ISCO. As we find in §4.Q46, the *orbital* frequency at the ISCO (for non-rotating black holes) is

$$\omega_{\rm ISCO} = \frac{c^3}{6^{3/2} G M_{\rm tot}} \tag{4.51}$$

So the gravitational wave frequency will be

$$\nu_{\rm gw} = \frac{\omega_{\rm ISCO}}{\pi} = \frac{c^3}{6^{3/2}\pi G M_{\rm tot}} \simeq \boxed{73\,\text{Hz}} \tag{4.52}$$

For any part of the inspiral after this, Newtonian gravity breaks down and we have to go to numerical GR to really get any accurate results. Using this chirp mass and frequency, we can calculate the luminosity at this point to be

$$L = \frac{32 G^{7/3}}{5c^5}(\mathscr{M}\pi\nu)^{10/3} \simeq \boxed{2\times 10^{55}\,\text{erg}\,\text{s}^{-1} \simeq 10^{21}\,L_\odot} \tag{4.53}$$

Notice that the frequency scales with $\nu \propto M^{-1}$—more massive binaries will have smaller coalescence frequencies, and the amplitude scales with $h_0 \propto r^{-1}$—closer binaries are easier to detect. But due to this dependence on distance, we cannot calculate h_0 without knowing r. Recall that gravitational

waves compress spacetime—that is, they change the proper distance between two particles by some difference ΔL. Then the amplitude, also called the **strain**, is typically on the order of

$$h_0 = \frac{\Delta L}{L} \sim \frac{\text{Diameter of H atom}}{1 \text{ AU}} \sim 10^{-21} \tag{4.54}$$

This means that for gravitational wave detectors with arm lengths on the order $L \sim 4$ km, we would have to measure a distance $\Delta L \sim 0.04$ fm[55].

Do gravitational waves really carry energy?[56]

Short answer: yes. We've already seen above the equations for the energy carried by a gravitational wave. But can we justify this physically? The first person known to do so was Richard Feynman at a conference in North Carolina in 1957. His argument goes something like this:

- Imagine a rigid rod with some rings looped around it that are free to slide along the rod.
- As a gravitational wave passes by, the intermolecular forces in the rod itself prevent it from being significantly stretched or squeezed by the waves, but the rings around the rod are free to slide along it.
- As the rings slide along the rod, they create friction and heat up the rod.
- This friction and heat has some energy associated with it. This energy must have come from somewhere, so we conclude that it must have come from the waves.

44. A double neutron star system in M31 is about to merge. What is the approximate energy emitted in gravitational radiation and what is the corresponding amplitude (strain) h observed here on Earth? Would LIGO be able to detect it?

Let's make some basic assumptions about our neutron star system:

- $\mu \sim 0.7\ M_\odot$, $M \sim 2.8\ M_\odot$, $\longrightarrow$ $\mathscr{M} = \mu^{3/5} M^{2/5} \sim 1.2\, M_\odot$
- $a \sim 6GM/c^2 \sim 24$ km
- $\omega_{\text{ISCO}} \sim c^3/6^{3/2}GM \sim 4900\,\text{rad}\,\text{s}^{-1}$ (from Kepler's 3rd law; *orbital* frequency)
- $\nu \sim \omega/\pi \sim 1.6$ kHz (*wave* frequency)
- $r = d_L \sim 750$ kpc (Luminosity distance to Andromeda)

As we found in §4.Q43, the power emitted from a gravitational wave is

$$L = \frac{32G}{5c^5}\mu^2 a^4 \omega^6 = \frac{32G^{7/3}}{5c^5}(\mathscr{M}\pi\nu)^{10/3} \tag{4.55}$$

In this equation, ω is the *orbital* frequency and ν is the *wave* frequency. Plugging everything in, we get

$$\boxed{L \sim 2 \times 10^{55}\,\text{erg}\,\text{s}^{-1} \sim 10^{21}\, L_\odot} \tag{4.56}$$

For a merger that lasts ~ 10 ms, the total energy released is

$$\boxed{E \sim 10^{53}\,\text{erg}} \tag{4.57}$$

This is about the same level as a supernova. Note that this is just approximate, to be rigorous we should really integrate L over time, using the expressions for $\omega(t)$ from §4.Q43.

Similarly, from §4.Q43, the strain is

$$h_0 = \frac{4}{r}\left(\frac{G\mathscr{M}}{c^3}\right)^{5/3} c\omega^{2/3} \tag{4.58}$$

Plugging in numbers, we obtain

$$\boxed{h_0 = \frac{\Delta L}{L} \sim 10^{-20}} \tag{4.59}$$

Would LIGO be able to detect it?

Yes, this is well within the detection limits of LIGO at 1.6 kHz (see Figure 4.8).

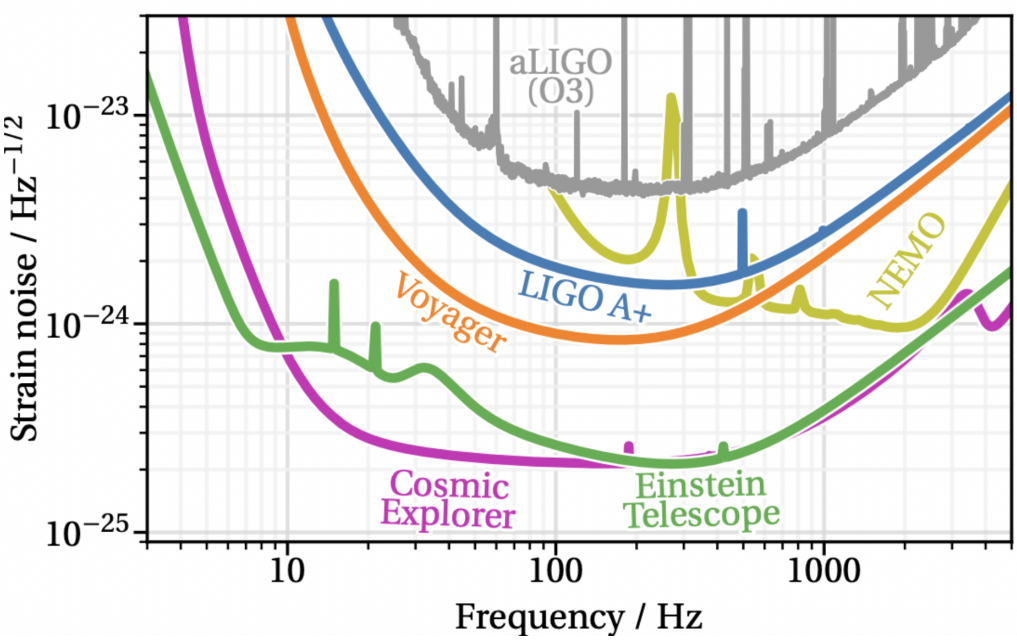


Figure 4.8: The sensitivity limits of LIGO and a few other gravitational wave instruments[57]. The rise at the low-frequency end is due to seismic noise (from the Earth), whereas the rise at the high-frequency end is due to photon Poisson noise and quantum noise (i.e. Heisenberg uncertainty in the amplitude/phase of the light). There is also a relatively ∼flat noise level attributed to thermal noise from the random motions of the electrons in the mirrors of each arm. The spikes are caused by power lines (60 Hz and harmonics thereof), suspension mechanical resonances, and intentionally added excitations for calibration & alignment purposes. Note the odd units of $\mathrm{Hz}^{-1/2}$ for h—this is because the luminosity must have the proportionality $L \propto \int h^2 \mathrm{d}\nu$, and the Hz units must cancel out.

45. Besides the merger of two compact objects, what other sources of gravitational waves are expected (though maybe not yet detected)?

Gravitational waves can be produced by any accelerating mass with a quadrupole moment. Already detected sources include:

- Binary neutron star mergers
- Neutron star–black hole mergers
- Binary stellar mass black hole mergers

Compact Binary Coalescence (CBC): Theoretically, there are other binary systems that should emit gravitational waves, but have not yet been detected:

- White dwarf–white dwarf merger
- White dwarf–neutron star merger
- White dwarf–black hole merger
- SMBH binary
- Extreme Mass Ratio Inspirals (EMRI)

Sources that do *not* involve the merger of two compact objects are

- **Continuous**: A rotating neutron star with asymmetric features on its surface (i.e. "mountains" and "valleys")
  - These have not been detected yet because the signal is extremely weak. Instead of a chirp, this should have a relatively constant frequency.
- **Bursts**: Supernovae and long GRBs
  - If they are not spherically symmetric, they could produce gravitational waves.
  - It is not well known what the wave signal would look like from such a source, but it would likely have to be bursty. This unknown quality makes it hard to search for such signals.
- **Stochastic**: Primordial fluctuations
  - GWs may have been produced during inflation and the early phase transitions of the universe right after the Big Bang. These should be analogous to the CMB in that they should be uniform and **stochastic** across the entire sky (stochastic meaning a random pattern that may be analyzed statistically but cannot be predicted precisely).
  - These can be searched for using Pulsar Timing Arrays (PTAs), which use many millisecond pulsars across the sky to search for primordial GW signals. Gravitational waves produce tiny perturbations in the pulsar period, and these signals should be correlated across different pulsars in the sky, whereas noise processes (from other effects in individual pulsars) will not be correlated.

The **NANOGrav collaboration** in June 2023 reported the first preliminary detection of a GW background signal using PTAs[58]. The correlation signal of the 15-year pulsar timing dataset follows closely to the expected **Hellings-Downs** pattern for a stochastic GW background signal. This is sort of like the analog to plotting the CMB power spectrum, which gives the correlation of the CMB at different angular scales (§8.Q128 and §8.Q129).

It is not confirmed whether the GW background signal detected by NANOGrav is actually due to primordial GWs. The nanohertz frequencies that PTAs probe could also be due to a population of SMBH mergers.

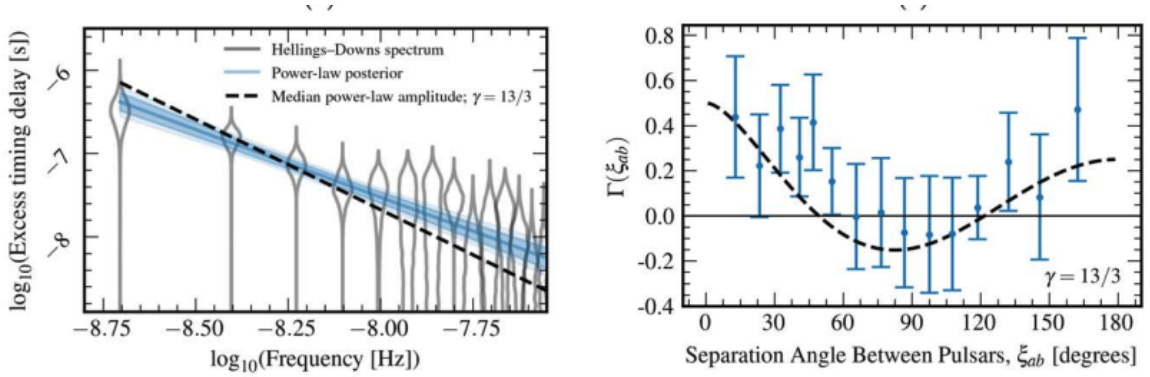


Figure 4.9: Left: The excess delay in pulsar timings as a function of frequency, showing a power-law relationship. For a pure SMBH background, the power law index should be 13/3 (dashed line). The posterior of the data is the blue line with an index closer to 3. Right: The correlation between pulsar signals as a function of the angular separation ξ_{ab}. The dashed line represents the pattern that would be expected if there were a GW background, while the flat solid line is the prediction for no GW background. The blue points are what has been measured by the NANOGrav collaboration[59,58].

46. How does the orbital frequency of the innermost stable circular orbit around a black hole scale with its (i) mass? (ii) spin?

For a non-rotating black hole, the Schwarzschild metric can be used to derive equations of motion, which produce an effective potential of the form

$$V_{\text{eff}}(r) = \left(1 - \frac{R_S}{r}\right)\left(c^2 + \frac{\ell^2}{r^2}\right) \tag{4.60}$$

where R_S is the Schwarzschild radius and ℓ is the angular momentum of the orbit. This produces a purely non-Newtonian effect where, under certain conditions, V_{eff} does not have a minimum. If we minimize V_{eff}, we get the condition

$$\frac{\mathrm{d}V_{\text{eff}}}{\mathrm{d}r} = 0 = R_S r^2 c^2 - 2\ell^2 r + 3R_S \ell^2 \tag{4.61}$$

This is a quadratic equation for $r = r_{\text{min}}$, which we can solve to obtain

$$r_{\text{min}} = \frac{\ell^2 \pm \sqrt{\ell^4 - 3R_S^2 c^2 \ell^2}}{R_S c^2} \tag{4.62}$$

We notice now that if the expression under the radical is less than 0, we get a complex result, meaning there is no minimum. Thus, if the expression is *exactly* 0, we get the boundary between the stable and unstable region. This happens when the angular momentum $\ell = \ell_{\text{min}}$ satisfies

$$\ell_{\text{min}}^2 = 3R_S^2 c^2 \tag{4.63}$$

At the same time, solving (4.61) for ℓ gives a relationship between ℓ and r

$$\ell^2 = \frac{R_S c^2 r^2}{2r - 3R_S} \tag{4.64}$$

which we can then use with (4.63) to obtain another quadratic equation, this time for $r_{\text{min}}(\ell = \ell_{\text{min}}) = r_{\text{ISCO}}$:

$$c^2 r_{\text{ISCO}}^2 - 6R_S c^2 r_{\text{ISCO}} + 9R_S^2 c^2 = 0 \tag{4.65}$$

Solving for r_{ISCO} reveals only one solution:

$$r_{\text{ISCO}} = 3R_S = \frac{6GM}{c^2} \tag{4.66}$$

We see that r_{ISCO} scales linearly with the mass[22]. This scaling generalizes to the spinning case, but the scaling with the spin parameter is more complicated. Let's just say that a spinning black hole produces an ISCO with a constant factor α such that $\alpha = 6$ in the non-spinning case, i.e.

$$\boxed{r_{\text{ISCO}} = \alpha \frac{GM}{c^2}} \tag{4.67}$$

And I'll just quote the solution for α in the spinning case ($\mp$ for prograde/retrograde orbits)[47]

$$\alpha = 3 + Z_2 \mp \left[(3 - Z_1)(3 + Z_1 + 2Z_2)\right]^{1/2} \tag{4.68}$$

Here, Z_1 and Z_2 are parameters that depend on a, the spin parameter:

$$Z_1 = 1 + (1-a^2)^{1/3}[(1+a)^{1/3} + (1-a)^{1/3}] \tag{4.69}$$

$$Z_2 = (3a^2 + Z_1^2)^{1/2} \tag{4.70}$$

We can see the behavior in some limiting cases

$$\begin{aligned} a = -1 &\longrightarrow Z_1 = 1 \ , \ Z_2 = 2 \ , \ \alpha = 9 \\ a = 0 &\longrightarrow Z_1 = 3 \ , \ Z_2 = 3 \ , \ \alpha = 6 \\ a = +1 &\longrightarrow Z_1 = 1 \ , \ Z_2 = 2 \ , \ \alpha = 1 \end{aligned}$$

We see that if the black hole is spinning rapidly in the *prograde* direction ($a = +1$), the ISCO becomes much closer ($\alpha = 1$) than the non-spinning case, whereas if it's spinning rapidly in the *retrograde* direction ($a = -1$), the ISCO moves out further ($\alpha = 9$) than the non-spinning case. See Figure 4.10 for a plot of the ISCO radius as a function of a.

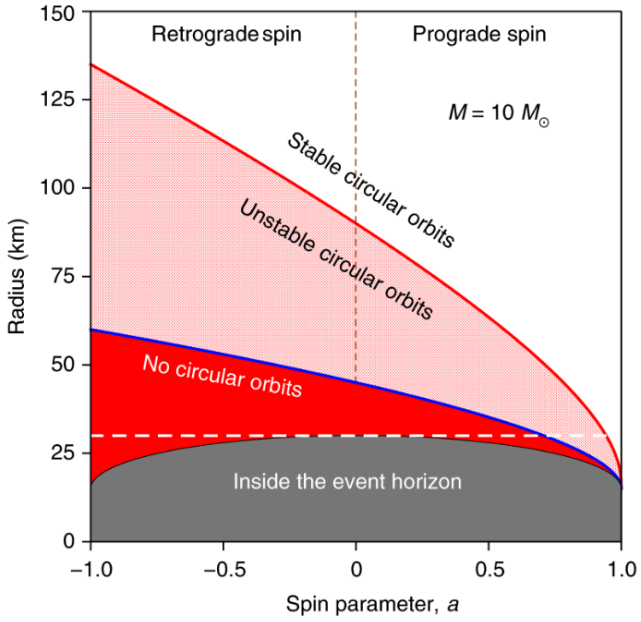


Figure 4.10: The ISCO radius as a function of the spin parameter a for a $10\,M_\odot$ black hole[47].

Converting this to an angular frequency using Kepler's 3rd law, we have

$$\omega_{\rm ISCO}^2 = \frac{GM}{r_{\rm ISCO}^3} \longrightarrow \boxed{\omega_{\rm ISCO} = \frac{c^3}{\alpha^{3/2} GM}} \tag{4.71}$$

Therefore, we have scalings of

- $\omega_{\rm ISCO} \propto M^{-1}$: More massive BHs produce larger ISCO radii and slower ISCO orbital frequencies
- $\omega_{\rm ISCO} \propto \alpha^{-3/2}$: More prograde spinning BHs produce smaller ISCO radii and faster ISCO orbital frequencies; more retrograde spinning BHs produce larger ISCO radii and slower ISCO orbital frequencies

47. What is the Eddington limit and how is it manifested in (i) ordinary stars? (ii) accreting X-ray sources?

The total pressure inside of a star is a combination of the gas pressure (given by the ideal gas law) and the radiation pressure:

$$P = \frac{\rho}{\mu m_p} kT + \frac{1}{3} aT^4 \tag{4.72}$$

We can see from the temperature scaling that if the temperature is hot enough and the density is low enough, then the radiation pressure can dominate over the gas pressure. If the radiation pressure is too high such that the star is driven out of hydrostatic equilibrium, mass loss can occur. This will happen at a threshold luminosity called the **Eddington luminosity**. We can solve for this by considering the pressure gradient due to radiation pressure from equation (3.74)

$$\frac{\mathrm{d}P_{\mathrm{rad}}}{\mathrm{d}r} = -\frac{\alpha}{c} F = -\frac{\kappa\rho}{c} \frac{L}{4\pi r^2} \tag{4.73}$$

and equating this to the pressure gradient for hydrostatic equilibrium

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{GM\rho}{r^2} \tag{4.74}$$

Solving for the luminosity, we obtain[11,22]

$$\boxed{L_{\mathrm{Edd}} = \frac{4\pi GMc}{\kappa} \simeq 3 \times 10^4 \left(\frac{M}{M_\odot}\right)\left(\frac{\kappa}{\kappa_\odot}\right)^{-1} L_\odot} \tag{4.75}$$

If we assume the opacity comes entirely from Thomson scattering, then $\kappa = \sigma_T / m_p$ and we have

$$\boxed{L_{\mathrm{Edd}} = \frac{4\pi GMm_p c}{\sigma_T}} \tag{4.76}$$

In normal stars

- As mentioned, if a star exceeds its Eddington limit, it experiences significant mass loss due to stellar winds.
- Since $L_{\mathrm{Edd}} \propto M$ and the main-sequence mass-luminosity relation is $L \propto M^{3.5}$, more massive stars' luminosities scale up faster than their Eddington luminosities, so we eventually reach a turnover point where $L_{\mathrm{Edd}} > L$.
- Stars with masses $\gtrsim 30\ M_\odot$ have such strong stellar winds that they lose mass and converge rapidly onto the track for a $30\ M_\odot$ star.
- Stars that have completely lost their outer layer of Hydrogen due to these winds and are fusing helium in their cores are called **Wolf-Rayet stars**.

In accreting X-ray sources

- In these objects, the source of luminosity is the release of gravitational energy as the gas loses angular momentum and falls onto the compact object. Thus, the luminosity is governed by the

accretion rate $\dot{M}$:

$$L = \frac{GM\dot{M}}{R} \tag{4.77}$$

Often this is parameterized in terms of η, the fraction of the rest mass energy that is released as gravitational energy during an infall:

$$L = \eta \dot{M} c^2 \tag{4.78}$$

Because of this, η is often called the accretion *efficiency*. Comparing these two equations, we see

$$\eta = \frac{GM}{Rc^2} \tag{4.79}$$

Which is usually on the order of 10%–15% for neutron stars and black holes. We have an Eddington limit for these sources, above which the radiation pressure exceeds the pull of gravity, just like stars.

- If the accretion disk is geometrically and optically thin, then typically $L/L_{\text{Edd}} \lesssim 0.5$
- If the accretion disk is geometrically and optically thick, then typically $L/L_{\text{Edd}} \gtrsim 0.5$
- The derivation of the Eddington limit assumes spherical symmetry, but usually these sources have accretion occurring in disks, which are more efficient. Therefore, it *is* possible to have super-Eddington luminosities $L > L_{\text{Edd}}$.

48. What is the Roche potential in a binary system? Describe carefully the assumptions that go into deriving it. Define the Roche limit.

(note in this question I say "star" often when I really just mean any mass)

The Roche Potential

Consider a binary system with masses (M_1, M_2) at positions ($\mathbf{r}_1$, $\mathbf{r}_2$) from the center of mass. We'll assume both masses have the same angular frequency Ω, so let's look in a reference frame that rotates with the binary. In this frame, the positions of M_1 and M_2 are fixed.

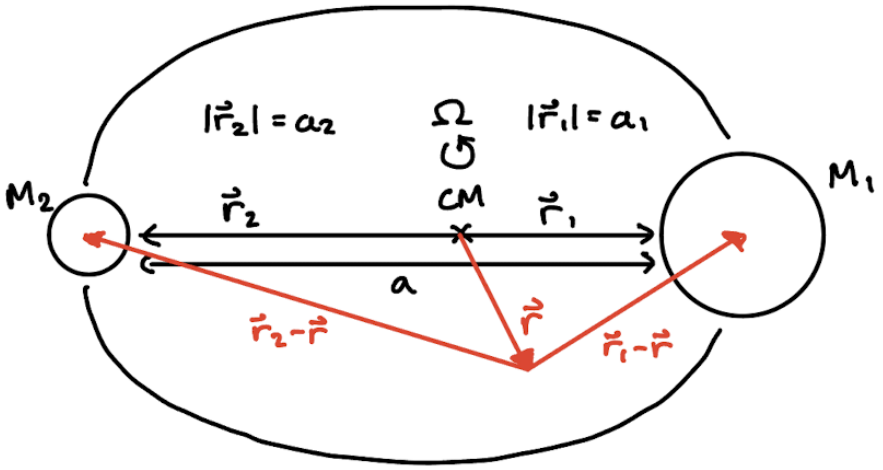


Figure 4.11: The setup of a rotating reference frame in a binary orbit to compute the potential at any arbitrary point $\mathbf{r}$.

Now we want to compute the potential at some arbitrary location $\mathbf{r}$ from the CM (see Figure 4.11). This will just be the sum of the potentials from each individual mass, plus a fictitious centrifugal term due to the rotation of the reference frame:

$$\boxed{\Phi(\mathbf{r}) = -\frac{GM_1}{|\mathbf{r}-\mathbf{r}_1|} - \frac{GM_2}{|\mathbf{r}-\mathbf{r}_2|} - \frac{1}{2}\Omega^2|\mathbf{r}|^2} \tag{4.80}$$

Note that this is the potential *per unit mass*. Say that the x-axis is aligned with the line that crosses M_1 and M_2, such that $\mathbf{r}_1 = (a_1, 0, 0)$ and $\mathbf{r}_2 = (-a_2, 0, 0)$. Then let's call our position vector $\mathbf{r} = (x, y, z)$. Then we can rewrite our potential as

$$\Phi(x,y,z) = -\frac{GM_1}{\sqrt{(x-a_1)^2+y^2+z^2}} - \frac{GM_2}{\sqrt{(x+a_2)^2+y^2+z^2}} - \frac{1}{2}\Omega^2(x^2+y^2+z^2) \tag{4.81}$$

Recall from equation (3.11) that putting our center of mass at the origin gives us the relationships

$$a_1 = \frac{M_2}{M}a \tag{4.82}$$

$$a_2 = \frac{M_1}{M}a \tag{4.83}$$

And from Kepler's 3rd law we have

$$\Omega^2 = \frac{G(M_1 + M_2)}{a^3} \tag{4.84}$$

Finally, our potential in cartesian coordinates is

$$\Phi(x,y,z) = -\frac{GM_1}{\sqrt{(x - M_2 a/M)^2 + y^2 + z^2}} - \frac{GM_2}{\sqrt{(x + M_1 a/M)^2 + y^2 + z^2}} - \frac{1}{2}\frac{GM}{a^3}(x^2 + y^2 + z^2) \tag{4.85}$$

What does this potential look like? Well the first term is at a minimum when $\mathbf{r} = \mathbf{r}_1$, and the second term is at a minimum when $\mathbf{r} = \mathbf{r}_2$, whereas the last term just decreases (gets more negative) as $|\mathbf{r}|$ increases. Therefore, we'll have two local wells around each mass superimposed on an overall paraboloid. See Figure 4.12.

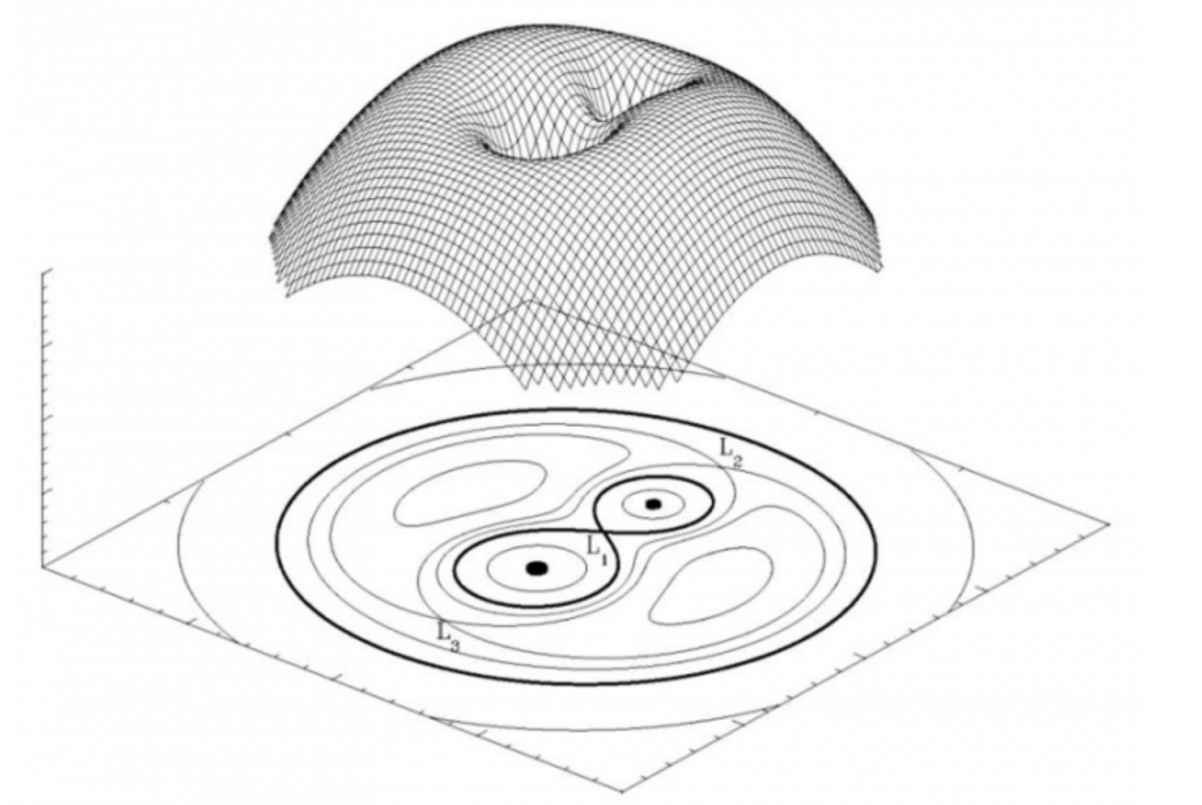


Figure 4.12: The Roche potential is visualized in two ways: first as a surface where the height depicts the potential, and second as a contour map, where each line represents an equipotential surface[60].

Recall that the force on a test mass m is $F = -m\nabla\Phi$, so points where the derivative of Φ vanishes will be points of (unstable) equilibrium. These are called **Lagrange points**, three of which are labeled in the Figure. There are also two (L_4 and L_5) which lie to either side. Since the two masses are so close together, their local potential wells partially overlap, and we get a Lagrange point (L_1) that coincides with the center of mass. This point also defines the intersection point of the "figure eight" shaped equipotential surface that surrounds both masses. This is the first such equipotential surface that is shared between both masses, and the sections of it that surround each mass are given a special name: the **Roche lobes**. Any test mass within a star's respective Roche lobe will be bound to that star.

The size of the Roche lobe around star 2 is given *very approximately* by

$$R_{L,2} \simeq \frac{a}{2}\left(\frac{M_2}{M}\right)^{1/3} \tag{4.86}$$

but remember, it is not spherical[22].

As a reminder, the assumptions we made in deriving the Roche potential, and thus the Roche lobes, are:

1. Stars can be treated as point masses
2. The orbits are synchronous ($\Omega = \Omega_1 = \Omega_2$) and circular
3. Newtonian gravity (no GR corrections)
4. The test mass is assumed to not be moving, otherwise it will experience a Coriolis force. We cannot include this in our model because the Coriolis force is not conservative, so it cannot be represented by a scalar potential.

Roche Lobe Overflow

In reality, stars are not point masses, and in fact if they are large enough (particularly in the red giant phase) they can **overflow their Roche lobes**, leading to mass from their outer layers being gravitationally attracted to their companion star and being accreted. There are a few different classes of binary systems defined by how their mass transfer evolves:

- **Detached binaries**: neither star fills its Roche lobe; no mass transfer
- **Semi-detached binaries**: one star fills its Roche lobe; mass is transferred to the other star
- **Contact binaries**: both stars fill their Roche lobes; mass can be transferred between both stars

The transfer of mass leads to a change in the orbital separation. If $\dot{a} < 0$, this is unstable since the masses will eventually collide. We can evaluate the sability of mass transfer by using the conservation of angular momentum (using the angular momentum from (1.9))

$$L = \mu\sqrt{GMa} \quad \longrightarrow \quad \frac{\mathrm{d}\ln L}{\mathrm{d}t} = \frac{\mathrm{d}\ln \mu}{\mathrm{d}t} + \frac{1}{2}\frac{\mathrm{d}\ln a}{\mathrm{d}t} = 0 \tag{4.87}$$

Now we look at the logarithmic derivative of the reduced mass (the total mass is conserved, so there is no $\mathrm{d}\ln M/\mathrm{d}t$ term). We'll also assume $\dot{M}_1 = -\dot{M}_2$, again due to mass conservation

$$\frac{\mathrm{d}\mu}{\mathrm{d}t} = \frac{1}{M}\frac{\mathrm{d}}{\mathrm{d}t}(M_1 M_2) = \frac{1}{M}(\dot{M}_1 M_2 + M_1 \dot{M}_2) = \dot{M}_2 \frac{M_1 - M_2}{M_1 + M_2} \tag{4.88}$$

$$\therefore \quad \frac{\mathrm{d}\ln \mu}{\mathrm{d}t} = \dot{M}_2 \frac{M_1 - M_2}{M_1 M_2} \tag{4.89}$$

Plugging back in and solving for the a derivative

$$\boxed{\frac{\dot{a}}{a} = -2\dot{M}_2 \frac{M_1 - M_2}{M_1 M_2}} \tag{4.90}$$

We assume $M_1 > M_2$, which means the sign of $\dot{a}$ is determined by the sign of $\dot{M}_2$. We have two possible situations here[22]:

1. $\dot{M}_2 > 0$ and $\dot{a} < 0$

   The secondary gains mass from the primary and the orbit shrinks. This is **unstable** since as the orbit shrinks mass transfer will continue until the two objects collide.

2. $\dot{M}_2 < 0$ and $\dot{a} > 0$

   The secondary loses mass to the primary and the orbit grows. This is **stable** since once the orbit expands enough the mass transfer will stop.

The Roche Limit[61]

This is conceptually related to the Roche potential, but not directly related in terms of the math. The **Roche limit** defines the distance from some massive object at which another approaching object will be completely ripped apart because the tidal forces are stronger than the object's own self-gravity.

We can derive the Roche limit by finding the distance where the tidal force equals the force of self-gravity. Call the primary mass M_1 and the secondary mass M_2. The radius of the secondary will be R, and the distance between the secondary and the primary will be d. We'll use a test mass m on the surface of M_2 to compare the tidal and gravitational forces. We use equation (1.33), evaluating at $\theta = 0$ (along the equator), to obtain the maximum tidal force:

$$F_g = F_{\text{tidal}}(0) \tag{4.91}$$

$$\frac{GM_2m}{R^2} = 2\frac{GM_1mR}{d^3} \tag{4.92}$$

This gives

$$\boxed{d_R = 2^{1/3}R\left(\frac{M_1}{M_2}\right)^{1/3}} \tag{4.93}$$

Side note: the maximum tidal force can be derived quickly (with a loss of generality from our derivation in §1.Q4) with

$$F_{\text{tidal}}(0) = \frac{GM_1m}{(d-R)^2} - \frac{GM_1m}{d^2} \tag{4.94}$$

$$\approx \frac{GM_1m}{d^2}\left(1 + 2\frac{R}{d}\right) - \frac{GM_1m}{d^2} \tag{4.95}$$

$$\approx \frac{2GM_1mR}{d^3} \tag{4.96}$$

However, this version of the Roche limit assumes that the secondary body is completely rigid. A more accurate expression that treats the secondary as a fluid (which is "easier" to deform and disrupt), derived by the man, the myth, the legend, Roche himself, has a different prefactor:

$$d_R \approx 2.44R\left(\frac{M_1}{M_2}\right)^{1/3} \tag{4.97}$$

After being tidally disrupted, the broken up material from the secondary gets spread out. Particles

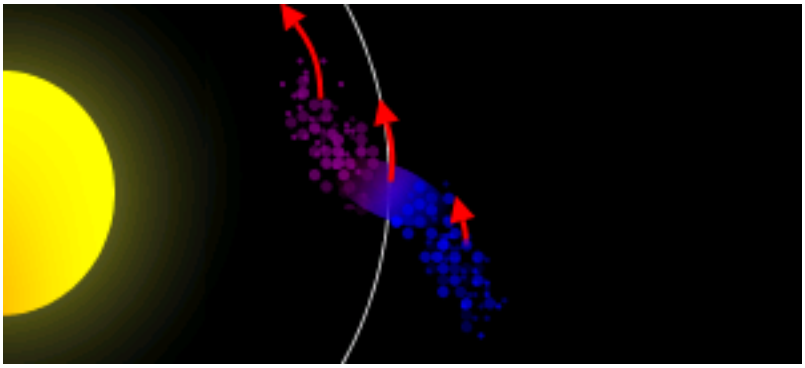


Figure 4.13: The material of a tidally disrupted secondary mass just after passing the Roche limit is now being spread out and starting to form a planetary ring.

that end up closer to the primary will orbit faster ($P^2 \propto a^3$), and particles further away will orbit slower (see Figure 4.13), so the material tends to get spread out. It's thought that this is how planetary **rings** form!

The Hill Sphere

While not asked, another related concept that may be worthwhile to know is the Hill sphere. Consider a system with a primary mass M_1, secondary mass M_2, and test mass m. The Hill sphere defines the region around the secondary mass for which the test mass will be gravitationally bound to it. Any orbit outside of the Hill sphere, then, is bound to the primary mass.

This can be derived by considering the balance of forces when the test particle lies exactly along the line connecting the primary and the secondary. There are gravitational forces from both bodies, as well as the centripetal force from the secondary's orbit around the primary. Let's say the distance between M_1 and M_2 is a, and the distance between M_2 and m is d_H. Then,

$$F_{g,1} + F_{g,2} + F_c = 0 \tag{4.98}$$

$$\frac{GmM_1}{(a-d_H)^2} - \frac{GmM_2}{d_H^2} - m\Omega^2(a-d_H) = 0 \tag{4.99}$$

Here, $\Omega = \sqrt{G(M_1+M_2)/a^3}$ is the Keplerian orbital velocity. We can then simplify by considering the limit $d_H \ll a$ and $M_2 \ll M_1$:

$$\frac{M_1}{(a-d_H)^2} - \frac{M_2}{d_H^2} - \frac{(M_1+M_2)(a-d_H)}{a^3} = 0 \tag{4.100}$$

$$\frac{M_1}{a^2}\left(1+2\frac{d_H}{a}\right) - \frac{M_1}{d_H^2} - \frac{M_1}{a^2}\left(1-\frac{d_H}{a}\right) \approx 0 \tag{4.101}$$

$$3\frac{M_1 d_H}{a^3} - \frac{M_2}{d_H^2} \approx 0 \tag{4.102}$$

Thus, the Hill sphere is approximately

$$\boxed{d_H \approx a\left(\frac{M_2}{3M_1}\right)^{1/3}} \tag{4.103}$$

49. What is the Shakura-Sunyaev (alpha-disk) model for accretion disks? What are the assumptions that go into its derivation?

What is an accretion disk?

We saw in the last question (§4.Q48) how mass transfer can be initiated between two objects in a binary configuration. This is just one scenario that can cause **accretion** of material onto a massive object (like a star or black hole) through a spinning disk called an **accretion disk**.

Why is matter accreted in a flat disk rather than a spherically symmetric distribution? Consider a system where we do start out with a spherically symmetric distribution of particles that has some net angular momentum. Near the equator, the centripetal force from the rotational motion will be large, and near the poles it will be small. All of the particles will experience a gravitational force towards the center of the sphere since this is the center of mass. Therefore, particles near the poles will fall inwards towards the center of mass much more easily than particles along the equator, causing the distribution to flatten out and form a disk. Particles with retrograde motions are more likely to collide, and over time they will be driven towards to prograde direction[1].

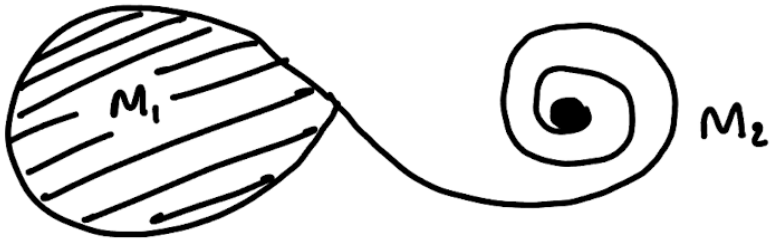


Figure 4.14: The inspiral of matter onto the secondary due to the Coriolis force during binary mass transfer.

In the case of accretion, particles falling inwards will gain angular momentum because the Coriolis force (which we neglected in §4.Q48) causes them to spiral inwards, as in Figure 4.14. Because this becomes a constant stream of gas rather than a single particle, it intersects itself and forms a ring. The ring is initially elliptical, but shocks cause energy loss while keeping the angular momentum constant, which circularizes the ring. The ring will also spread out in radius due to viscous coupling with gas in adjacent annuli. Rings at further radii have slower angular speeds, while rings at closer radii have faster angular speeds:

- Each ring tries to speed up its exterior neighbor
- Each ring tries to slow down its interior neighbor

The net effect is that the rings spread out and form an **accretion disk**. The innermost parts of the disk are collected by the massive object at the center, while the outermost parts of the disk move outwards to conserve angular momentum. As matter falls towards the center of the disk, viscous stresses cause it to radiate away some of its gravitational energy. However, this by itself is not enough to explain why angular momentum is transported to the outer parts of the disk. There must also be some mechanism that causes matter near the center of the disk to lose angular momentum—this is thought to be due to turbulent eddies in the disk, although the origin and microphysics of this turbulence are not well understood[22,29]. It is likely related to magnetic fields and magnetorotational instabilities.

The α-Disk Model

Developed by Shakura and Sunyaev in 1973[62], the so-called "α-disk" model describes the steady state solution for the accretion of matter onto a central compact object at a constant accretion rate $\dot{M}$ via a geometrically thin accretion disk. The unknown microphysics of the viscosity and turbulence are all absorbed into a single dimensionless parameter α. This model has a number of coupled equations that describe the macroscopic physics (as in many fluid flow problems). We'll work in cylindrical coordinates (r, φ, z). The geometry of the accretion disk is shown in Figure 4.15.

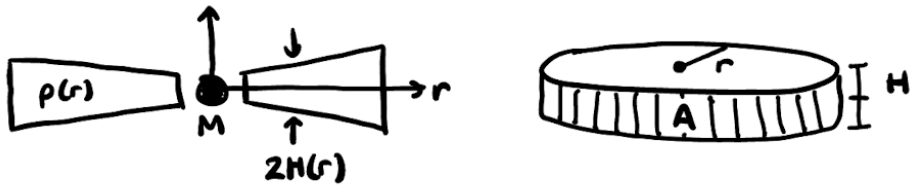


Figure 4.15: Left: A side-view of the accretion disk, such that we're seeing a cross-section normal to the plane of the disk. The central mass M, radial coordinate r, and disk half-height H are labeled. Notice that the hight of the disk increases with r. Right: A depiction of the cross-sectional area A for some radius.

1. **Conservation of mass**

   We have, from the continuity equation, for some mass density ρ and velocity $\mathbf{v}$

   $$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{4.104}$$

   Assuming the flow is purely in the radial direction, we can write the divergence as

   $$\nabla \cdot (\rho \mathbf{v}) = \frac{1}{r} \frac{\partial}{\partial r} (\rho r v_r) \tag{4.105}$$

   However, this assumes the disk is the same thickness for all r, which is not necessarily true. To allow for a changing thickness, we have to replace r with the cross-sectional area $A = 2\pi r \times 2H(r)$, where $H(r)$ is the half-height of the disk at a radius r. Also, since we're in the steady-state, we assume $\dot{\rho} = 0$, so we have

   $$\frac{1}{4\pi r H(r)} \frac{\mathrm{d}}{\mathrm{d}r} (4\pi r H(r) \rho v_r) = 0 \tag{4.106}$$

   We see that the quantity inside the derivative must be a constant. We can identify this quantity as the mass flow rate $\dot{M}$:

   $$|\dot{M}| = \left| \frac{\Delta M}{\Delta t} \right| = M \frac{\Delta r}{\Delta t} \frac{A}{V} = 4\pi r H(r) \rho v_r \tag{4.107}$$

   Therefore, we obtain our equation for conservation of mass

   $$\boxed{\dot{M} = -4\pi r H(r) \rho v_r} = \text{constant} \tag{4.108}$$

2. **Conservation of angular momentum**

   We can apply all of the same logic using the continuity equation with the density of angular momentum $\ell = \rho\Omega r^2$. However, this time we have a source term (the net torque per unit volume, $\tilde{\tau}_{\rm net}$)

$$\frac{1}{4\pi r H(r)}\frac{\mathrm{d}}{\mathrm{d}r}(4\pi r H(r)(\rho\Omega r^2)v_r) = \tilde{\tau}_{\rm net} \tag{4.109}$$

   Subsituting in $\dot{M}$ from the mass continuity equation

$$\boxed{\frac{1}{4\pi r H(r)}\frac{\mathrm{d}}{\mathrm{d}r}(-\dot{M}\Omega r^2) = \tilde{\tau}_{\rm net}} \tag{4.110}$$

3. **Conservation of vertical momentum (pressure balance)**

   We assume that hydrostatic equilibrium holds in the z-direction. The z-component of the gravitational acceleration is multiplied by a factor of $\sin\theta = z/r$, so we have

$$\frac{\mathrm{d}P}{\mathrm{d}z} = -\frac{GM\rho z}{r^3} \tag{4.111}$$

   which we can integrate to obtain

$$\boxed{P_0 = \frac{GM}{r^3}\int_0^H \rho z\,\mathrm{d}z \simeq \frac{GM\rho H(r)^2}{2r^3}} \tag{4.112}$$

4. **Conservation of energy**

   We assume that all of the viscous energy released locally within the disk is trasported vertically and emitted from the surface of the disk as blackbody radiation. Then, the power per unit surface area on the disk, $\dot{Q}$, is

$$\boxed{\dot{Q} = \sigma_{\rm SB}T_{\rm eff}^4} \tag{4.113}$$

5. **Radiative transport**

   Recall that the radiative flux can be written in the form (i.e. equation (3.74))

$$F = -\frac{c}{\rho\kappa_R}\frac{\mathrm{d}P_{\rm rad}}{\mathrm{d}z} \tag{4.114}$$

   We approximate

$$\frac{\mathrm{d}P_{\rm rad}}{\mathrm{d}z} \simeq \frac{P_{\rm rad}(H) - P_{\rm rad}(0)}{H} \simeq -\frac{P_{\rm rad}(0)}{H} \tag{4.115}$$

   Also recall the definitions

$$P_{\rm rad} = \frac{1}{3}aT^4 = \frac{4\sigma_{\rm SB}}{3c}T^4 \quad , \quad F = \sigma_{\rm SB}T^4 \tag{4.116}$$

Rearranging, our equation for the temperature at the center of the disk ($z = 0$) becomes

$$\boxed{T_0^4(r) = \frac{3}{4} H(r) \rho \kappa_R T_{\text{eff}}^4 = \frac{3}{4} \tau T_{\text{eff}}^4} \tag{4.117}$$

Where τ is the optical depth, which is $\gg 1$, so $T_0(r) \gg T_{\text{eff}}$.

6. **Torque from the viscous stress**

The $r\varphi$-component of the stress-energy tensor describes the force in the φ direction per unit area in the r direction. This is the viscous shear stress, which generates torque:

$$|T_{r\varphi}| = \rho \mathscr{V} r \left| \frac{\mathrm{d}\Omega}{\mathrm{d}r} \right| \tag{4.118}$$

Here, $\mathscr{V}$ is the **kinematic viscosity coefficient**, which is the product of some scale length λ and some scale velocity $\tilde{v}$: $\mathscr{V} \sim \lambda \tilde{v}$. For a turbulent eddy, the maximum possible size is the disk scale height $\lambda \sim H$, and the maximum possible velocity is the sound speed $\tilde{v} \sim c_s$. Therefore, if we include an order-unity constant α, we can write

$$\boxed{\mathscr{V} = \alpha c_s H} \tag{4.119}$$

This reduces the shear stress to $|T_{r\varphi}| \simeq \alpha P$. We can then write the torque (force × level arm) as

$$\tau(r) = -|T_{r\varphi}|(4\pi r H(r)) r = -4\pi \alpha P(r) H(r) r^2 \tag{4.120}$$

Which produces a net torque per unit volume

$$\tilde{\tau}_{\text{net}}(r) = \lim_{\Delta r \to 0} \frac{-4\pi\alpha P(r+\Delta r) H(r+\Delta r)(r+\Delta r)^2 + 4\pi\alpha P(r) H(r) r^2}{4\pi r H(r) \Delta r} \tag{4.121}$$

$$\boxed{\tilde{\tau}_{\text{net}}(r) = -\frac{\alpha}{r H(r)} \frac{\mathrm{d}}{\mathrm{d}r} \left(r^2 P(r) H(r) \right)} \tag{4.122}$$

7. **Power from the viscous stress**

For a force acting at a constant velocity, the power generated is $\mathbf{F} \cdot \mathbf{v}$. The angular equivalent of this is $\boldsymbol{\tau} \cdot \Delta\boldsymbol{\Omega}$ where $\Delta\Omega$ is for a small annulus in the disk. This becomes

$$\dot{Q} = \frac{\text{power}}{\text{area}} = \frac{|\tau(r)|}{2 \times 2\pi r \Delta r} \cdot \left| \frac{\mathrm{d}\Omega}{\mathrm{d}r} \right| \Delta r = \frac{4\pi\alpha P(r) H(r) r^2 \Delta r}{4\pi r \Delta r} \left| \frac{\mathrm{d}\Omega}{\mathrm{d}r} \right| \tag{4.123}$$

From Kepler's 3rd law, we have

$$\Omega = \sqrt{\frac{GM}{r^2}} \quad \longrightarrow \quad \frac{\mathrm{d}\Omega}{\mathrm{d}r} = -\frac{3\Omega}{2r} \tag{4.124}$$

Therefore

$$\boxed{\dot{Q} = \frac{3}{2}\alpha P(r)H(r)\Omega} \tag{4.125}$$

8. **Angular momentum part 2: Electric Boogaloo**

   Combining equations (4.110) and (4.122), we find

$$\frac{1}{4\pi r H(r)}\frac{\mathrm{d}}{\mathrm{d}r}(-\dot{M}\Omega r^2) = -\frac{\alpha}{rH(r)}\frac{\mathrm{d}}{\mathrm{d}r}(r^2 P(r)H(r)) \tag{4.126}$$

   Integrating,

$$\dot{M}\Omega r^2 = 4\pi\alpha r^2 P(r)H(r) + \text{constant} \tag{4.127}$$

   The constant can be found by noting that $\mathrm{d}\Omega/\mathrm{d}r = 0$ at the innermost part of the disk that meets the star. This means that $P(r_{\text{in}}) = 0$, so our constant must be

$$\text{constant} = \dot{M}\Omega(r_{\text{in}})r_{\text{in}}^2 \tag{4.128}$$

   Then, solving our equation for $P(r)H(r)$ gives

$$\boxed{P(r)H(r) = \frac{\dot{M}\Omega r^2 - \dot{M}\Omega_{\text{in}} r_{\text{in}}^2}{4\pi\alpha r^2} = \frac{\dot{M}\Omega}{4\pi\alpha}\left(1 - \sqrt{\frac{r_{\text{in}}}{r}}\right)} \tag{4.129}$$

9. **Effective temperature**

   Finally, we can combine equations (4.113), (4.125), and (4.129) to get an expression for the effective temperature

$$\dot{Q} = \sigma_{\text{SB}} T_{\text{eff}}^4 = \frac{3}{2}\alpha\Omega P(r)H(r) \tag{4.130}$$

$$= \frac{3\dot{M}\Omega^2}{8\pi}\left(1 - \sqrt{\frac{r_{\text{in}}}{r}}\right) \tag{4.131}$$

   Rearranging for T_{eff}:

$$\boxed{T_{\text{eff}}(r) = \left[\left(\frac{3GM\dot{M}}{8\pi\sigma_{\text{SB}} r^3}\right)\left(1 - \sqrt{\frac{r_{\text{in}}}{r}}\right)\right]^{1/4}} \tag{4.132}$$

   This equation determines the emission spectrum of the disk. Notice that T_{eff} is **independent of α**. From here, we would need to assume some equation of state and some opacity law to get a full profile for the pressure and density.

Let's review the assumptions we made in deriving all of these equations:

- Geometrically thin ($H \ll r$) / optically thick ($\tau \gg 1$)
- Azimuthal symmetry in the disk
- Subsonic turbulence ($v \ll c_s$)
- Rotational speed is much larger than radial speed ($v_\varphi \gg v_r$)

- Local thermodynamic equilibrium (LTE) / Disk radiates heat efficiently (i.e. blackbody)
- Steady-state (constant accretion rate $\dot{M}$)
- The disk mass $M_{\text{disk}} \ll M_{\text{star/BH/etc.}}$, allowing us to neglect the self-gravity of the disk

And let's summarize all of the important results[22,29]:

$$T_{\text{eff}}^4(r) = \left(\frac{3GM\dot{M}}{8\pi\sigma_{\text{SB}}r^3}\right)\left(1-\sqrt{\frac{r_{\text{in}}}{r}}\right)$$

$$T_0^4(r) = \frac{3}{4}H(r)\rho\kappa_R T_{\text{eff}}^4(r) = \frac{3}{4}\tau T_{\text{eff}}^4(r)$$

$$P(r)H(r) = \frac{\dot{M}}{4\pi\alpha}\sqrt{\frac{GM}{r^3}}\left(1-\sqrt{\frac{r_{\text{in}}}{r}}\right)$$

$$P_0(r) = \frac{GM\langle\rho\rangle H^2(r)}{2r^3} = \begin{cases} \dfrac{\rho_0(r)}{\mu m_p}kT_0(r) & \text{(ideal gas)} \\ \dfrac{4\sigma_{\text{SB}}}{3c}T_0^4(r) & \text{(radiation pressure)} \end{cases}$$

$$\kappa(r) = \begin{cases} \kappa_0\rho(r)T^{-3.5}(r) & \text{(Kramer)} \\ \sigma_T/m_p & \text{(Thomson scattering)} \end{cases}$$

50. Write down the fluid equations for conservation of mass and momentum that would describe a spherically symmetric, expanding supernova remnant. How would you derive the "jump conditions" for a strong adiabatic shock.

Conservation equations[22,29]

Mass conservation comes from the standard continuity equation

$$\boxed{\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0} \tag{4.133}$$

The **Jeans equation** (§8.Q130) describes the conservation of momentum (this is essentially just a continuity equation for $\rho\mathbf{v}$, with source terms from the force and pressure gradient)

$$\boxed{\rho \frac{\partial \mathbf{v}}{\partial t} + (\rho \mathbf{v} \cdot \nabla)\mathbf{v} = -\rho \nabla \Phi - \nabla P} \tag{4.134}$$

For energy, we can use another continuity equation, where this time our source terms will come from the first law of thermodynamics:

$$\mathrm{d}E = \mathrm{d}Q + \mathrm{d}W \tag{4.135}$$

where the work is

$$\frac{\mathrm{d}W}{\mathrm{d}t} = -\mathbf{F} \cdot \mathbf{v} = -\int_A P\mathbf{v} \cdot \mathrm{d}\mathbf{A} = -\int_V \nabla \cdot (P\mathbf{v}) \mathrm{d}V \tag{4.136}$$

Therefore, our energy conservation equation is

$$\frac{\partial \varepsilon}{\partial t} + \nabla \cdot (\varepsilon \mathbf{v}) = \dot{Q} - \nabla \cdot (P\mathbf{v}) \tag{4.137}$$

where

$$\varepsilon = \frac{1}{2}\rho v^2 + \frac{3}{2}nkT + n\Phi_g \tag{4.138}$$

But if we also assume the process is adiabatic, then $\dot{Q} = 0$, leaving us with

$$\boxed{\frac{\partial \varepsilon}{\partial t} + \nabla \cdot (\varepsilon \mathbf{v}) = -\nabla \cdot (P\mathbf{v})} \tag{4.139}$$

Shock from a supernova remnant[22,29]

Now let's apply these equations to an expanding supernova remnant. The enormous energy released forces the gas to supersonic speeds, creating a **shock front**. Behind the shock front, shocked gas moves at supersonic speeds v_s, and in front of the shock front, gas is not moving. However, it's helpful to analyze this problem in the frame of the shock front. Say that the shock front itself moves as a speed $u_s \gg c_s$. In the frame of the shock front then, the gas in front of it approaches at a speed $-u_s$, and the gas behind it moves at a speed $v = u_s - v_s$. See Figure 4.16.

There will be a jump in density at the shock front, where gas behind the shock front is much more

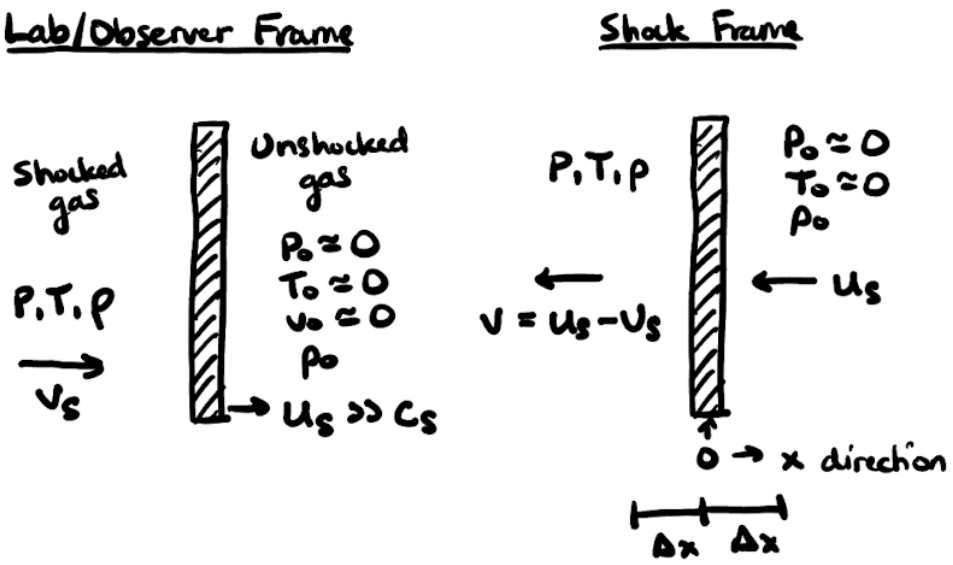


Figure 4.16: A shock front viewed in the observer frame and the shock frame.

dense than the gas in front of it, simply because the unshocked gas cannot respond quickly enough to change its density (see Figure 4.17).

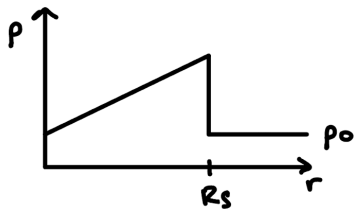


Figure 4.17: The density as a function of distance. R_s is the position of the shock front, where there is a large discontinuity between the shocked and unshocked gas densities.

Let's set our origin at the shock front and analyze this problem purely in a 1D sense along the x direction. We'll also assume we're in a steady-state, so all time derivatives go to 0. In this case, our mass continuity equation gives

$$\frac{\mathrm{d}}{\mathrm{d}x}(\rho v) = 0 \quad \longrightarrow \quad \rho v = \rho_0 u_s \tag{4.140}$$

And our momentum continuity equation, ignoring any gravity/external forces, gives

$$\rho v \frac{\mathrm{d}}{\mathrm{d}x}(v) = -\frac{\mathrm{d}P}{\mathrm{d}x} \tag{4.141}$$

We can pull the ρv inside the derivative, since we know from the mass continuity equation that it's just a constant.

$$\frac{\mathrm{d}}{\mathrm{d}x}(\rho v^2 + P) = 0 \quad \longrightarrow \quad P + \rho v^2 = \cancel{P_0}^{\,0} + \rho_0 u_s^2 \tag{4.142}$$

Finally, from the energy conservation equation we have

$$\frac{\mathrm{d}}{\mathrm{d}x}[(\varepsilon+P)v]=0 \;\longrightarrow\; \varepsilon v+Pv=\varepsilon_0 u_s+\cancelto{0}{P_0 u_s} \tag{4.143}$$

Again, we ignore gravity, so the energies become

$$\varepsilon=\frac{1}{2}\rho v^2+\frac{3}{2}P \;\;,\;\; \varepsilon_0=\frac{1}{2}\rho_0 u_s^2+\cancelto{0}{\frac{3}{2}P_0} \tag{4.144}$$

This gives

$$\frac{1}{2}\rho v^3+\frac{5}{2}Pv=\frac{1}{2}\rho_0 u_s^3 \tag{4.145}$$

We now have a system of equations (4.140), (4.142), and (4.145), which we can use to solve for v_s, ρ, and P in terms of u_s and ρ_0. First let's take (4.142), solve for P, and replace u_s with v using (4.140):

$$P=\rho_0 u_s^2-\rho v^2=\rho v^2\left(\frac{\rho}{\rho_0}-1\right) \tag{4.146}$$

Next, let's plug this into (4.145), while also replacing all of the u_s instances using (4.140)

$$\frac{1}{2}\rho\cancel{v^3}+\frac{5}{2}\rho\cancel{v^3}\left(\frac{\rho}{\rho_0}-1\right)=\frac{1}{2}\frac{\rho^3\cancel{v^3}}{\rho_0^2} \tag{4.147}$$

$$\frac{5}{2}\frac{\rho\cancel{\rho^1}}{\rho_0}-2\cancel{\rho}=\frac{1}{2}\frac{\rho\cancel{\rho^2}}{\rho_0^2} \tag{4.148}$$

$$\left(\frac{\rho}{\rho_0}\right)^2-5\frac{\rho}{\rho_0}+4=0 \tag{4.149}$$

$$\left(\frac{\rho}{\rho_0}-4\right)\left(\frac{\rho}{\rho_0}-1\right)=0 \tag{4.150}$$

We see we have two solutions for $\rho/\rho_0=1,4$. Obviously the case where this is 1 is trivial, since this means there is no boundary/shockwave at all. So let's ignore that and look at the $\rho/\rho_0=4$ case. We can see immediately from (4.140) that our velocities must satisfy

$$v=\frac{1}{4}u_s \;\longrightarrow\; v_s=u_s-v=\frac{3}{4}u_s \tag{4.151}$$

Finally, we can plug these back into our equation for the pressure to obtain

$$P=\rho v^2\left(\frac{\rho}{\rho_0}-1\right)=4\rho_0\frac{1}{4^2}u_s^2(4-1)=\frac{3}{4}\rho_0 u_s^2 \tag{4.152}$$

These are called the **Rankine-Hugoniot shock jump conditions**. In summary:

$$\boxed{v_s=\frac{3}{4}u_s} \qquad \boxed{\rho=4\rho_0} \qquad \boxed{P=\frac{3}{4}\rho_0 u_s^2} \tag{4.153}$$

This leads to a temperature of

$$T = \frac{P\mu m_p}{k\rho} = \frac{3}{16}\frac{\mu m_p}{k}u_s^2 \tag{4.154}$$

For a supernova shock, $u_s \sim 1000\,\mathrm{km\,s^{-1}}$, which leads to coronal temperatures:

$$T \simeq 2\times 10^7\left(\frac{u_s}{1000\,\mathrm{km\,s^{-1}}}\right)^2\ \mathrm{K} \tag{4.155}$$

51. Describe the various stages of evolution of a supernova remnant. What are the relevant physical processes during each phase? Explain why in the Sedov-Taylor phase of a supernova remnant, the radius expands at $t^{2/5}$.

A supernova remnant generally evolves in 3 phases, where the radius scales differently with time in each phase. This is summarized in Figure 4.18, and also in the numbered list below[63,29].

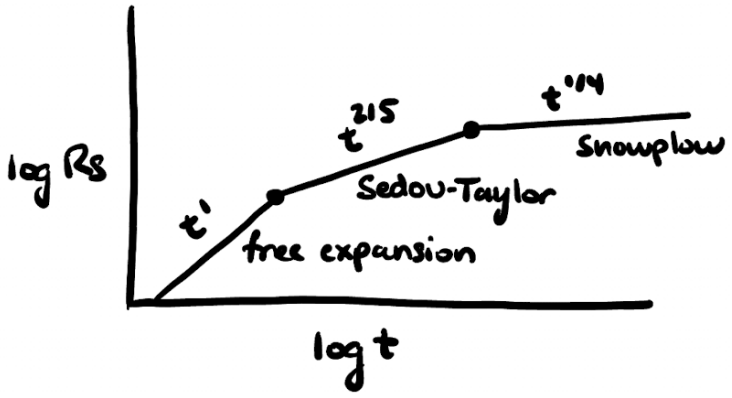


Figure 4.18: The different scaling sizes of the expanding supernova remnant over time, showing the three distinct phases of expansion: free expansion, the Sedov-Taylor phase, and the snowplow phase.

0. **Initial Explosion**

   During the supernova, an initial mass of $M_0 \sim 10\ M_\odot$ is ejected at velocities of $u_s(t=0) \equiv v_{\rm ej} \sim 3000\ {\rm km\,s^{-1}}$ into the surrounding medium, which has an ambient density ρ_0.

1. **Free expansion**

   Early on, the shock expands at a nearly constant velocity since the pressure from the ISM is negligible. As the shock expands outwards, it carries with it the ejecta mass M_0 from the supernova, which transfers momentum into the ISM gas and sweeps it up as well—$M_{\rm swept} = (4\pi/3)R_s^3\rho_0$. This means the ejecta mass must itself lose momentum, so it actually experiences a **reverse shock** that propagates inwards (in the rest frame of the ejecta). The ejecta itself is a site of **r-process nucleosynthesis** (see §12.8). The shock front continues expanding at a nearly constant speed until the swept up mass is approximately equal to the ejected mass, $M_{\rm swept} \sim M_0$. This usually lasts on the order of

   $$t \simeq \left(\frac{3M_0}{4\pi\rho_0 v_{\rm ej}^3}\right)^{1/3} \simeq 10^3\,{\rm yr} \tag{4.156}$$

   During this phase, the shock expands linearly

   $$R_s = v_{\rm ej} t \tag{4.157}$$

   With typical sizes of $R_s \sim 5$ pc.

2. **Sedov-Taylor phase (adiabatic phase)**

   Now that the swept-up mass dominates ($M_{\rm swept} \gg M_0$), the ISM pressure is no longer negligible and the shock velocity starts slowing down. At this point the reverse shock has also crossed most

of the ejecta material, so the ejecta gets decoupled from the shock front and it gets left behind as the shock continues to propagate outwards into the ISM. The shock is pushed forwards by the thermal pressure of the shock-heated material behind it. The gas, at this point, is still too hot to experience significant radiative energy losses (its cooling time is much larger than its dynamical time), so it expands adiabatically and energy is conserved. We have a combination of kinetic and thermal energy. Using the shock jump conditions from (4.153):

$$E_{\rm kin} = \frac{1}{2}Mv^2 = \frac{1}{2}\frac{4}{3}\pi R_s^3 \rho u_s^2 = \frac{2}{3}\pi R_s^3 \rho_0 u_s^2 \tag{4.158}$$

$$E_{\rm thm} = \frac{3}{2}PV = \frac{3}{2}\frac{3}{4}\rho_0 u_s^2 \frac{4}{3}\pi R_s^3 = \frac{3}{2}\pi R_s^3 \rho_0 u_s^2 \tag{4.159}$$

$$\therefore \quad E_0 = E_{\rm kin} + E_{\rm thm} = \frac{13}{6}\pi R_s^3 \rho_0 u_s^2 = \text{constant} \tag{4.160}$$

Since E_0 is constant, we can solve for u_s, the shock velocity, and see how it changes as the shock radius R_s increases

$$u_s = \frac{{\rm d}R_s}{{\rm d}t} = \left(\frac{6E_0}{13\pi\rho_0}\right)^{1/2} R_s^{-3/2} \tag{4.161}$$

This is a separable differential equation

$${\rm d}R_s R_s^{3/2} = \left(\frac{6E_0}{13\pi\rho_0}\right)^{1/2} {\rm d}t \quad \longrightarrow \quad R_s = \left(\frac{5}{2}\right)^{2/5}\left(\frac{6E_0}{13\pi\rho_0}\right)^{1/5} t^{2/5} \tag{4.162}$$

We see that $R_s \propto t^{2/5}$. This phase lasts until radiative losses start becoming important, which we can find by looking at the velocity and temperature dependence

$$u_s = \frac{{\rm d}R_s}{{\rm d}t} = \left(\frac{2}{5}\right)^{3/5}\left(\frac{6E_0}{13\pi\rho_0}\right)^{1/5} t^{-3/5} = \left(\frac{6E_0}{13\pi\rho_0}\right)^{1/2} R_s^{-3/2} \tag{4.163}$$

$$T_s = \frac{3}{16}\frac{\mu m_p}{k}u_s^2 = \frac{3}{16}\frac{\mu m_p}{k}\frac{6E_0}{13\pi\rho_0}R_s^{-3} \tag{4.164}$$

$$u_s \simeq 1.8\times 10^4 \left(\frac{E_{51}}{\rho_{01}}\right)^{1/2}\left(\frac{R_s}{1\,{\rm pc}}\right)^{-3/2} {\rm km\,s^{-1}} \tag{4.165}$$

$$T_s \simeq 7.3\times 10^9 \left(\frac{E_{51}}{\rho_{01}}\right)\left(\frac{R_s}{1\,{\rm pc}}\right)^{-3} {\rm K} \tag{4.166}$$

Radiative losses start becoming important at $T \sim 10^6$ K, so we get typical ages of $t \sim 4\times 10^4$ yr. At this point, typical values for the radius and velocity are $R_s \sim 20$ pc and $u_s \sim 200$ km s^{-1}.

3. **Snowplow phase (radiative phase)**

The cooling time of the shell of shocked gas is now fast and the shell starts rapidly losing energy. This happens because electrons and hydrogen ions (protons) recombine into neutral hydrogen, which dramaticaly decreases electron scatterings and causes the shocked gas to transition from optically thick to optically thin. The shell of gas now loses most of its internal thermal pressure and contracts into a thinner shell. However, the hot shocked gas *inside* the shell still has a lot of thermal pressure, which continues pushing the shell outwards. As the shell expands, the

internal pressure decreases adiabatically: $PV^\gamma = \text{constant}$. In other words

$$PV^\gamma \propto PR_s^{3\gamma} \propto PR_s^5 = \text{constant} \tag{4.167}$$

where we used $\gamma = 5/3$. Thus, we can write the pressure as

$$P(t) = P_0\left(\frac{R_0}{R_s}\right)^5 \tag{4.168}$$

This changes the momentum like

$$P = \frac{1}{A_s}\frac{\mathrm{d}p}{\mathrm{d}t} \quad \longrightarrow \quad \frac{\mathrm{d}}{\mathrm{d}t}(Mv) = 4\pi R_s^2(t)P(t) = 4\pi P_0 R_0^5 R_s^{-3}(t) \tag{4.169}$$

With $Mv \propto R_s^3\dot{R}_s$, we get a differential equation for R_s:

$$R_s^2{\dot{R}_s}^2 + R_s^3\ddot{R}_s \propto R_s^{-3} \tag{4.170}$$

Assuming a power law $R_s \propto t^\eta$, we find

$$t^{4\eta-2} \propto t^{-3\eta} \quad \longrightarrow \quad \eta = \frac{2}{7} \tag{4.171}$$

So, $R_s \propto t^{2/7}$ in the snowplow phase[64,65]. Note that, if we had ignored the internal pressure of the bubble and just considered the shell to be undergoing a perfectly inelastic collision with the surrounding environment (so momentum is conserved), we would obtain a slightly smaller scaling relation of $R_s \propto t^{1/4}$.

4. **Merger phase**

   Eventually, enough ISM mass is swept up that internal instabilities cause the expanding shell to break up and dissipate into the surrounding ISM, becoming indistinguishable.

52. What is a white dwarf star? Why is the radius of a white dwarf a decreasing function of its mass? What is the basic physics that leads to the upper limit on the mass of a white dwarf (i.e. the Chandrasekhar limit)?

A **white dwarf** is a type of stellar remnant that is produced in the end stages of stellar evolution for a low-mass star ($M \lesssim 8\ M_\odot$; see §4.Q40). It is the leftover Carbon-Oxygen core of the star after the outer layers have been blown away in the AGB and planetary nebula phases (see §3.Q22). It is extremely dense and is only held up against its own gravity by **electron degeneracy pressure**. This is pressure that exists purely due to quantum mechanical effects: electrons are fermions, which means they must obey the **Pauli exclusion principle**—i.e. they cannot occupy the same quantum state as any other electron in the same interacting system. This means that in the extremely dense conditions of the cores of stars, all of the lowest energy states get filled up, and the remaining electrons are forced to inhabit states with higher energies, which inherently have more pressure than the lower energy states.

Typical properties of a white dwarf

- $M \sim 0.1$–$1.3\ M_\odot$
- $R \sim 0.01\ R_\odot \sim R_\oplus$
- $\rho \sim 10^4$–$10^7\ \mathrm{g\,cm^{-3}} \ll \rho_{\text{neutron star}}$
- Thought to be the end state of every star with $M \lesssim 8\ M_\odot$

Depending on the initial mass of the star, the composition of the white dwarf will be different. CO white dwarfs are by far the most common:

Progenitor mass ($M_\odot$)	White dwarf mass ($M_\odot$)	White dwarf composition
0.5–1.5	0.1–0.4	Mostly He
1.5–8	0.5–1.3	Mostly C/O
6–8	1–1.3	Mostly O/Ne/Mg

Electron degeneracy pressure

Electron degeneracy pressure works exactly the same way as **neutron degeneracy pressure**, which is explained in §4.Q41. However, there is one key difference. Neutron degeneracy pressure only kicks in at *much higher* densities than electron degeneracy pressure, and at these high densities the neutrons will be interacting with each other via the strong nuclear force. In contrast, the electrons can be assumed to be essentially non-interacting, even in electron-degenerate conditions. This means that our derivation for the degeneracy pressure in §4.Q41 is actually a much better model for electron degeneracy pressure than it is for neutron degeneracy pressure. Without redoing the derivation, I'll just restate the results here for the non-relativistic (NR) and ultra-relativistic (UR) cases:

$$P_{\text{NR}} = \frac{1}{20}\left(\frac{3}{\pi}\right)^{2/3} \frac{h^2}{m_e(\mu_e m_p)^{5/3}} \rho^{5/3} \quad (P_{\text{NR}} \propto \rho^{5/3}) \tag{4.172}$$

$$P_{\text{UR}} = \frac{1}{8}\left(\frac{3}{\pi}\right)^{1/3} \frac{hc}{(\mu_e m_p)^{4/3}} \rho^{4/3} \quad (P_{\text{UR}} \propto \rho^{4/3}) \tag{4.173}$$

Mass-Radius Relationship

Degeneracy pressure will be the dominant pressure mechanism. So if we take the non-relativistic limit for the degeneracy pressure and compare this to the pressure that would be required to keep the white dwarf in hydrostatic equilibrium at the core (3.100), we immediately obtain a scaling relation[22,29,11]

$$P_{\text{HSE}}(0) \propto \langle P_{\text{NR}}(r) \rangle \tag{4.174}$$

$$\frac{M^2}{R^4} \propto \langle \rho \rangle^{5/3} \propto \left(\frac{M}{R^3}\right)^{5/3} \tag{4.175}$$

$$\boxed{R \propto M^{-1/3}} \tag{4.176}$$

Note that in the first step we are allowed to compare the *central* pressure for HSE to the *average* pressure for degeneracy because the central density is related to the average density by just a constant, which we're ignoring for this analysis.

This scaling relationship tells us that as we increase the mass, the radius must decrease. Intuitively, this happens because when we add mass, we require more pressure to counteract the increased gravitational force. But degeneracy pressure is proportional to the density, so we have to shrink in order to increase the density and increase the pressure. There is no other mechanism, such as nuclear fusion, that can provide pressure support without requiring an increase in density.

Chandrasekhar limit

The ultra-relativistic limit has a polytropic index of $4/3$, which if we recall from §3.Q31, implies that this limit is dynamically unstable. As a white dwarf gets more massive, it necessarily moves from the non-relativistic limit to the relativistic limit since electrons are forced to occupy more energetic states and increase their speeds to approach c. If we do the same analysis as above in the ultra-relativistic limit, we'll notice something stunning: the dependence on the radius will completely cancel out! This implies that there is some maximum mass, above which it cannot be supported by electron degeneracy pressure anymore. If we're more careful about not dropping our constants, we can try to derive what this upper mass limit is. We take the pressure as a function of radius from equation (3.59), using $r = 0$ for the central pressure and $\rho = 3M/4\pi R^3$ for the density

$$\frac{9}{24\pi}\frac{GM^2}{R^4} \simeq K_{\text{UR}}\left(\frac{3M}{4\pi R^3}\right)^{4/3} = \frac{1}{8}\left(\frac{3}{\pi}\right)^{1/3}\frac{hc}{(\mu_e m_p)^{4/3}}\left(\frac{3M}{4\pi R^3}\right)^{4/3} \tag{4.177}$$

After a lot of algebra that is not particularly illuminating, we achieve a mass limit of[11]

$$M_{\text{Ch}} \simeq \frac{3}{16\pi}\frac{1}{(\mu_e m_p)^2}\left(\frac{hc}{G}\right)^{3/2} \simeq 0.44\, M_\odot \tag{4.178}$$

But this is an underestimate of the true limit. The scalings with $\mu_e m_p$ and hc/G are all correct, but the prefactor of $3/16\pi$ is not. This is because we made the simplifying assumption that the density is always equal to the average density $\rho = 3M/4\pi R^3$. If one does a more careful analysis using the **Lane-Emden equation** (see §12.10.2) with a polytrope index of $n = 3$, we get a more precise limit of[66]

$$\boxed{M_{\text{Ch}} = 0.197\frac{1}{(\mu_e m_p)^2}\left(\frac{hc}{G}\right)^{3/2} \simeq 1.44\, M_\odot} \tag{4.179}$$

This is called the **Chandrasekhar limit**, named after Subrahmanyan Chandrasekhar, who discovered it at age 21. Don't ask me what I was doing at 21, it'll only be disappointing in comparison.

The mass-radius relationship in the non-relativistic and relativistic limits is shown in Figure 4.19.

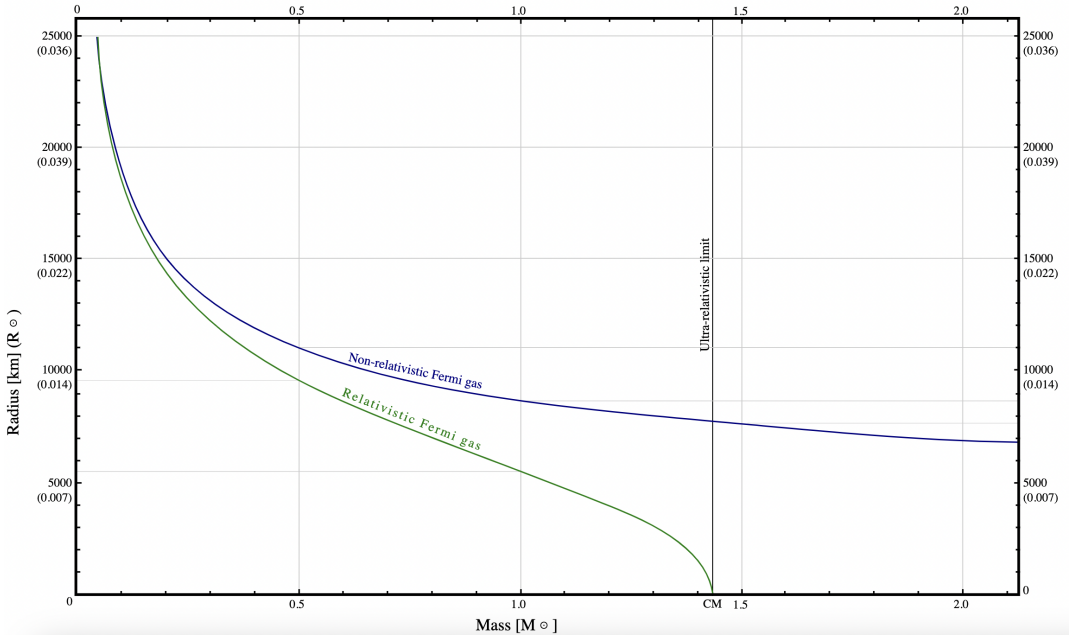
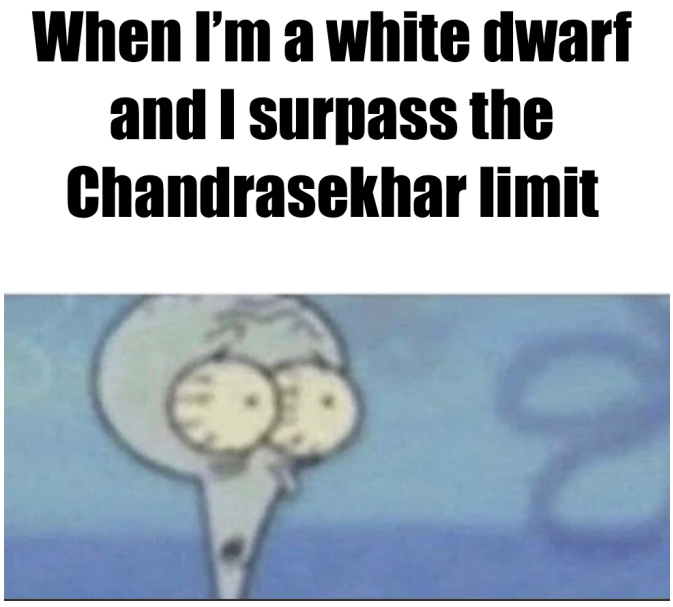


Figure 4.19: The mass-radius relationship for a degenerate Fermi gas in the non-relativistic and relativistic limits. Notice the vertical cutoff in the relativistic case at the Chandrasekhar limit of $\sim 1.4\ M_\odot$.



53. What is a "millisecond pulsar"? What is the shortest spin period known for such an object? Estimate a lower bound to its mean density.

First of all, a **pulsar** (pulsating radio source; PSR) is a highly magnetized ($B \sim 10^{12}$ G) neutron star that emits narrow beams of synchrotron radiation along the poles of its magnetic field, which is misaligned with its spin axis. This means that the beams of radiation sweep around the sky as it rotates. If they are aligned in such a way that they pass through our line of sight, we see pulses with a very consistent period.

A **millisecond pulsar** is a pulsar with a period of $\lesssim 10$ ms. Pulsars are expected to have periods on the order of milliseconds shortly after their birth, but they spin down over time since they lose energy to the synchrotron radiation. Most pulsars have periods of $\sim$ seconds. It is thought that most millisecond pulsars are not young systems, but rather they are pulsars in binary systems that have gained angular momentum through accretion, spinning them up. See Figure 4.20[29,22].

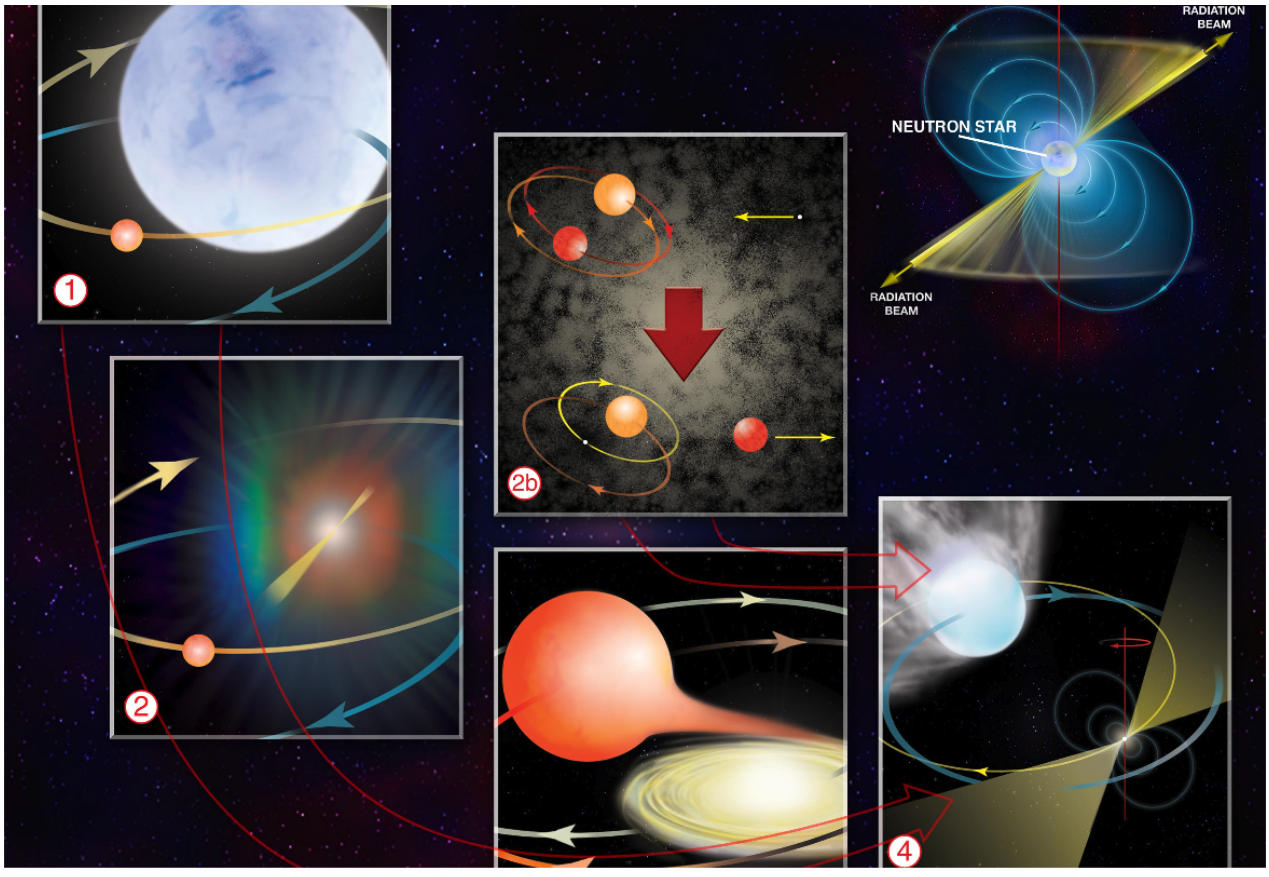


Figure 4.20: How a millisecond pulsar is created: (1) A main sequence star and a supergiant star are in a binary together. (2) The giant star goes supernova and leaves behind a pulsar. (3) A few billion years later, the other star evolves into the red giant phase and overflows its Roche lobe. The pulsar stars accreting material from its giant companion. (4) The pulsar gains angular momentum from the accretion and ends up spinning very rapidly. It also produces strong winds that erode the companion star. (2b) In an alternative scenario, globular clusters have so many stars packed together that a free neutron star could interact with a binary system and, in a process called "exchange", become a part of the binary while kicking out the lowest mass star. The process then continues with steps 3 and 4 in the normal scenario.

Pulsars in LMXBs get spun up through an accretion disk that builds up due to the Roche lobe overflow of its companion star. In this case, the companion star must have a long enough lifetime to still exist for the lifetime of a pulsar, so it must be a cool red star (unless the companion was captured later). In HMXBs, on the other hand, accretion can happen through strong stellar winds and/or a "decretion" disk that builds up around the star as it ejects materials.

Shortest spin period known

Currently the pulsar with the shortest known spin period is PSR J1748-2446ad in the Terzan 5 globular cluster. Its period is 1.4 ms (716 Hz)[67].

Lower bound to mean density

The pulsar must not spin too fast, lest it break up due to the centripetal force. Thus, we can set these forces equal to find

$$\frac{F_{\rm cf}}{m} = \frac{F_g}{m} \quad \longrightarrow \quad \Omega^2 R = \frac{GM}{R^2} \tag{4.180}$$

For a given mass and radius, we have a minimum period

$$\Omega_{\rm max} = \sqrt{\frac{GM}{R^3}} \quad \longrightarrow \quad P_{\rm min} = 2\pi\sqrt{\frac{R^3}{GM}} = \sqrt{\frac{3\pi}{G\bar{\rho}}} \tag{4.181}$$

Alternatively, we can think of this as a minimum density for a given period

$$\boxed{\bar{\rho}_{\rm min} = \frac{3\pi}{GP^2}} \tag{4.182}$$

Using the shortest known period of 1.4 ms, this corresponds to a minimum density of

$$\boxed{\bar{\rho}_{\rm min} \simeq 7.2 \times 10^{13}\,{\rm g\,cm^{-3}}} \tag{4.183}$$

54. Show from a simple dimensional analysis how the effective temperature of an accretion disk depends on accretion rate and distance from the central object.

Gravitational energy goes like GM^2/r, so per unit time we have $GM\dot{M}/r$. By the Stefan-Boltzmann law, the flux is proportional to T^4. Since flux is energy per time per area (and area $\propto r^2$), we have

$$\frac{M\dot{M}}{r}\frac{1}{r^2} \propto T^4 \quad \longrightarrow \quad T \propto \dot{M}^{1/4} r^{-3/4} \tag{4.184}$$

Now, for a slightly more formal argument. The accretion of mass causes a release of gravitational potential energy

$$\frac{\mathrm{d}U}{\mathrm{d}t} = -\frac{GM\dot{M}}{r} \tag{4.185}$$

Using a simple virial argument (§12.9.2)

$$2K + U = 0 \quad \longrightarrow \quad K = -\frac{1}{2}U \tag{4.186}$$

Half of the released potential energy must go into kinetic energy, so the remaining half goes into radiation, and thus the luminosity is

$$L = -\frac{1}{2}\frac{\mathrm{d}U}{\mathrm{d}t} = \frac{GM\dot{M}}{2r} \tag{4.187}$$

And by the Stefan-Boltzmann law, if we assume the disk is optically thick, it's related to the effective temperature by

$$L = 2\pi r^2 \sigma_{\mathrm{SB}} T_{\mathrm{eff}}^4 \tag{4.188}$$

(with a factor of 2 since there are 2 sides of the accretion disk). Equating these two, we can solve for T_{eff} in terms of $\dot{M}$ and r:

$$\boxed{T_{\mathrm{eff}} = \left(\frac{GM\dot{M}}{4\pi\sigma_{\mathrm{SB}} r^3}\right)^{1/4}} \tag{4.189}$$

We get a scaling of

$$T_{\mathrm{eff}} \propto M^{1/4} \dot{M}^{1/4} r^{-3/4} \tag{4.190}$$

Comparing equation (4.189) to the alpha-disk model (4.132) shows that we got the scalings right, but the prefactor is slightly different, and the inner disk radius r_{in} creates a hard cutoff[22,29]. Also, notice that $R_{\mathrm{ISCO}} \propto M$ and $\dot{M} \propto L_{\mathrm{Edd}} \propto M$, so the temperature at the ISCO scales with the mass like

$$T_{\mathrm{eff,ISCO}} \propto M^{-1/4} \tag{4.191}$$

In other words, *less massive* compact objects have *hotter* accretion disks (because the ISCO is closer)!

55. What evidence is there for "superluminal" jets from black holes, and how does one explain the superluminal motion?

Some black holes produce relativistic jets of radiation that appear to have **superluminal** velocities ($v > c$). See, e.g., Figure 4.21. where a jet travels 30 lightyears in 6 years (5 lightyears per year).

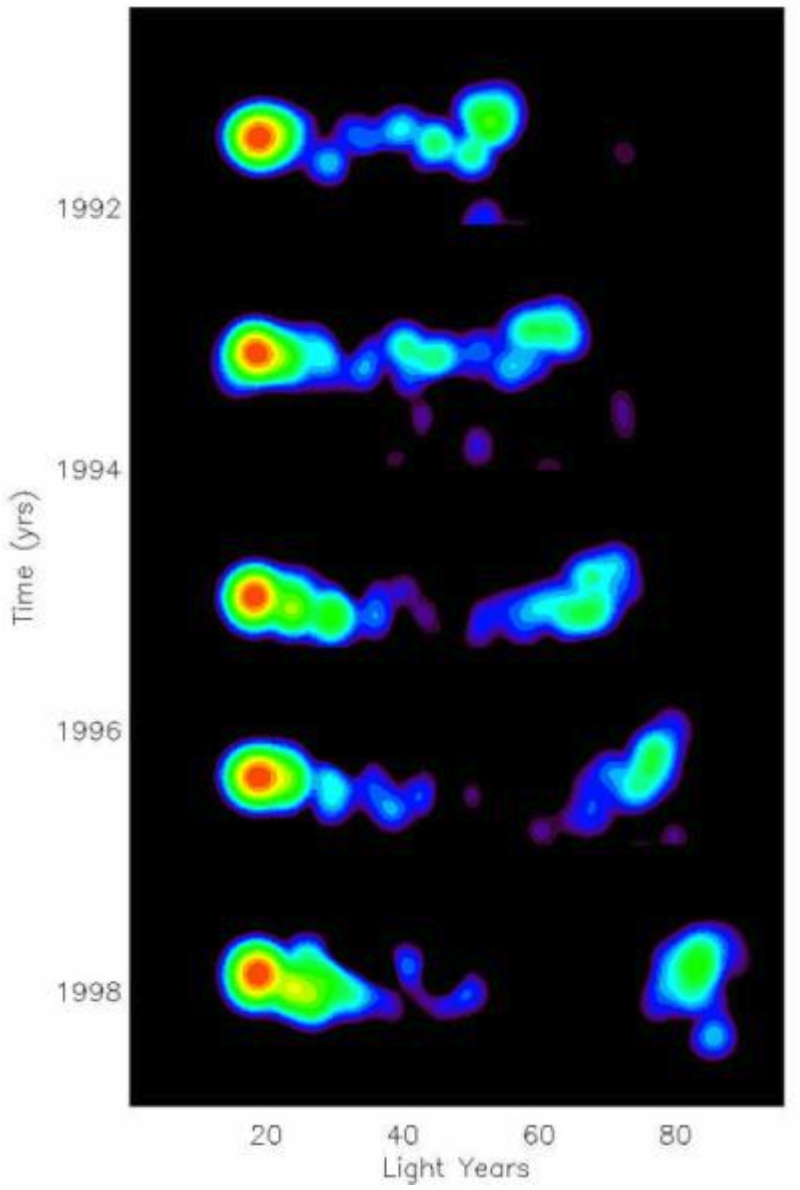


Figure 4.21: The apparent superluminal motion of an AGN radio jet from 3C 279

SR forbids any velocities $> c$, so how is this possible? This is actually a projection effect—the jets *appear* to be moving faster than c from our perspective, but they aren't really. We can see how this works by considering some blob of material that starts out at point P_1 at time t_1 and moves to point P_2 at time t_2 at speed v. It moves towards us along our line of sight, but at some angle θ so that it also has some proper motion. This setup is shown in Figure 4.22.

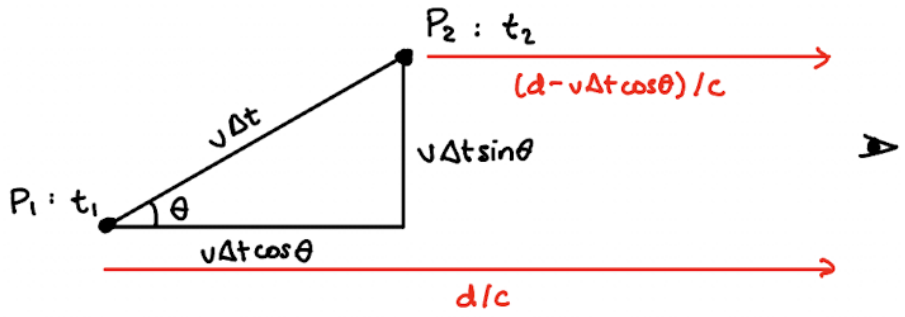


Figure 4.22: Diagram showing the observed time and distance between two events P_1 and P_2 along the path of motion of some blob of material.

In the rest frame of the blob, the time and distance it travels are:

$$\Delta t = t_2 - t_1 \tag{4.192}$$
$$\Delta d = v\Delta t \tag{4.193}$$

However, in the observed frame, since the object is moving towards us relativistically, the perceived time difference between the two events is shorter (the second event doesn't take as long to reach us as the first event did):

$$t_1' = \frac{d}{c} \tag{4.194}$$
$$t_2' = \Delta t + \frac{d - v\Delta t\cos\theta}{c} \tag{4.195}$$
$$\Delta t' = t_2' - t_1' = \Delta t(1 - \beta\cos\theta) \tag{4.196}$$

Of course, because of the angle, the perceived distance is also shorter:

$$\Delta x' = v\Delta t\sin\theta \tag{4.197}$$

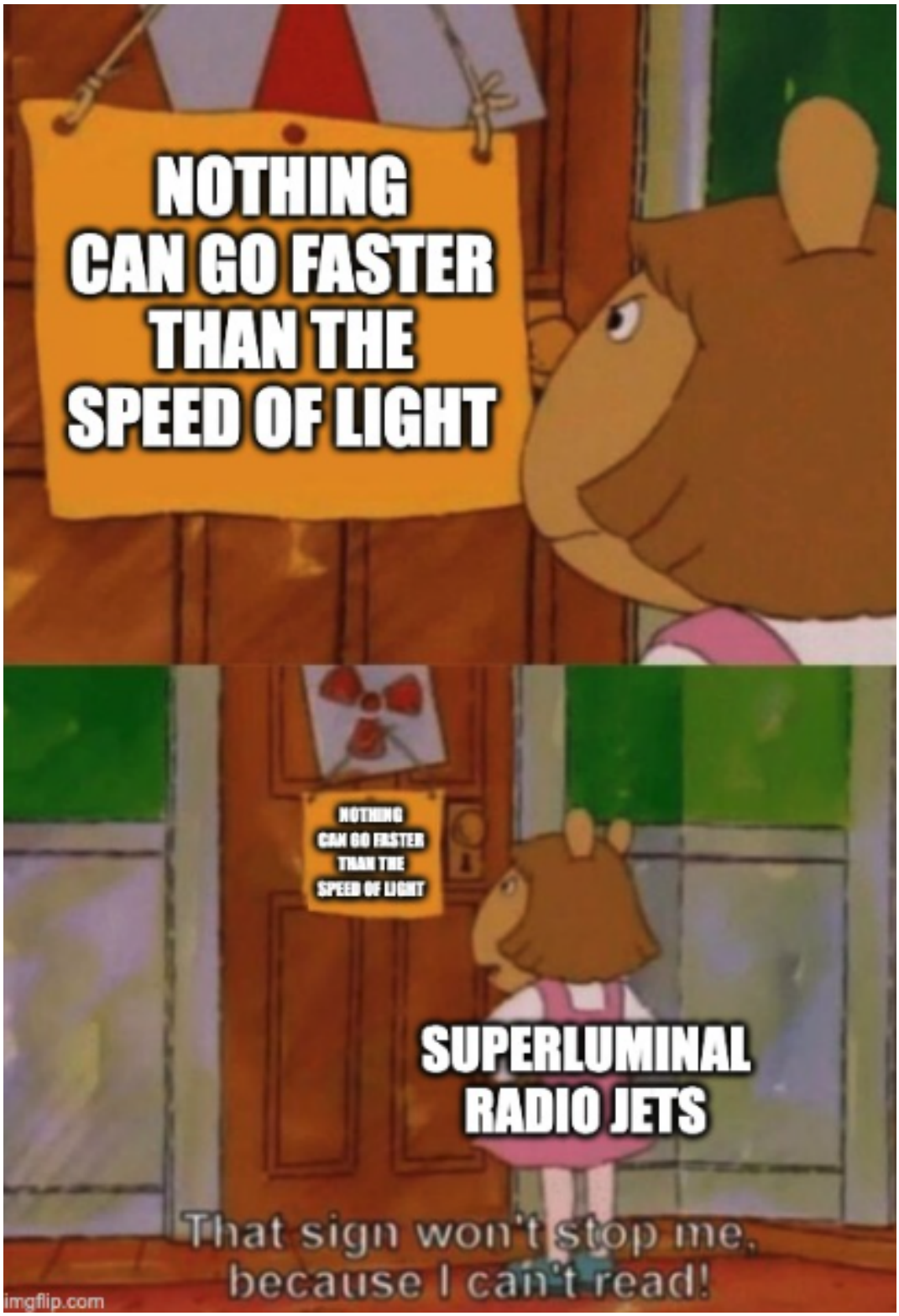
But this effect is not as strong as the effect on the time delay. The apparent speed ends up being:

$$\boxed{v_{\text{app}} = \frac{\Delta x'}{\Delta t'} = c\frac{\beta\sin\theta}{1 - \beta\cos\theta}} \tag{4.198}$$

Where $\beta = v/c$. Notice that the fraction can be greater than 1, depending on the values of θ and β. We can see that in the small angle limit this becomes

$$\frac{v_{\text{app}}}{c} \simeq \frac{\beta\theta}{1 - \beta} \tag{4.199}$$

As $\beta \to 1$, $v_{\text{app}}/c \to \infty$, allowing apparent "superluminal" motion.


NOTHING
CAN GO FASTER
THAN THE
SPEED OF LIGHT
NOTHING
CAN GO FASTER
THAN THE
SPEED OF LIGHT
SUPERLUMINAL
RADIO JETS
That sign won't stop me,
because I can't read!
imgflip.com

56. What is the historical significance of the Hulse-Taylor binary radio pulsar to physics?

The **Hulse-Taylor pulsar** was discovered in 1974 by Hulse and Taylor[68], and was the first known pulsar in a binary system. They were able to determine this by looking at the variations in its pulsation period over time. They found the pulsation period to be 59 ms, varying by 0.078 ms over the course of 7.75 hr, due to the orbital motion of the binary causing the pulses to take longer/shorter to reach us. From this, they got estimates of the radial velocities, and from those they could estimate the mass of the companion (i.e. §3.Q20). They determined the companion must also be a neutron star (but not a pulsar).

Having two compact objects in a close orbit like this is a prime laboratory for producing gravitational radiation (as it has a varying quadrupole moment, §4.Q42), and thus this system acted as an important test for Einstein's general relativity. This causes the system to lose energy, which causes the two neutron stars to come closer together. Recall in the Newtonian limit:

$$E = -\frac{GM_1M_2}{2a} \tag{4.200}$$

So as E decreases (gets more negative), a must also decrease. Figure 4.23 shows the measurements of the cumulative shifts in the pulsation period over time, matching perfectly what would be predicted from GR. This was one of the first pieces of evidence for the existence of gravitational waves[69]!

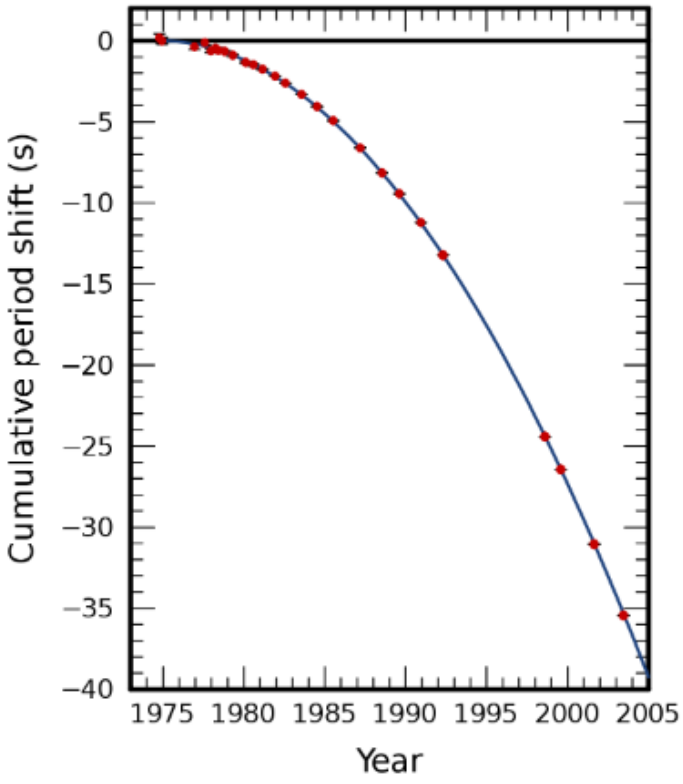


Figure 4.23: The cumulative shift of the Hulse-Taylor pulsar's pulsation period over time, in comparison to the predictions of GR[69].

57. What minimum mass is required for a black hole powering a quasar of luminosity $10^{12}\ L_\odot$? What accretion rate is required?

For a mass M, the maximum luminosity that can be produced is the Eddington luminosity $L_{\rm Edd}$. Therefore, if we have a given luminosity L, the *minimum* mass that could produce this luminosity must be the mass for which this is the Eddington luminosity. If the mass were any smaller, the Eddington luminosity would have to be lower, otherwise radiation pressure would dominate over gravity. Recalling our derivation of the Eddington limit from §4.Q47:

$$L_{\rm Edd} \simeq 3\times 10^4\left(\frac{M}{M_\odot}\right)L_\odot \quad\longrightarrow\quad M \simeq \frac{1}{3\times 10^4}\left(\frac{L_{\rm Edd}}{L_\odot}\right)M_\odot \tag{4.201}$$

This gives a minimum mass of

$$\boxed{M_{\rm min} \simeq 3\times 10^7\, M_\odot} \tag{4.202}$$

We can relate the luminosity to an accretion rate if we assume an efficiency of $\eta \sim 0.1$:

$$L = \eta \dot{M} c^2 \quad\longrightarrow\quad \dot{M} = \frac{L}{\eta c^2} \tag{4.203}$$

Which results in

$$\boxed{\dot{M} \simeq 0.67\, M_\odot\,{\rm yr}^{-1}} \tag{4.204}$$

58. How much energy is typically released in a type II supernova? What fractions of that energy are in the forms of neutrinos, visible light, gas kinetic energy and gravitational radiation?

For a core with a mass $M_{\rm core} \sim 1.4\,M_\odot$ contracting from an initial radius $r_i \sim 5000$ km to a final radius $r_f \sim 12$ km, the gravitational binding energy released is

$$\Delta E_g \simeq \frac{3}{5}\frac{GM_{\rm core}^2}{r_f} - \frac{3}{5}\frac{GM_{\rm core}^2}{r_i} \simeq \boxed{3\times 10^{53}\,\text{erg}} \tag{4.205}$$

where the factor of $3/5$ comes in due to the spherical symmetry.

Where does the energy go?[29,11]

1. **Photodisintegration of ^{56}Fe (7%)**

   The binding energy per nucleon in ^{56}Fe is around 9 MeV. So, the total energy that goes into photodisintegration will be 9 MeV × number of nucleons per Fe (56) × number of Fe atoms. Assuming a core mass of $M_{\rm core} \sim 1.4\,M_\odot$, we have $1.4\,M_\odot/56m_n$ Fe atoms, so

   $$\Delta E_{\rm nuc} \simeq 9\,\text{MeV} \times 56 \times \frac{1.4\,M_\odot}{56m_n} \simeq 2\times 10^{52}\,\text{erg} \tag{4.206}$$

2. **Ejecting the stellar envelope (2%)**

   The ejection of the outer stellar envelope from the initial radius r_i out to infinity requires gravitational energy. Say the total mass of the envelope and the core is $M \sim 10\,M_\odot$. Then the energy released is

   $$\Delta E_{\rm env} \simeq \frac{GM_{\rm core}(M - M_{\rm core})}{r_i} \simeq 6\times 10^{51}\,\text{erg} \tag{4.207}$$

3. **Driving the ejecta (3%)**

   Gas outflows involved with the ejection of the stellar envelope have observed speeds on the order $v \sim 10^4\ \text{km}\,\text{s}^{-1}$, giving them energies of

   $$\Delta E_{\rm kin} \simeq \frac{1}{2}(M - M_{\rm core})v^2 \simeq 9\times 10^{51}\,\text{erg} \tag{4.208}$$

4. **Electromagnetic radiation (0.3%, only 0.01% in the visible)**

   The shock produced by the supernova ionizes the Hydrogen-rich envelope, which leads to a long period of recombination that results in radiation. The radioactive decay of elements produced in the shock also produces radiation. The total energy is of order $\Delta E_{\rm rad} \simeq 10^{51}$ erg.

5. **Neutrinos (88%)**

   Neutrinos are produced from electron capture reactions ($\text{p}^+\text{e}^- \longrightarrow \text{n} + \nu_e$), which accounts for the remaining $\sim$ 88% of the energy released.

6. **Gravitational radiation? (0.1%?)**

   Gravitational radiation from supernovae has not yet been observed, but theoretically it should be produced by any asymmetries in the system.

59. Describe a scenario whereby a neutron star can be formed in a binary system and have the system remain bound

To remain bound, the energy of the orbit must remain less than 0 after one of the stars goes supernova (recall $E = 0$ corresponds to parabolic orbits, and $E > 0$ are hyperbolic orbits). We'll work in the reduced problem (i.e. §3.Q20) to simplify things. In this case the total energy in the initial configuration is

$$E = \frac{1}{2}\mu v^2 - \frac{GM\mu}{r} \tag{4.209}$$

If we assume the orbits are circular so $r = a$ (which is not necessary but just makes the math easier), then we can use Kepler's 3rd law to get the relative velocity

$$v^2 = \frac{GM}{a} \tag{4.210}$$

Now, after the supernova, the mass of one star will change, say by Δm, so the total mass changes to $M \to M - \Delta m$. This will also change the reduced mass $\mu \to \mu'$. Supernovae occur on timescales much faster than the orbital timescales, so we can look at the time immediately after the explosion, before the relative velocity v and orbital separation a have time to change. Our new energy becomes

$$E' = \frac{1}{2}\mu' v^2 - \frac{G(M-\Delta m)\mu'}{a} \tag{4.211}$$

$$= \frac{1}{2}\frac{GM\mu'}{a} - \frac{G(M-\Delta m)\mu'}{a} \tag{4.212}$$

Setting this to be < 0, we obtain the constraint

$$\frac{M}{2} < M - \Delta m \quad \longrightarrow \quad \boxed{\Delta m < \frac{M}{2}} \tag{4.213}$$

In other words, we must not lose more than half of the original total mass in order to remain bound[22,29,70]. This comes with a few consequences:

- **Low Mass X-ray Binaries (LMXBs)** are binary systems with a neutron star (or black hole) and a companion with a mass $\lesssim 2\ M_\odot$. By our above analysis, these systems should be nearly *impossible* to form. The progenitor stellar mass for a neutron star to form is $8\ M_\odot$. Let's assume the companion has a mass $2\ M_\odot$, the maximum allowed for an LMXB classification. Then, when the massive star goes supernova, it becomes a neutron star with a mass of $\sim 1.4\ M_\odot$, shedding a $\Delta m = 8 - 1.4 = 6.6\ M_\odot$. This is more than half of the original total mass ($8+2 = 10\ M_\odot$), so the supernova should unbind the system! Nevertheless, many LMXB systems are observed. These systems likely captured a companion star *after* the neutron star went supernova, meaning they require dense environments where interactions are more common (i.e. globular clusters; see §4.Q65). These systems can be long-lived, with ages $\gtrsim 10^7$ yr.
- **High Mass X-ray Binaries (HMXBs)** have a companion with a mass $\gtrsim 8\ M_\odot$. By the converse argument to LMXBs, HMXBs *are* possible to stay bound after a supernova. Since they require massive young stars, they are most often found in the stellar disk. These systems experience unstable mass transfer since mass is transferred from the more massive companion to the less massive NS/BH, which shrinks the orbit (§4.Q48), so they have short ages of $\lesssim 10^6$ yr.

- **Intermediate Mass X-ray Binaries** would theoretically have companion masses between 2–8 $M_\odot$. However, these systems are almost never observed. This is because Roche lobe overflow in these systems proceeds very quickly, so it's unlikely for us to observe such events. Stellar winds are also not very strong in these intermediate-mass companions.

60. What is a "cataclysmic variable"? What are nova explosions, and what is the basic physics underlying these events?

A **cataclysmic variable** is a binary system in which a white dwarf accretes material from a companion star that fills its Roche lobe and experiences occasional, unpredictable bursts in energy. Note: this is distinct from a type Ia supernova, in which an accreting white dwarf exceeds its Chandrasekhar limit and is destroyed. In a cataclysmic variable, the white dwarf does not exceed the Chandrasekhar limit, and the outbursts are caused by different mechanisms called **novae**. These novae are subdivided into two categories: **dwarf novae** and **classical novae**.

Typical parameters of a cataclysmic variable[11]:

- Orbital period $P \sim 20$ min–5 day (most between 1–12 hr)
- Primary (WD) mass $M \sim 0.86\,M_\odot$
- Secondary star typically a spectral type of G or later

Novae

1. **Dwarf Novae**[11]

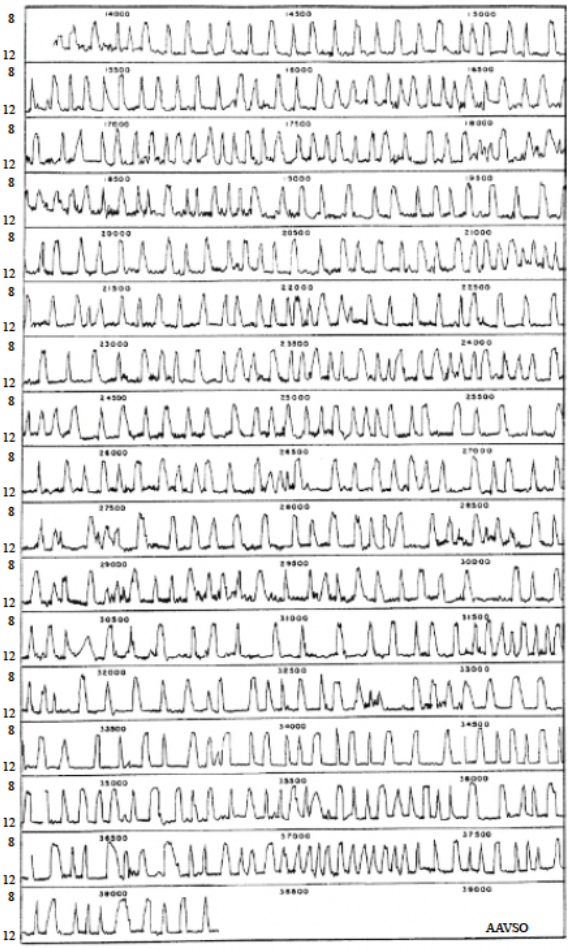


Figure 4.24: An example of a dwarf nova light curve (SS Cygni), compiled by the AAVSO[11]

These are caused by a sudden increase in the rate at which mass flows down through the accretion disk. These periods of increased mass flow typically last from 5–20 days, separated by quiescent intervals of 30–300 days. The mass transfer rates increase from typical values of $10^{-11} - 10^{-10}\,M_\odot\,\mathrm{yr}^{-1}$ up to values of $10^{-9} - 10^{-8}\,M_\odot\,\mathrm{yr}^{-1}$. Since the disk luminosity is proportional to $\dot{M}$, this increases the luminosity by a factor of 10–100. The mechanism that causes the change in $\dot{M}$ is not well known, but there are two possible explanations.

Instabilities in the mass transfer rate:

- The Hydrogen partial ionization zone in a star ($T \sim 10^4$ K) periodically dams up and released energy. This is sort of like the κ mechanism (see §3.Q24).
- This causes the outer layers of a star to periodically overflow its Roche lobe, turning mass transfer on and off.

Instabilities in the accretion disk itself:

- The chain of reasoning is similar, as this also uses the Hydrogen partial ionization zone. We also use the fact that the viscosity of the disk, which determines how fast mass flows through it, is proportional to its temperature.
- Below 10^4 K, hydrogen is mostly neutral, so opacity is low, cooling is efficient, temperature is low, and viscosity is low.
- Above 10^4 K, hydrogen is mostly ionized, so opacity is high (see Figure 3.33), cooling is inefficient, temperature is high, and viscosity is high.
- Therefore, periodic ionization and recombination of hydrogen can periodically change the viscosity of the disk, which likewise changes how fast mass flows through it.

See an example light curve in Figure 4.24.

2. **Classical Novae**[11]

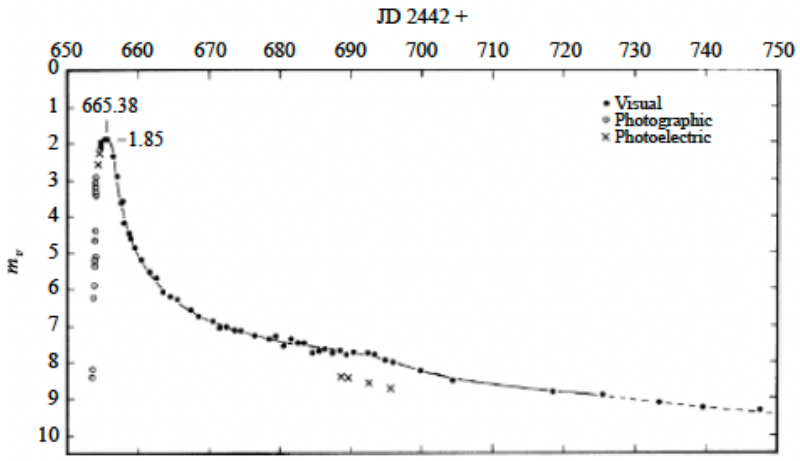
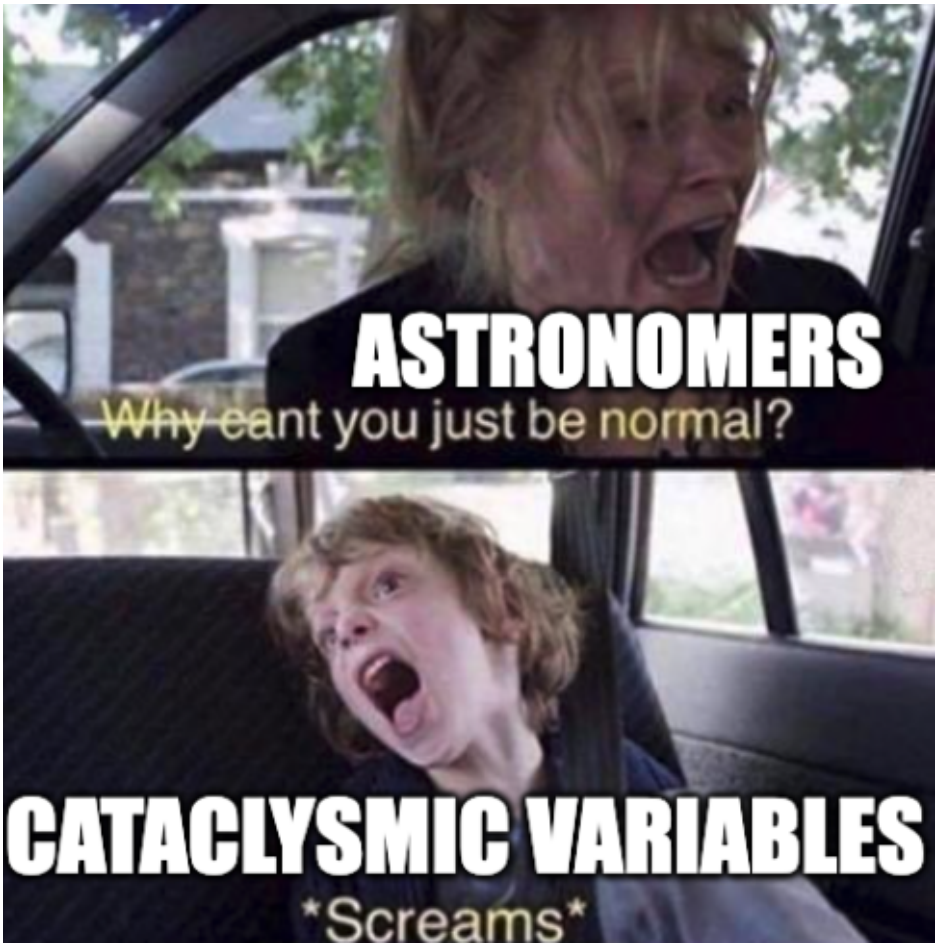


Figure 4.25: An example of a classical nova light curve (V1500 Cyg)[71,11].

These are caused by Hydrogen-rich gases accumulating on the surface of the white dwarf, heating up and mixing with carbon and oxygen from the white dwarf, and igniting hydrogen burning through the CNO cycle under degenerate conditions. This causes a runaway thermonuclear reaction, heating up to 10^8 K before the electrons lose their degeneracy pressure.

The luminosity eventually exceeds the Eddington limit, causing radiation pressure to expel the accreted material and returning the WD to a state before the nova. This process is able to happen multiple times in the same system. The luminosity typically increases by 7–20 magnitudes in the initial burst, followed by a slow decline back to the pre-nova levels over the course of weeks to months. The hot gas shell is typically expelled at 1000s of $\mathrm{km\,s^{-1}}$. See an example light curve in Figure 4.25



61. What is the gravitational redshift from the surface of a neutron star?

From the definition of redshift, we have

$$z = \frac{\lambda_{\text{obs}} - \lambda_{\text{emit}}}{\lambda_{\text{emit}}} = \frac{\lambda_{\text{obs}}}{\lambda_{\text{emit}}} - 1 = \frac{\nu_{\text{emit}}}{\nu_{\text{obs}}} - 1 = \frac{\delta t_{\text{obs}}}{\delta t_{\text{emit}}} - 1 \tag{4.214}$$

where δt_{obs} and δt_{emit} are the small time differences between two adjacent peaks/troughs that define the wavelength/frequency. We want to take δt_{obs} to be the time measured by an observer at ∞, and compare this to the δt_{emit} from an observer at a distance r_{emit}.

Using the Schwarzschild metric for a spherically symmetric gravitational field (non-rotating):

$$\mathrm{d}s^2 = -\left(1 - \frac{R_S}{r}\right)\mathrm{d}t^2 + \left(1 - \frac{R_S}{r}\right)^{-1}\mathrm{d}r^2 + r^2\mathrm{d}\Omega^2 \tag{4.215}$$

Then, δt_{emit} will be the proper time $\mathrm{d}\tau$, and δt_{obs} will be the coordinate time $\mathrm{d}t$, which are related by:

$$\mathrm{d}\tau^2 = \left(1 - \frac{R_S}{r}\right)\mathrm{d}t^2 \quad \longrightarrow \quad \mathrm{d}\tau = \left(1 - \frac{R_S}{r}\right)^{1/2}\mathrm{d}t \tag{4.216}$$

Thus, we have

$$z = \frac{\mathrm{d}t}{\mathrm{d}\tau} - 1 = \frac{1}{(1 - R_S/r_{\text{emit}})^{1/2}} - 1 \tag{4.217}$$

Simplifying,

$$\boxed{z = \left(1 - \frac{R_S}{r_{\text{emit}}}\right)^{-1/2} - 1} \tag{4.218}$$

If we take typical neutron star parameters of $M \sim 1.4\,M_\odot$ and $R \sim r_{\text{emit}} \sim 10$ km, we get a redshift of

$$\boxed{z \sim 0.31} \tag{4.219}$$

62. How much rotational kinetic energy can be stored in a neutron star? How does this help resolve the "energy budget" for the Crab nebula?

First off, some numbers for the Crab pulsar. Just remember, Crabs come in threes:

$$P \simeq 33 \text{ ms} \tag{4.220}$$

$$\dot{P} \simeq 10^{-13} \tag{4.221}$$

Rotational kinetic energy is given by

$$E_{\text{rot}} = \frac{1}{2} I \omega^2 \tag{4.222}$$

And the moment of inertia for a sphere is $I = (2/5)MR^2$, so

$$E_{\text{rot}} = \frac{4\pi^2 MR^2}{P^2} \tag{4.223}$$

Using numbers appropriate for the crab pulsar, we have $M \sim 1.4\,M_\odot$, $R \sim 10$ km, and $P \sim 33$ ms[72], so the rotational energy is

$$\boxed{E_{\text{rot}} \simeq 10^{50}\,\text{erg}} \tag{4.224}$$

The neutron star spins down over time, which releases energy. Taking the derivative of (4.223):

$$\frac{\mathrm{d}E_{\text{rot}}}{\mathrm{d}t} = -\frac{8\pi^2 MR^2}{P^3}\dot{P} = -2E_{\text{rot}}\frac{\dot{P}}{P} \tag{4.225}$$

The crab pulsar has a spin-down rate of $\dot{\nu} \simeq -3.7 \times 10^{-10}$ Hz/s[72], corresponding to $\dot{P} = -P^2\dot{\nu} \simeq 4 \times 10^{-13}$

$$\boxed{\frac{\mathrm{d}E_{\text{rot}}}{\mathrm{d}t} \simeq -2 \times 10^{39}\,\text{erg}\,\text{s}^{-1}} \tag{4.226}$$

Compared to the observed luminosity of the entire Crab nebula:

$$\boxed{L \simeq 4 \times 10^{38}\,\text{erg}\,\text{s}^{-1}} \tag{4.227}$$

We can see that $|\,\mathrm{d}E_{\text{rot}}/\mathrm{d}t\,| > L$, so the Crab nebula can be entirely powered by the spin-down of the Crab pulsar[24].

Fun note about the Crab nebula: the supernova that created this remnant was observed by Chinese astronomers in 1054, making it the first supernova remnant that corresponds to a historically observed supernova event[72].

63. Explain why you might expect most of the emission of an accreting neutron star to be in the form of X-rays. How does an X-ray pulsar pulse?

Recall that the luminosity of an accretion disk is related to its effective temperature by

$$L = 2\pi R^2 \sigma_{\text{SB}} T_{\text{eff}}^4 \tag{4.228}$$

where the factor of 2 is from the 2 sides of the disk. If we assume the radiation is Eddington-limited, then for a $1.4\,M_\odot$ neutron star, the luminosity is

$$L = L_{\text{Edd}} = 3 \times 10^4 \left(\frac{M}{M_\odot}\right) L_\odot \simeq 10^{38}\,\text{erg}\,\text{s}^{-1} \tag{4.229}$$

Solving for the temperature, assuming the radiation is emitted at the surface of the neutron star at 10 km

$$T_{\text{eff}} = \left(\frac{L}{2\pi R^2 \sigma_{\text{SB}}}\right)^{1/4} \simeq 2 \times 10^7\,\text{K} \tag{4.230}$$

Using Wein's displacement law (§12.11), this corresponds to a peak wavelength of

$$\lambda_{\text{peak}} = \frac{2898\,\mu\text{m}\,\text{K}}{2 \times 10^7\,\text{K}} \simeq 10^{-4}\,\mu\text{m} \simeq \boxed{1\,\text{Å}} \tag{4.231}$$

This is solidly in the hard X-ray range (see §12.2)[24]. Technically, we will see in the second part of this question and in the next question §4.Q64 that the accretion disk in these systems does not exist all the way down to the neutron star's surface at 10 km, so if we wanted to get the luminosity just from the disk itself, we should really be using the *magnetospheric radius* (which is $\sim$ an order of magnitude larger). We can also use the α-disk model (4.132) for a more accurate solution. If we evaluate at an example radius of $r = 2r_{\text{in}}$, with $r_{\text{in}} = 100$ km, $M = 1.4\,M_\odot$, and $\dot{M} = L_{\text{Edd}}/\eta c^2$ (with $\eta = 0.1$), we get a temperature and wavelength of

$$T_{\text{eff}} = \left[\left(\frac{3GM\dot{M}}{8\pi\sigma_{\text{SB}} r^3}\right)\left(1 - \sqrt{\frac{r_{\text{in}}}{r}}\right)\right]^{1/4} \simeq 2 \times 10^6\,\text{K} \tag{4.232}$$

$$\lambda_{\text{peak}} = \frac{2898\,\mu\,\text{K}}{2 \times 10^6\,\text{K}} \simeq \boxed{10\,\text{Å}} \tag{4.233}$$

which is now in the soft X-ray regime.

How does an X-ray pulsar pulse?

The pulsar has a strong dipolar magnetic field which disrupts and truncates the inner accretion disk well above the surface of the neutron star. The accretion flow is channeled along the pulsar's magnetic field lines onto its poles. See a diagram in Figure 4.26.

The accretion on the poles of the neutron star now looks like a column, and it's traveling at supersonic speeds. This produces a combination of bremsstrahlung (§5.Q74) and synchrotron (§5.Q75) radiation from electrons in the accretion flow, and blackbody radiation from the surface of the neutron star. These three processes produce so-called "seed" photons, which have mostly soft X-ray energies. These photons are then upscattered to higher (hard) X-ray energies by **inverse Compton scattering** with the relativistic electrons in the accretion flow, in a process called **Comptonization**. This whole process is shown diagramatically in Figure 4.27.

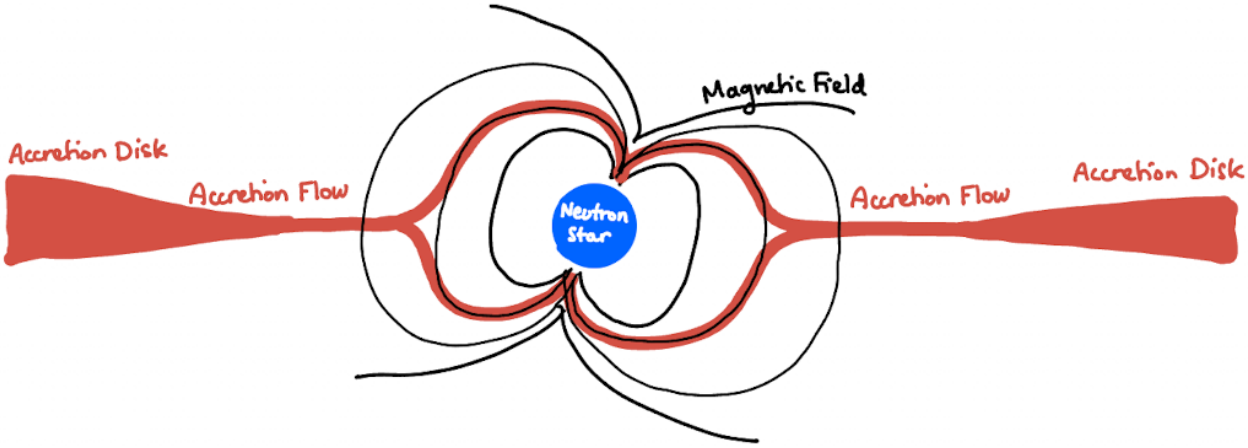


Figure 4.26: The accretion flow of an X-ray pulsar. Notice where it begins following the magnetic field lines and coalesces onto the poles of the neutron star.

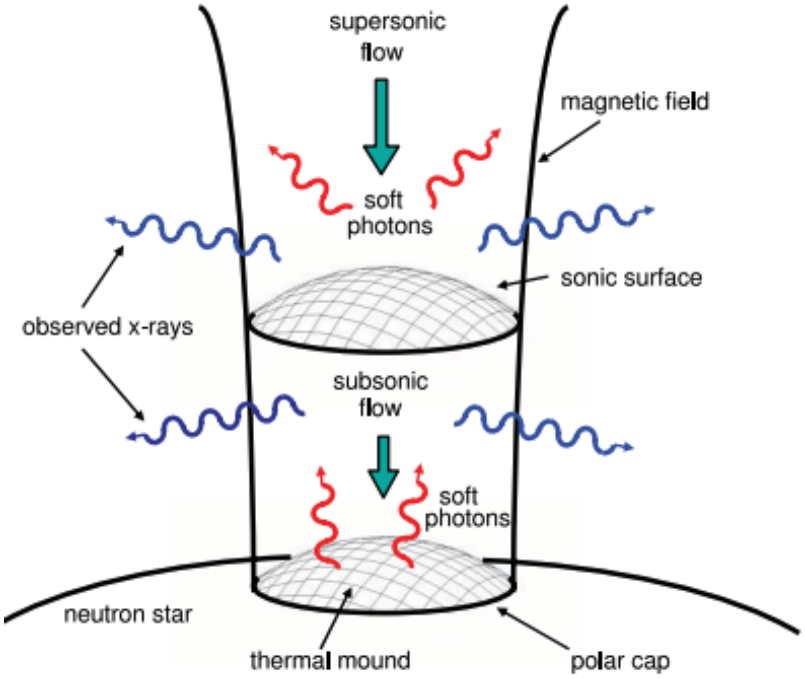


Figure 4.27: The accretion column on the pole of a neutron star. Soft seed photons are created through bremsstrahlung, synchrotron, and blackbody radiation, which are then upscattered via inverse Compton scattering. Figure from Ref. 73.

The spin axis of the pulsar is often misaligned with the magnetic axis, so as the pulsar rotates, the magnetic hotspots pass in and out of our field of view, creating pulses. However, this simple picture is likely not the full story, as recent observations from NICER have revealed hotspots on pulsars that are not dipolar, meaning the magnetic field configuration is more complicated than what we have been imagining[74].

64. Derive a plausible relation between the luminosity of an X-ray pulsar and its spin-up rate. How would you calculate the magnetospheric radius of an accreting neutron star?

This question is a bit tricky and requires thinking about the physics of the situation and how exactly the luminosity is being generated. I'll start with a simpler example of an isolated pulsar that is not accreting any material and is just slowing down over time. This is also called a **radio pulsar** to distinguish it from accreting **X-ray pulsars**. Then, we'll upgrade to the more complicated situation that the question actually asks about.

Luminosity for a radio pulsar

For an **isolated** pulsar, the luminosity would just be equal to the change in rotational kinetic energy (where, in this case $\dot{\omega}$ is a spin-**down** rate):

$$L = \frac{\mathrm{d}E_{\mathrm{rot}}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}I\omega^2\right) = I\omega\dot{\omega} \tag{4.234}$$

Then recall $I = (2/5)MR^2$ and $\omega = \sqrt{GM/R^3}$:

$$L = \frac{2}{5}MR^2 \times \sqrt{\frac{GM}{R^3}} \times \dot{\omega} \quad \longrightarrow \quad L = \frac{2}{5}\sqrt{GM^3R}\dot{\omega} \tag{4.235}$$

We see that L and $\dot{\omega}$ are directly proportional to each other.

Luminosity for an accreting X-ray pulsar[22]

In contrast to the previous case, we now want to consider the case of an X-ray pulsar which is **accreting**. In this case, the equation above no longer applies because the radiated energy is not coming from the rotation, but rather accretion. The infalling material in the accretion disk will reach some inner radius r_m (called the **magnetospheric radius**) before being channeled along the magnetic field lines onto the poles of the pulsar (see §4.Q64). This exerts a torque on the pulsar, spinning it up. Each chunk of material with mass m has an angular momentum ℓ and torque τ:

$$\ell = mr_m^2\omega_{\mathrm{disk}} \tag{4.236}$$

$$\tau = \dot{\ell} = \dot{m}r_m^2\omega_{\mathrm{disk}} = \dot{m}\sqrt{Gmr_m} \tag{4.237}$$

Note: in calculating τ, we don't need a $\dot{\omega}_{\mathrm{disk}}$ term because the rotation of the *disk* is constant. It's only the rotation of the *pulsar* that speeds up. We can then replace $m \to M$ and $\dot{m} \to \dot{M}$ to get the total torque from the disk. Assuming all of the angular momentum is transferred into the rotation of the pulsar, we get

$$\tau = I\dot{\omega} = \frac{2}{5}MR^2\dot{\omega} = \dot{M}\sqrt{GMr_m} \quad \longrightarrow \quad \dot{\omega} = \frac{5}{2}\frac{\dot{M}}{R^2}\sqrt{\frac{Gr_m}{M}} \tag{4.238}$$

Now, the luminosity generated from accretion will just be proportional to $\dot{M}$: $L = \eta\dot{M}c^2$. And, as I will prove in the next section, the magnetospheric radius is proportional to $\dot{M}^{-2/7}$. Therefore, we have

$$\dot{\omega} \propto \dot{M}^{6/7} \quad \longrightarrow \quad \boxed{L \propto \dot{\omega}^{7/6}} \tag{4.239}$$

The Alfvén Radius[22]

The magnetospheric radius r_m, also sometimes called the **Alfvén radius**, is the radius at which the accretion disk is disturbed and the infalling material starts being channeled along the magnetic field lines (as mentioned above). To make a more precise definition, we call this the radius at which the magnetic energy density u_B is equal to the kinetic energy density u_K of the accretion flow:

$$u_B = \frac{B^2(r)}{8\pi} \quad , \quad u_K = \frac{1}{2}\rho v^2 \tag{4.240}$$

Now, in a steady state, our mass flow rate $\dot{M}$ will be the mass flux density (ρv_r) times the cross-sectional area of the disk ($2\pi r \times r\Delta\theta$), where $\Delta\theta$ is the angle subtended by the inner disk as seen by the pulsar. Rearranging for the density:

$$\rho = \frac{\dot{M}}{2\pi v_r r^2 \Delta\theta} \tag{4.241}$$

Now let's also assume the total velocity $v^2 = v_K^2 + v_r^2$ is dominated by the orbital Keplerian velocity ($v_K \gg v_r$). We can then write the inwards radial velocity v_r using the small parameter $\eta = v_r/v_K \ll 1$:

$$v_r = \eta v_K = \eta\sqrt{\frac{GM}{r}} \tag{4.242}$$

Then the kinetic energy density is

$$u_K = \frac{1}{2}\frac{\dot{M}}{2\pi r^2\Delta\theta}\eta\sqrt{\frac{GM}{r}} = \frac{\eta}{4\pi\Delta\theta}\frac{\dot{M}\sqrt{GM}}{r^{5/2}} \tag{4.243}$$

Now, for the magnetic energy density, the magnetic field of a dipole goes like $B(r) = \mu/r^3$. Therefore, if we define the magnetic field strength at the poles of the neutron star as $B(R) = B_p$, we can write

$$B(r) = B_p\left(\frac{R}{r}\right)^3 \tag{4.244}$$

Giving a magnetic energy density of

$$u_B = \frac{B_p^2}{8\pi}\left(\frac{R}{r}\right)^6 \tag{4.245}$$

Setting (4.243) and (4.245) equal and solving for $r = r_m$, we obtain

$$\boxed{r_m = \left(\frac{\Delta\theta^2 B_p^4 R^{12}}{4\eta^2 GM\dot{M}^2}\right)^{1/7}} \tag{4.246}$$

Notice the proportionalities:

$$\boxed{r_m \propto B_p^{4/7} R^{12/7} M^{-1/7} \dot{M}^{-2/7}} \tag{4.247}$$

Also note in the less realistic case of full spherical accretion, $v_K \to 0$ (free-fall), $\eta \to \sqrt{2}$, and $2\pi\Delta\theta \to$

4π, so our constant prefactor in r_m becomes $\Delta\theta^2/4\eta^2 \rightarrow 1/2$ (see Ref. 75).

See a diagram of the Alfvén radius in Figure 4.28.

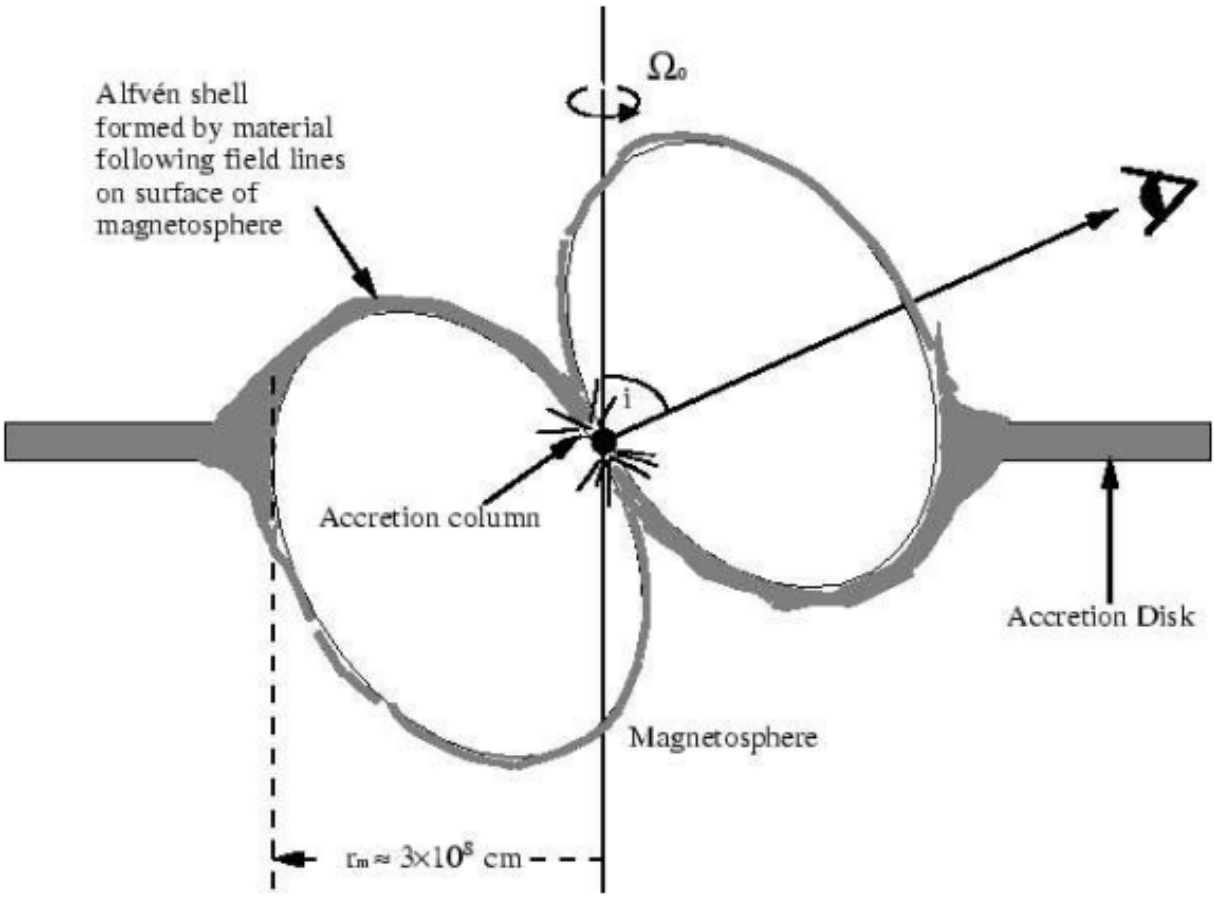


Figure 4.28: A rough sketch of the accretion flow of a disk being picked up by the magnetic field lines at the Alfvén radius[75].

65. Why are more X-ray binaries and radio pulsars found (per unit mass) in globular clusters than in other parts of the galaxy?

There are a couple of reasons why X-ray binaries and radio pulsars are more common in globular clusters:

1. Globular clusters have old, evolved populations of stars, meaning there is a higher number density of stellar remnants, including neutron stars and radio pulsars.
2. Globular clusters have a higher number density of stars, which means dynamical interactions are more common. Certain systems like LMXBs can only be formed by dynamical captures, since a supernova explosion would unbind any binary system with low enough mass (see §4.Q59). Therefore, a dense globular cluster makes it easier to form these systems through dynamical captures (either tidal captures between individual objects, or binary exchange interactions where a massive compact object interacts with a pre-existing binary and kicks out the lowest mass object, taking its place as a part of the binary).

Note that HMXBs are *less* common in globular clusters since they require young, high mass O/B type stars which only exist in the galactic disk.

66. What are "magnetars", and what evidence do we have that they exist?

A **magnetar** is a neutron star with an extremely strong magnetic field, $B \sim 10^{15}$ G compared to the typical neutron star field strength of $B \sim 10^{12}$ G. They also have relatively slow rotation periods of 5–8 seconds.

Evidence for their existence

Magnetars are the proposed origins for **soft gamma repeaters** (SGRs), objects that emit bursts of hard X-rays and soft gamma-rays with energies up to 100 keV. The source of these bursts are stresses in the magnetic fields that cause the surface of the neutron star to "crack", readjusting the surface and producing a super-Eddington release of energy.

There are only a few that are currently known (24, with 6 additional candidates) they all correlate with very young ($\lesssim 10^4$ yr) supernova remnants, suggesting that they may be a transient phenomenon that become "extinct" some time after a supernova[11].

The "smoking gun" evidence for magnetars came when we obtained measurements of both P and $\dot{P}$ for one of these systems, which, from $B \propto \sqrt{P\dot{P}}$ (see §4.Q70), gave the extremely large magnetic field expected from a magnetar.

Another piece of evidence in favor of magnetars is the existence of **Anomalous X-ray Pulsars (AXPs)**. These are pulsars that emit in the X-rays but are not in a binary accreting system. Emission from the loss of rotational energy is not strong enough to produce X-rays, so the source of energy in these systems is thought to be the decay of the strong magnetic fields over time, requiring the presence of a magnetar[76].

67. Describe the basic features of the "fireball" model of gamma-ray burst afterglows. What is the emission mechanism? What inputs are required?

See question §4.Q68 for a description of what a GRB is. The important point, which we'll start the discussion here with, is that a GRB is thought to be produced by a relativistic jet from a black hole that forms either during a supernova or during a merger with a neutron star, and that happens to be oriented towards us.

The Fireball Model[77] (see Figure 4.29)

1. The immense energies initially produced in a compact region during the supernova/merger event fuel the creation of electron-positron (e^-e^+) pairs. This creates a "fireball"—a plasma that is optically thick (due to electron scattering), preventing radiation from escaping.
2. The fireball expands due to its own pressure, which decreases its temperature. It also sweeps up baryons surrounding the initial explosion. Eventually it gets cool enough for e^-e^+ to recombine, which reduces the opacity, but the extra baryons that have been swept up can keep the fireball optically thick, preventing energy from escaping.
3. Most of the energy within the fireball is then converted from radiation into bulk kinetic energy of the baryons, since they are accelerated with the fireball. This is called the *matter-dominated phase* to distinguish it from the previous *radiation-dominated phase*.
4. Different shells in the fireball can have different relativistic γ factors, causing them to catch up with each other, which produces internal shocks in the jet.
5. Internal shocks within the jet accelerate electrons through shock drift acceleration[3], which causes them to emit synchrotron radiation. This produces a distinctly non-thermal distribution.
6. These photons are then inverse-Compton (IC) scattered up to gamma-ray energies. The jet is now optically thin because it has expanded and decreased in density, so this produces the observed GRB spectra.
7. Some time later, the jet collides with the surrounding ISM and produces external shocks. This produces synchrotron radiation in the same way as the initial GRB. As the shock loses energy over time, the emission from these external shocks gradually shifts from the X-rays to lower energy bands (UV, optical, IR, radio) before dissipating. This can last days to weeks.
8. Eventually the shock front expands so it covers a solid angle greater than just our field of view, so we suddenly see the flux start dropping since we are no longer receiving all of the radiation. This creates a "knee" in the light curve, i.e. Figure 4.30.

[3] See a cool visualization of this effect here!

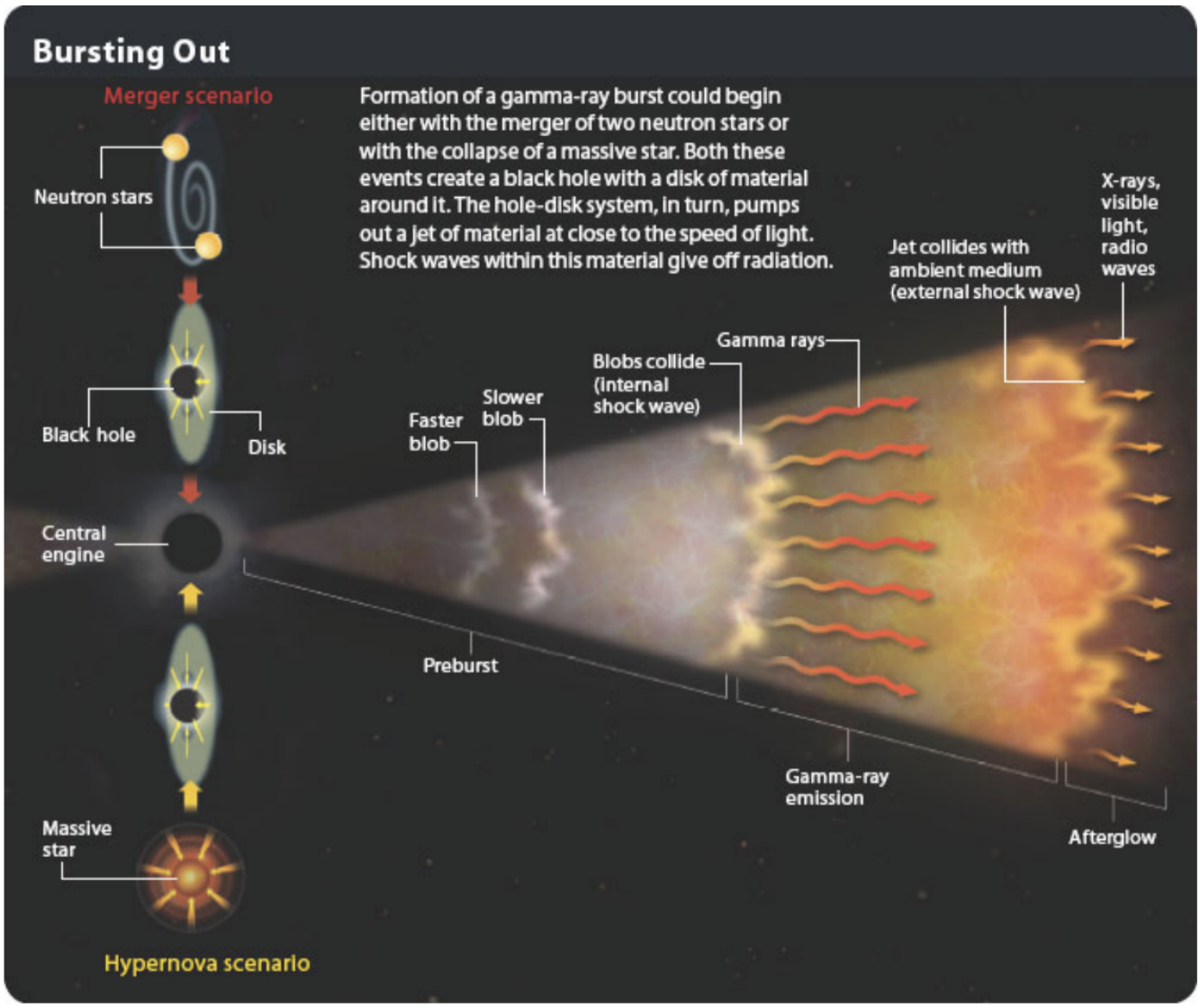


Figure 4.29: The fireball model for a GRB afterglow[78].

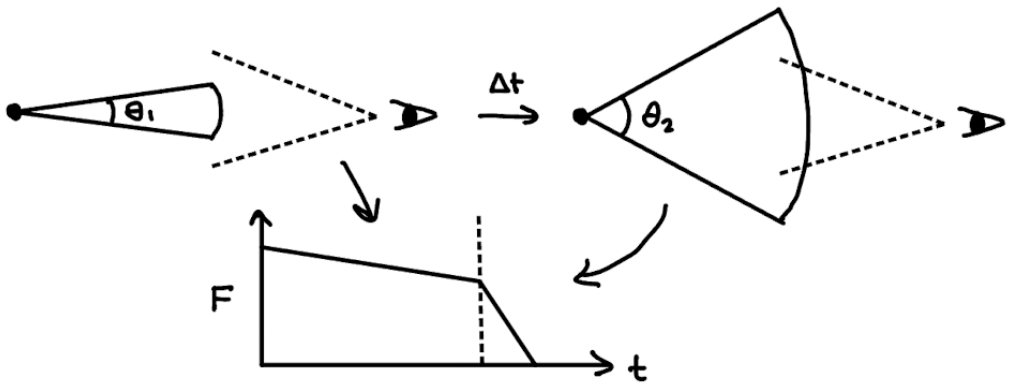


Figure 4.30: The effect of the jet opening angle increasing larger than our field of view, which lowers the flux.

68. What are the differences between short- and long-duration gamma ray bursts (GRBs)? What do we know about the physical origins of these energetic phenomena?

A **gamma-ray burst (GRB)** is a transient event characterized by a sudden burst of gamma-ray photons with energies ranging from 1 keV to many GeV. These bursts last from 10^{-2}–10^{3} s. They have rapid rises as fast as 10^{-4} s followed by a slower exponential decay. The actual profiles of these bursts can be complex, with multiple peaks, and there is no "typical" burst profile. Nevertheless, an example light curve is shown in Figure 4.31. GRBs are distributed isotropically across the sky, suggesting an extragalactic origin source (see §8.Q151).

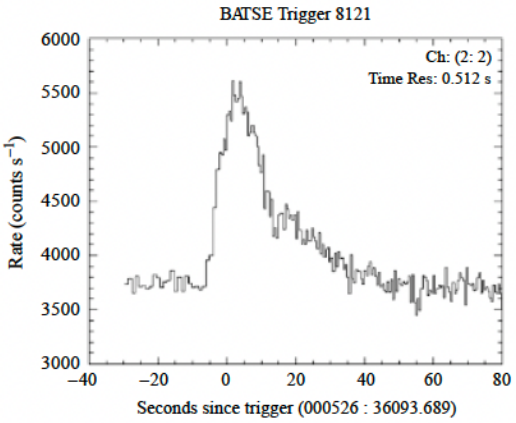


Figure 4.31: An example of a GRB: GRB 000526 observed by BATSE on the Compton Gamma-Ray Observatory[11].

The actual origin of the GRB depends on the subclassification. The population of GRBs is bimodal in duration space—there are **short (/hard) GRBs** with durations $\lesssim 2$ s, and **long (/soft) GRBs** with durations $\gtrsim 2$ s. See the distribution in Figure 4.32.

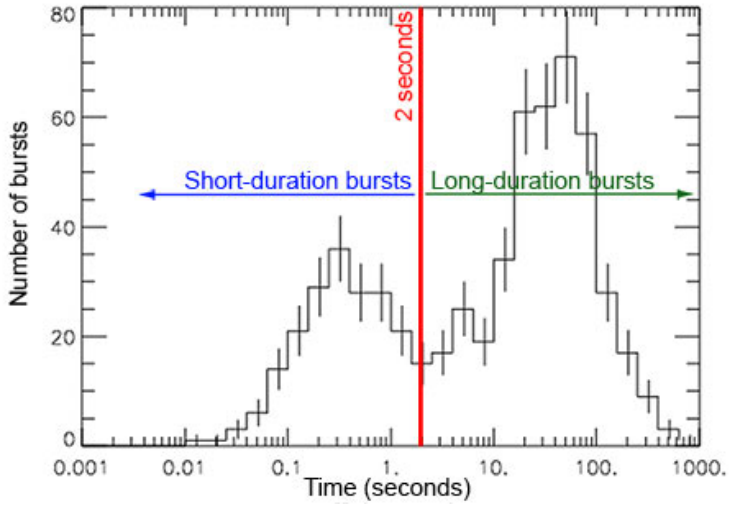


Figure 4.32: The population of GRBs as a function of duration. The distribution is bimodal, with the population being separated into short and long bursts.

GRBs were originally discovered serendipitously in the 1960s in an attempt to monitor the Soviet Union with the Vela military satellites. They were looking for sudden bursts of gamma rays that would presumably be produced during nuclear tests, which were banned with the 1963 nuclear test ban treaty. Instead, they found these anomalous events that appeared to have an extra-terrestrial origin.

The sub-categorizations of GRBs are as follows.

1. **Long/Soft GRBs**[11,79]

   These events are attributed to core-collapse supernovae. There are two main models that explain the gamma-ray emission.

   - The **collapsar** model (AKA the **hypernova** model) proclaims that the most massive stars will form a black hole with a debris disk in the supernova phase, while the star is still collapsing. The debris disk has an associated magnetic field that produces a collimating effect, causing a jet to emanate from the center of the supernova and relativistically beam gamma-rays. If we happen to lie along the line-of-sight of this jet, we will see a GRB.
   - The **supranova** model instead suggests that there is a delay mechanism in the black hole formation: initially, a neutron star is produced with a rapid rotation rate that allows it to stay together at masses higher than the typical upper limit for a non-rotating neutron star. This neutron star then slows down in the weeks–months after the supernova and becomes unstable against gravitational collapse, forming a black hole, which may have a debris disk. The GRB is then produced by the same jet mechanism as above.

   Both of these models invoke a relativistic beaming effect, which suggests that our measured energies are much larger than the intrinsic energies ($E \propto \gamma$, with $\gamma \sim 100$). Additionally, the interaction of these relativistic outflows with the surrounding medium produces synchrotron radiation with a much longer duration than the gamma-ray pulses. This is called the **afterglow** (see §4.Q67).

2. **Short/Hard GRBs**[11,79]

   These are thought to be due to the merger of two compact objects (NS–NS or NS–BH), which produces a kilonova with a jet and a relativistic outflow similar to a long GRB. These also can produce afterglows, but typically with lower energies than long GRBs.

   Note: the Hulse-Taylor binary (§4.Q56) is destined to produce a short GRB, although we may or may not be able to see it depending on the orientation of the jet.

   Another important object here is GW170817, which was the first NS–NS binary merger that was detected both in gravitational waves (with LIGO) and in electromagnetic waves as a GRB (with Fermi), with a lag time of 1.7 s[80]. The lag time is thought to be due to a lag between the emission mechanisms of the gravitational waves and the GRB, with the gravitational waves being produced during the collision and the gamma-rays coming shortly after, but the fact that it is so short despite the vast distance to the source proves that gravitational waves must travel at the speed of light to better than 1 part in 10^{15}. This is one type of **multi-messenger** study which uses a combination of EM and GW signals to provide a fuller picture of the underlying physics. The GWs give us information about the masses of the neutron stars and their distance, while the GRB gives us information on the produced black hole and relativistic jet.

69. How would you go about estimating the number of binary systems in the galaxy that contain a black hole? What observational constraints do we have on this number?

We can do this in two steps. First, we can estimate the fraction of stars that will ultimately become a black hole by using the **initial mass function (IMF)**, which gives the number of stars as a function of mass during their initial formation episode. In other words, $\phi(M)\mathrm{d}M$ is the number of stars formed with masses between $(M, M + \mathrm{d}M)$.

$$\phi(M) = \frac{1}{N_0}\frac{\mathrm{d}N}{\mathrm{d}M} \tag{4.248}$$

Where N_0 is a normalization constant. This is often chosen such that the total stellar mass formed is $1\,M_\odot$. In other words:

$$\int_{m_l}^{m_u} \mathrm{d}M\, M\phi(M) \equiv 1\,M_\odot \tag{4.249}$$

This requires that our normalization constant N_0 be the total number of solar masses that was actually formed:

$$N_0 = \frac{M_*}{M_\odot} \quad\longrightarrow\quad \phi(M) = \frac{M_\odot}{M_*}\frac{\mathrm{d}N}{\mathrm{d}M} \tag{4.250}$$

Observational constraints of the IMF suggest it generally follows a power law-like distribution. The first and most simple IMF is the **Salpeter IMF**[81 22,29]:

$$\phi(M) \propto \left(\frac{M}{M_\odot}\right)^{-2.35} \tag{4.251}$$

where there is some normalization related to N_0. Note that often people will also look at the IMF in logarithmic mass space, which modifies the power law index by 1:

$$\phi(\log_{10} M)\mathrm{d}\log_{10} M = \phi(M)\mathrm{d}M \tag{4.252}$$

$$\phi(\log_{10} M) = \ln(10) M\phi(M) \tag{4.253}$$

$$\phi(\log_{10} M) \propto \left(\frac{M}{M_\odot}\right)^{-1.35} \tag{4.254}$$

Later IMFs were developed by Kroupa[82] and Chabrier[83] which level off the slope at the low-mass end (see Figure 4.33).

For the purposes of this problem, we'll just consider the Salpeter IMF. So, to find the fraction of stars that will end up becoming black holes (requiring initial masses of $\gtrsim 20\ M_\odot$), we just integrate the IMF above 20 $M_\odot$, and normalize by integrating the IMF over all possible starting masses (m_l, m_u) = (0.08, 100) $M_\odot$. The lower limit corresponds to the Hydrogen burning limit, and the upper limit corresponds to the Eddington limit.

$$f_{\mathrm{BH}} = \frac{\int_{20}^{100} \mathrm{d}M\, \phi(M)}{\int_{0.08}^{100} \mathrm{d}M\, \phi(M)} \tag{4.255}$$

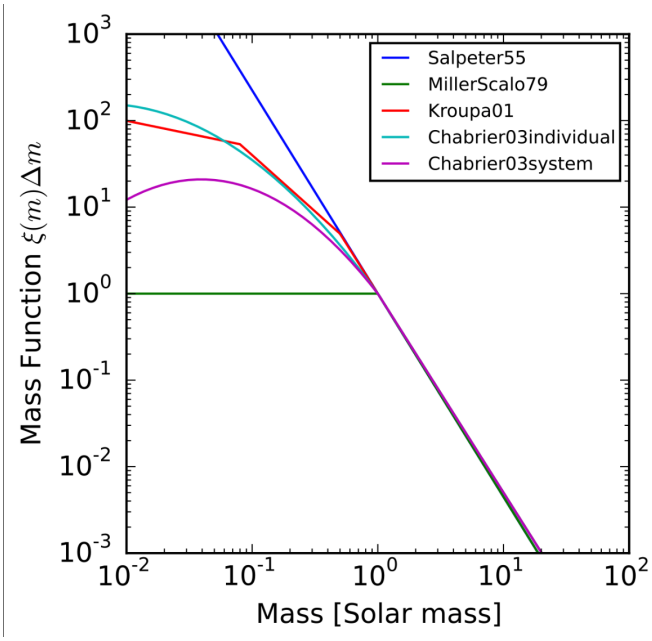


Figure 4.33: Various initial mass functions. Notice the primary differences lie at the low-mass end of the distribution.

We can approximate the upper bound as infinity since $\phi(100\,M_\odot)$ is negligible. Integrating, we find

$$f_{\rm BH} = \frac{M^{-1.35}\Big|_{20}^{\infty}}{M^{-1.35}\Big|_{0.08}^{\infty}} = \frac{20^{-1.35}}{0.08^{-1.35}} \simeq 6\times 10^{-4} \tag{4.256}$$

Now we have to consider how many of these stars that form black holes will be in binary systems. We can make a naïve assumption that half of the stars formed will be a part of a binary system. If this is the case, then we just divide our fraction by 2:

$$f_{\rm BHB} \simeq 3\times 10^{-4} \tag{4.257}$$

Now we need an estimate of the total number of stars in the galaxy to convert this fraction into a number. The average star formation rate of the Milky Way is $\langle\psi\rangle \simeq 3\,M_\odot\,{\rm yr}^{-1}$, and the average mass of stars formed is

$$\langle M\rangle = \frac{\int_{0.08}^{\infty} {\rm d}M\, M\phi(M)}{\int_{0.08}^{\infty} {\rm d}M\,\phi(M)} = \frac{\dfrac{M^{-0.35}}{-0.35}\Big|_{0.08}^{\infty}}{\dfrac{M^{-1.35}}{-1.35}\Big|_{0.08}^{\infty}} = \frac{0.08^{-0.35}/0.35}{0.08^{-1.35}/1.35} \simeq 0.31\,M_\odot \tag{4.258}$$

Therefore, we can find the number of stars the Milky Way forms per year on average

$$\left\langle \frac{\psi}{M}\right\rangle \simeq 10 \text{ stars yr}^{-1} \tag{4.259}$$

Over the lifetime of the Milky Way, $\sim 10^{10}$ yr, we therefore formed about 10^{11} stars, meaning we have

$$\boxed{N_{\rm BHB} \simeq 10^{11}\times f_{\rm BHB} \simeq 3\times 10^{7}} \tag{4.260}$$

To get a better estimate, one would need to use stellar population synthesis models and binary evolution models. Doing so gives results of the same order of magnitude (i.e. Ref. 84).

Observational Constraints

There are a few different tracers of black holes in binaries that can be used

- XRBs (LMXBs/HMXBs): These objects must be actively accreting for us to see them, which means the number of XRBs ($\sim$ 100) is much lower than the predicted number of black holes in binaries.
- Gravitational wave events (i.e. with LIGO): These require a merger between two compact objects. Often we can tell whether they are neutron stars or black holes from the masses, but since both objects need to be compact objects, we are only probing a subset of all black holes in binaries. They also need to have existed long enough to merge.
- Microlensing events: These are sensitive to black holes regardless of whether they are in a binary or not.

70. How are the magnetic fields of neutron stars estimated in (i) X-ray binaries? (ii) radio pulsars?

(i) X-ray binaries

Recall from question §4.Q63 that material from the accretion disk is channeled along the magnetic field lines and onto the poles of the pulsar. This creates an accretion column where resonant scattering of photons by electrons can produce "**cyclotron resonant scattering features**." These are line-like features that show up in absorption (since photons are preferrentially scattered out of our line of sight) at energies corresponding to[85]

$$E = n\hbar\omega = n\frac{\hbar eB}{m_e c} \tag{4.261}$$

There is also a gravitational redshift from the surface of the neutron start that adds a factor of $1/(1+z)$, but the important point is that $E \propto B$, so the magnetic field strength can be directly measured from the strength of these features.

(ii) Radio pulsars

A rotating magnetic dipole produces a power of

$$P = \frac{\mathrm{d}E}{\mathrm{d}t} = -\frac{2}{3}\frac{1}{c^3}|\ddot{\boldsymbol{\mu}}_\perp|^2 \tag{4.262}$$

where the magnetic moment $\boldsymbol{\mu}$ is related to the magnetic field by

$$\mathbf{B}(r,\theta) = \frac{|\boldsymbol{\mu}|}{r^3}(2\cos\theta\hat{\mathbf{r}} + \sin\theta\hat{\boldsymbol{\theta}}) \;\longrightarrow\; B_p = 2\frac{|\boldsymbol{\mu}|}{R^3}\;,\; B = \frac{|\boldsymbol{\mu}|}{R^3} \tag{4.263}$$

where B_p is the polar field at $\theta = 0$ and B is the equatorial field at $\theta = \pi/2$. Now, say that the magnetic moment is inclined relative to the axis of rotation (polar axis) by some angle α, such that $|\boldsymbol{\mu}|\sin\alpha = |\boldsymbol{\mu}_\perp|$. Then, differentiating twice, we find

$$|\ddot{\boldsymbol{\mu}}_\perp| = \sin\alpha\frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{BR^3}{\sin\theta} = BR^3\sin\alpha\csc\theta(\cot^2\theta + \csc^2\theta)\dot{\theta}^2\Big|_{\theta=\pi/2} = BR^3\Omega^2\sin\alpha \tag{4.264}$$

where $\Omega = \dot{\theta}$. This makes our power

$$\dot{E} = -\frac{2}{3c^3}B^2R^6\Omega^4\sin^2\alpha \tag{4.265}$$

If the source of this energy is purely from the rotation of the pulsar, then we also have

$$\dot{E} = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(I\Omega^2) = I\Omega\dot{\Omega} \tag{4.266}$$

Replacing $\Omega = 2\pi/P$ and $\dot{\Omega} = -2\pi\dot{P}/P^2$ and solving for B gives[22]

$$B = \sqrt{\frac{3c^3 I}{8\pi^2 R^6\sin^2\alpha}}\sqrt{P\dot{P}} \;\longrightarrow\; \boxed{B \propto \sqrt{P\dot{P}}} \tag{4.267}$$

We see that B is related to the product $\sqrt{P\dot{P}}$. This is the primary way that we can estimate the

magnetic fields of these objects, but it comes with a few caveats:

- We assume a dipolar magnetic field, which is likely a simplification of the real field
- We have the ignored the possibility that the field itself evolves over time
- We need to have some idea of the inclination angle α

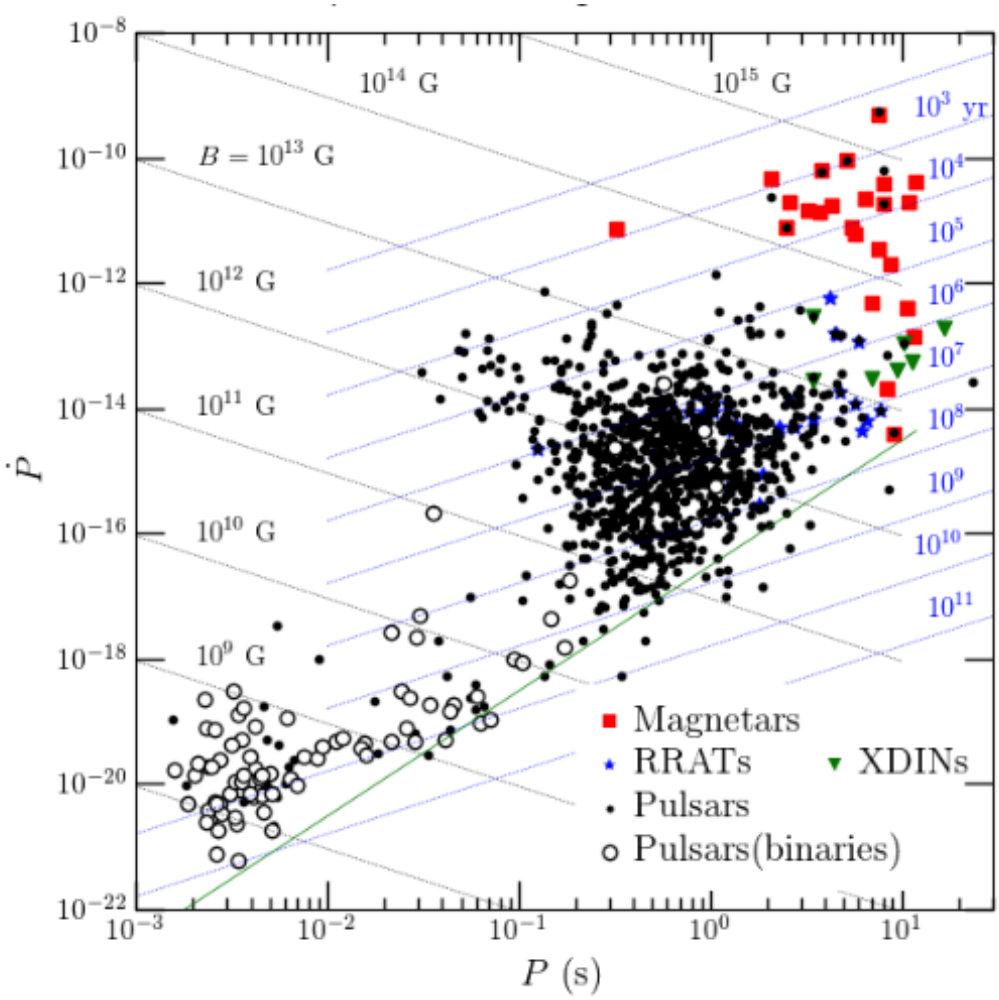


Figure 4.34: The P-$\dot{P}$ diagram for pulsars.

Also, notice that if we assume the field is constant, we can get a relationship between Ω and $\dot{\Omega}$:

$$\dot{\Omega} = -k\Omega^n \tag{4.268}$$

In this case $n = 3$, but in the general case n is called the *braking index*. We can write n in terms of derivatives of Ω like so[22]:

$$\ddot{\Omega} = -kn\Omega^{n-1}\dot{\Omega} = -kn\left(\frac{\Omega^n}{\Omega}\right)\dot{\Omega} = n\frac{\dot{\Omega}^2}{\Omega} \quad \longrightarrow \quad \boxed{n = \frac{\ddot{\Omega}\Omega}{\dot{\Omega}^2}} \tag{4.269}$$

The braking index is important when considering the ***P*-$\dot{P}$ diagram** for pulsars, as it gives the slope for lines of constant magnetic field ($\dot{P} \propto P^{2-n}$). In our dipole case, $n = 3$ so lines of constant B have a slope of -1.

The characteristic age of a pulsar can be estimated using the spin-down timescale[22]:

$$\left|\frac{P}{\dot{P}}\right|_{\text{final}} \simeq \tau \frac{\mathrm{d}}{\mathrm{d}t}\left|\frac{P}{\dot{P}}\right|_{\text{initial}} \simeq \tau(n-1) \longrightarrow \boxed{\tau = \frac{P}{2\dot{P}}} \tag{4.270}$$

Therefore, lines of constant age τ will have a slope of $+1$ on the P-$\dot{P}$ diagram.

The P-$\dot{P}$ diagram is shown in Figure 4.34. Notice the lines of slope -1 (constant B) and $+1$ (constant τ). There is also a solid green line called the *death line*. It is thought that below this line, the rotational speeds are not fast enough to power magnetic fields enough to produce observable radiation. Magnetars tend to be clustered on the top right, having strong magnetic fields and long periods—this forces them to have short lifetimes. Binary pulsars lie at the bottom left since accretion spins them up and decreases their periods (this is where millisecond pulsars lie). Normal pulsars lie in the middle between these two extremes[86]. Typical values for magnetic fields and ages for these populations are, as read from the plot:

	P	$\dot{P}$	B	τ
Magnetars	10 s	10^{-10}	10^{15} G	10 kyr
Pulsars	1 s	10^{-14}	10^{12} G	Myr–Gyr
Millisecond pulsars	10 ms	10^{-20}	10^{9} G	10 Gyr

71. Sketch the frequency distribution of the emission of a typical QSO from radio to X-ray frequencies.

The spectral energy distribution (SED) of QSOs is complicated and involves many physical processes.

- There is thermal emission from the accretion disk, which produces mostly optical/UV photons.
- Some of the photons from the disk are inverse-Compton scattered to soft X-ray energies by the hot corona above the disk.
- Other photons from the disk can be absorbed by dust in the torus and re-emitted in the IR.
- X-ray photons generated in the corona can be reflected off the accretion disk, producing a reflection component as well.
- Radio synchrotron (non-thermal) emission can vary depending on the radio-loudness of the AGN. This depends on if the AGN has a jet that is providing mechanical feedback into the surrounding medium.
- IGM absorption (see §8.Q142) makes EUV–soft X-ray emission largely unobservable.
- The main difference between type I and type II SEDs is the obscuration of the accretion disk component in the optical/UV.
- X-rays also have a penetrating power that makes them less susceptible to obscuration, and IR is similarly much less susceptible to dust obscuration, so these components are similar in type I and type II SEDs.

A model of a type I AGN SED is shown in Figure 4.35, and sketches of type I and type II AGN SEDs are shown in Figures 4.36 and 4.37.

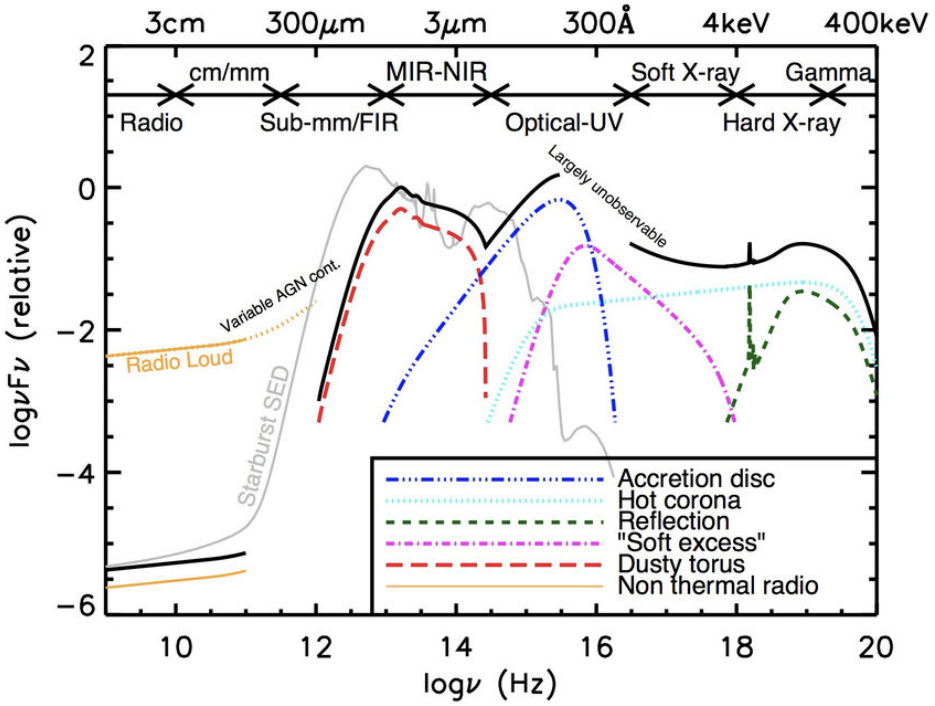


Figure 4.35: The SED of a type I AGN, with the different emission components labeled and color-coded[87]. Note that this is a log-log plot: the abscissa is $\log_{10}(\nu)$ and the ordinate is $\log_{10}(\nu F_\nu)$.

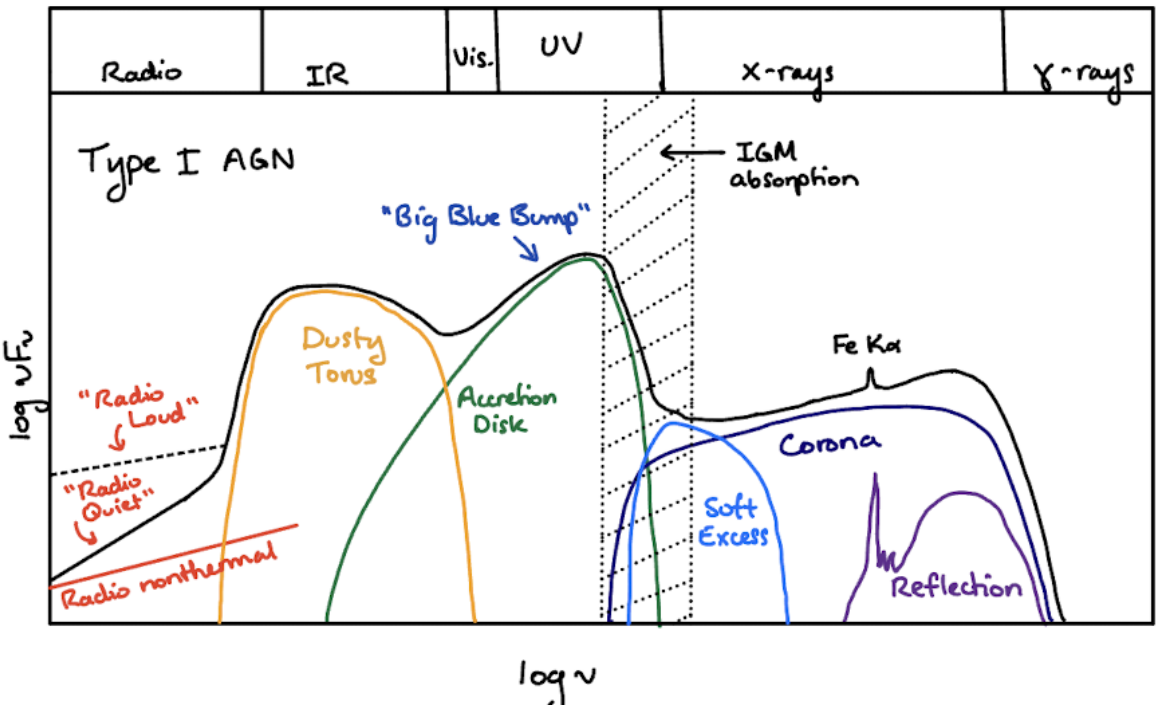


Figure 4.36: A sketch of the SED of a type I AGN, with the different emission components labeled and color-coded[88,89].

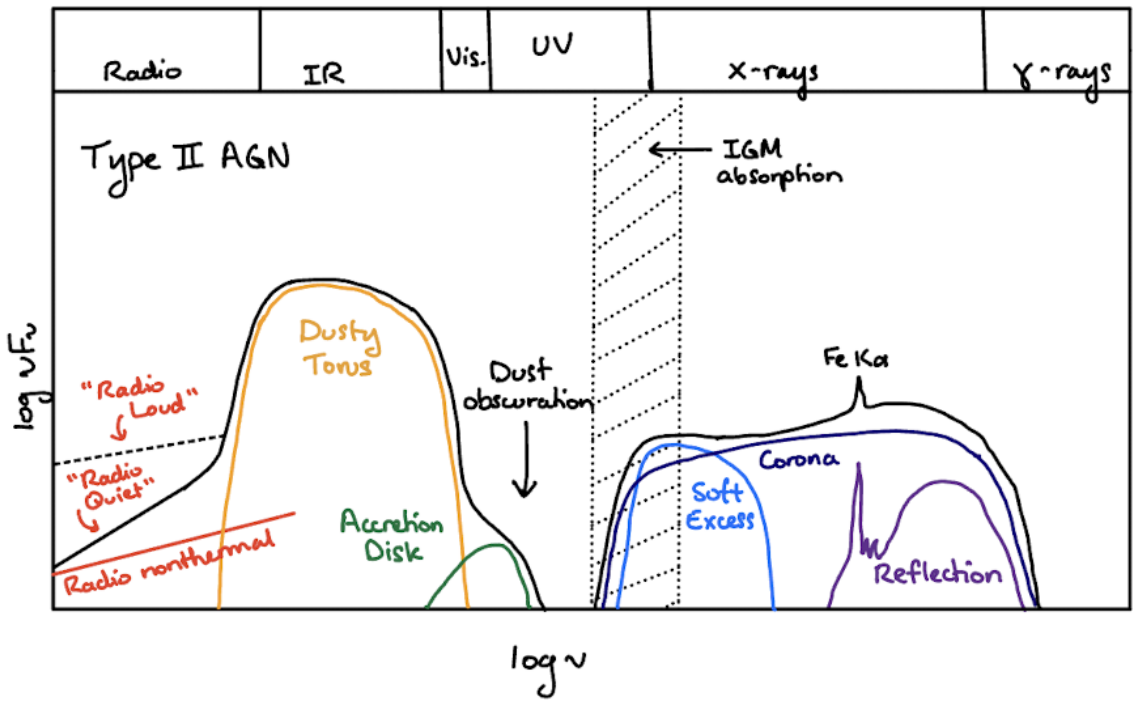


Figure 4.37: A sketch of the SED of a type II AGN, with the different emission components labeled and color-coded[88,89].

5 ISM/Emission

72. Discuss the various "phases" of gas in the interstellar medium. Are these phases in pressure equilibrium?

Phase	T (K)	n_H (cm^{-3})	Description
Coronal gas (HIM) $f_V \approx 0.5$? $\langle n_H \rangle f_V \approx 0.002$ cm^{-3} Fills most of the volume of the halo	$\gtrsim 10^{5.5}$	~ 0.004	Heating by shocks Collisionally ionized Cooling by: •Adiabatic expansion •X-ray emission Observed by: •UV and X-ray emission •Radio sychrotron emission
H II gas $f_V \approx 0.1$ $\langle n_H \rangle f_V \approx 0.02$ cm^{-3} Contains: •**Dense H II regions** •**Diffuse WIM**	10^4	0.3–10^4	Heating by photoelectrons from H, He Photoionized Cooling by: •Optical line emission •Free-free emission •Fine-structure line emission Observed by: •Optical line emission (H recombination) •Thermal radio continuum
Warm H I (WNM) $f_V \approx 0.4$ $n_H f_V \approx 0.2$ cm^{-3}	~ 5000	0.6	Heating by photoelectrons from dust Ionized by starlight, cosmic rays Cooling by: •Optical line emission •Fine-structure line emission Observed by: •H I 21 cm emission, absorption •Optical, UV absorption lines
Cool H I (CNM) $f_V \approx 0.01$ $n_H f_V \approx 0.3$ cm^{-3}	~ 100	30	Heating by photoelectrons from dust Ionized by starlight, cosmic rays Cooling by: •Fine-structure line emission Observed by: •H I 21 cm emission, absorption •Optical, UV absorption lines
Diffuse H_2 $f_V \approx 0.001$ $n_H f_V \approx 0.1$ cm^{-3}	~ 50	~ 100	Heating by photoelectrons from dust Ionized by starlight, cosmic rays Cooling by: •Fine-structure line emission Observed by: •H I 21 cm emission, absorption •CO 2.6 mm emission •Optical, UV absorption lines
Dense H_2 (molecular clouds) $f_V \approx 10^{-4}$ $\langle n_H \rangle f_V \approx 0.2$ cm^{-3}	10–50	10^3–10^6	Heating by photoelectrons from dust Ionized and heating by cosmic rays Self-gravitating: $p > p$(ambient ISM) Cooling by: •CO line emission •C I fine-structure line emission Observed by: •CO 2.6 mm emission •Dust FIR emission

Notes:

- Most of this table is copied from Ref. 64.
- f_V = volume filling factor (fraction of the total disk volume contained in this phase)
- Typical density of the Earth's atmosphere is $\rho \sim 10^{-3}$ g cm^{-3}, corresponding to $n \sim 10^{21}$ cm^{-3}.
- All phases are in **rough pressure equilibrium** with each other which keeps them distinct from each other. See a phase diagram in Figure 5.1.
- They are **not** in **thermodynamic equilibrium**, but are able to maintain this state by constantly receiving energy in the form of UV radiation from stars and gaseous outflows from supernovae and losing energy through radiative cooling into the IGM.

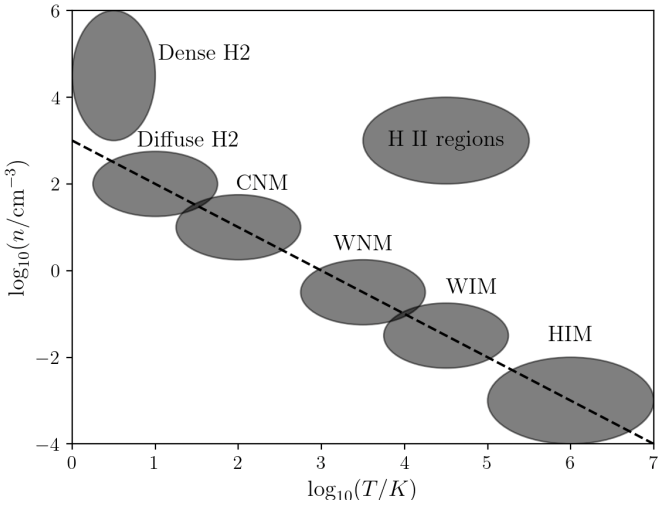


Figure 5.1: A phase diagram of the ISM showing the rough temperatures and densities of each phase. The dashed line shows a line of $n \propto T^{-1}$, meaning pressure is constant along this line.

Elemental composition:

- Majority H and He from the Big Bang
- A small fraction (∼1% by mass) of metals (C to U) reprocessed by stars
- The abundances of metals is the highest at the galactic center and decreases outwards, reaching about 50% at the distance of the Sun.

Structure of the ISM (i.e. how is it spatially distributed)[65]:

- In general, structure is highly uncertain and very poorly constrained. Why?
  - Determining distances is difficult for interstellar material.
  - Emission from the ISM can be absorbed/extincted by closer ISM.
  - High resolution required to resolve small structures in the ISM.
  - Numerous complicated physical processes are involved, including compressible turbulence.
- McKee & Ostriker (1977)[90] developed a detailed model of the ISM which relies on supernovae feedback. In the model, they consider the CNM, WNM, WIM, and HIM to have the same pressure ($P/k \approx 3600$ K cm^{-3}). The ISM consists mostly of the HIM (by volume), with embedded and disconnected clouds of WIM/WNM/CNM. See a demonstration in Figure 5.2.

- Problems with the McKee & Ostriker model:
  - Most real clouds appear as sheets or filaments rather than the spherical ones the model assumes
  - Observed H I distribution is smoother than the model predicts
  - Seriously underestimates the amount of WNM
  - Effects of magnetic fields and cosmic rays are neglected
  - Predicts more SNe shells than are observed
  - Completely ignores molecular clouds

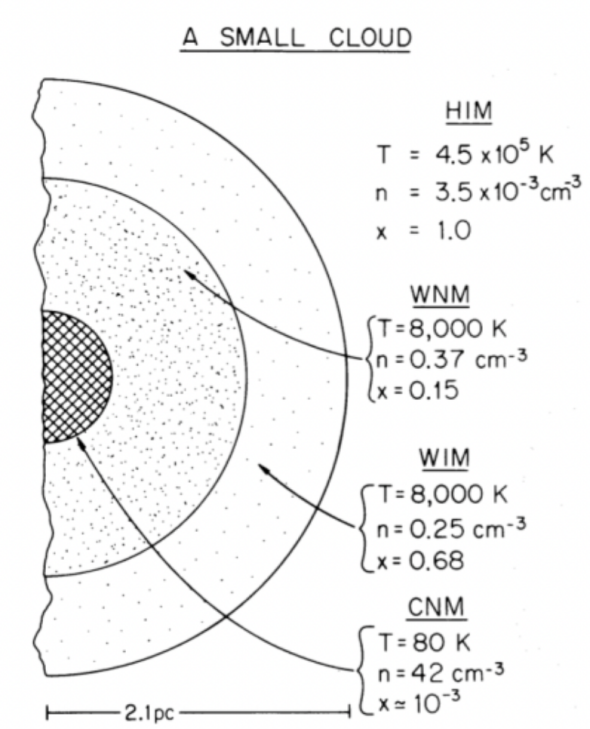


Figure 5.2: How the ISM is distributed locally around a molecular cloud

Other media, in increasing order of spatial extent:

- **The Interstellar Medium (ISM)**: The gas and dust within a galaxy, located cospatially with its stars (hence, inter*stellar*).
- **The Circumgalactic Medium (CGM)**: The gas and dust in the immediate vicinity of a galaxy that is still gravitationally bound to it, but exists on a larger spatial scale than its stars.
- **The Intracluster Medium (ICM)**: The gas and dust that is gravitationally bound to a galaxy cluster, but not to any particular galaxy in the cluster. It exists between the cluster's constituent galaxies.
- **The Intergalactic Medium (IGM)**: This is a more general term for the gas and dust between galaxies in any environment. It is not gravitationally bound to anything. The ICM is a subset of the IGM that pertains to galaxies within clusters, but the IGM also exists outside of clusters, along filaments.

73. What is an H II region? Estimate how the Strömgren radius scales with the luminosity of the ionizing source and with the ambient density.

An **H II region** refers to the photoionized gas surrounding a hot, luminous star. They have gas temperatures from 7,000–15,000 K and densities of 10^2–10^4 cm^{-3}.

We can picture the H II region as an idealized, fully ionized, spherical region of uniform density (known as a **Strömgren sphere**) made up of pure Hydrogen. Then, the **Strömgren radius** is the point at which photoionization from the central star is balanced by hydrogen recombination. Say Q_0 is the rate at which hydrogen-ionizing photons (with $h\nu > 13.6$ eV) are emitted (thus it is the ionization rate). Then, if we have a number density n_H of hydrogen ions and n_e of electrons, the recombination rate per unit volume is $\alpha_B n_H n_e$, where α_B is the case B recombination coefficient*.

We can equate these two rates, multiplying by the volume to get the total recombination rate over the full H II region, to get

$$Q_0 = \frac{4}{3}\pi R_S^3 \alpha_B n_H n_e \tag{5.1}$$

where R_S is the Strömgren radius[64]. For a fully ionized hydrogen gas, $n_H = n_e$, and $\alpha_B \approx 2.56 \times 10^{-13} T_4^{-0.83}\,\mathrm{cm^3\,s^{-1}}$. Solving for R_S:

$$\boxed{R_S = \left(\frac{3Q_0}{4\pi n_H^2 \alpha_B}\right)^{1/3}} \tag{5.2}$$

The luminosity of the ionizing source is directly proportional to Q_0, so R_S should scale like[64,22]

$$R_S \propto L^{1/3} n_H^{-2/3} \tag{5.3}$$

*__Recombination Coefficients__

The recombination coefficient, in general, is the average of the cross section for recombination and the relative velocities of the particles:

$$\alpha = \langle \sigma v \rangle \tag{5.4}$$

Case A recombination occurs when the gas is optically thin to ionizing radiation, so every ionizing photon emitted during the recombination process escapes. Case A recombination is an excellent approximation for low density, high temperature ($T \gtrsim 10^6$ K) regions where H is collisionally ionized and shock-heated, as ionizing photons emitted during recombination are likely to escape the nebula before being reabsorbed.

Case B recombination, in contrast, occurs when the gas is optically thick to radiation above 13.6 eV, so that the ionizing photons emitted during the recombination are immediately reabsorbed, creating another ion and free electron by photoionization. In this case, recombinations to the ground state ($n = 1$) do not reduce the ionization of the gas—only recombinations to $n \geqslant 2$ will reduce the ionization. Case B recombination is an excellent approximation for high density nebulae like H II regions.

74. Explain what bremsstrahlung radiation is. Draw a typical bremsstrahlung spectrum. From what kinds of astrophysical objects is such radiation observed?

TL;DR Bremsstrahlung radiation is **free-free radiation**, that is, it originates from electrons and ions in a thermal plasma scattering off of each other due to electromagnetic forces ("bremsstrahlung" is German for "braking"). The electrons, which have much smaller masses than the ions, will experience larger accelerations and thus will dominate the free-free emission. The emission is a continuum extending from the radio up to frequencies that correspond to the thermal energy kT of the plasma. This gives an emissivity which has a proportionality of $\varepsilon_\nu \propto n_i n_e T^{-1/2} \exp(-h\nu/kT)$, meaning ε_ν is relatively constant up to a cutoff frequency where $h\nu \sim kT$ where it drops off steeply[64].

The following derivation is probably not necessary to know, but it is interesting and shows where the proportionality comes from.

To see what kind of spectrum we will observe from bremsstrahlung emission, we can start with the spectrum that we expect to see from a **single** scattering event with an electron at speed v passing by an ion with charge Ze at impact parameter b:

$$\frac{\mathrm{d}E}{\mathrm{d}\omega} = \begin{cases} \dfrac{8Z^2e^6}{3\pi m_e^2 c^3 b^2 v^2} & \omega\tau \ll 1 \\ 0 & \omega\tau \gg 1 \end{cases} \tag{5.5}$$

This is derived in §5.Q86. Now we want to consider a plasma that will contain many of these scattering events. Let the number density of electrons be n_e and the number density of ions be n_i. Then the flux of electrons onto one ion is $n_e v$ (electrons per unit area per unit time), and the number of scatterings per unit time within a thin shell of width $\mathrm{d}b$ from a single ion is $n_e v 2\pi b \mathrm{d}b$. Then, multiplying by n_i gives us the number of scatterings from *all* ions per unit volume per unit time. Then we just multiply this by the energy per scattering event per frequency, $\mathrm{d}E/\mathrm{d}\omega$, and integrate over $\mathrm{d}b$ to get:

$$\frac{\mathrm{d}E}{\mathrm{d}\omega\,\mathrm{d}V\,\mathrm{d}t} = 2\pi n_i n_e v \int_{b_{\min}}^{b_{\max}} \mathrm{d}b\, b \frac{\mathrm{d}E}{\mathrm{d}\omega} = \frac{16e^6}{3m_e^2c^3v} Z^2 n_i n_e \int_{b_{\min}}^{b_{\max}} \frac{\mathrm{d}b}{b} \tag{5.6}$$

$$= \frac{16e^6}{3m_e^2c^3v} Z^2 n_i n_e \ln\left(\frac{b_{\max}}{b_{\min}}\right) \tag{5.7}$$

where in the second step we took the asymptotic behavior of $\mathrm{d}E/\mathrm{d}\omega$ in the limit $\omega\tau \ll 1$, which can be used as a good approximation for the full result. We define the *Gaunt factor* g_{ff} as

$$g_{\mathrm{ff}}(v,\omega) = \frac{\sqrt{3}}{\pi} \ln\left(\frac{b_{\max}}{b_{\min}}\right) \tag{5.8}$$

Such that the result can be written in the form

$$\frac{\mathrm{d}E}{\mathrm{d}\omega\,\mathrm{d}V\,\mathrm{d}t} = \frac{16\pi e^6}{3\sqrt{3}m_e^2c^3v} Z^2 n_i n_e g_{\mathrm{ff}}(v,\omega) \tag{5.9}$$

In general, the Gaunt factor depends on the energy of the electron and the frequency of the emission. Now we have a result for the spectrum if all electrons moved at a single speed v. For **thermal bremsstrahlung radiation**, we consider a thermal distribution of speeds that should be typical of a

plasma. In particular, we want to use a Maxwell-Boltzmann speed distribution

$$f(v)\mathrm{d}^3v \propto \exp\left(-\frac{E}{kT}\right)\mathrm{d}^3v = \exp\left(-\frac{mv^2}{2kT}\right)\mathrm{d}^3v \tag{5.10}$$

We can assume the velocities are isotropic, i.e. $\mathrm{d}^3v = 4\pi v^2\mathrm{d}v$, to obtain

$$f(v)\mathrm{d}v \propto v^2\exp\left(-\frac{m_e v^2}{2kT}\right)\mathrm{d}v \tag{5.11}$$

We average (5.9) over this speed distribution to get it as a function of the plasma temperature T

$$\frac{\mathrm{d}E(T,\omega)}{\mathrm{d}\omega\,\mathrm{d}V\,\mathrm{d}t} = \frac{\int_{v_{\min}}^{\infty}\mathrm{d}v\,\dfrac{\mathrm{d}E(v,\omega)}{\mathrm{d}\omega\,\mathrm{d}V\,\mathrm{d}t}v^2\exp\left(-m_e v^2/2kT\right)}{\int_0^{\infty}\mathrm{d}v\,v^2\exp(-m_e v^2/2kT)} \tag{5.12}$$

The denominator is

$$\int_0^{\infty}\mathrm{d}v\,v^2\exp\left(-\frac{m_e v^2}{2kT}\right) = \frac{\sqrt{\pi}}{4}\left(\frac{m_e}{2kT}\right)^{-3/2} \tag{5.13}$$

And the numerator is

$$\frac{16\pi e^6}{3\sqrt{3}m_e^2c^3}Z^2n_in_e\langle g_{\mathrm{ff}}\rangle\int_{v_{\min}}^{\infty}\mathrm{d}v\,v\exp\left(-\frac{m_e v^2}{2kT}\right) \tag{5.14}$$

$$=\frac{16\pi e^6}{3\sqrt{3}m_e^2c^3}Z^2n_in_e\langle g_{\mathrm{ff}}\rangle\left[-\frac{kT}{m_e}\exp\left(-\frac{m_e v^2}{2kT}\right)\right]_{v_{\min}}^{\infty} \tag{5.15}$$

$$=\frac{16\pi e^6}{3\sqrt{3}m_e^2c^3}Z^2n_in_e\langle g_{\mathrm{ff}}\rangle\frac{kT}{m_e}\exp\left(-\frac{h\nu}{kT}\right) \tag{5.16}$$

Here, notice we needed to impose a minimum speed $v_{\min}$ because of the quantization of photons. At any speed v, the electron can only create photons with energies below $h\nu \leqslant m_e v^2/2$. So we specify $v_{\min} = \sqrt{2h\nu/m_e}$. Putting it all together, we have

$$\frac{\mathrm{d}E(T,\omega)}{\mathrm{d}\omega\,\mathrm{d}V\,\mathrm{d}t} = \frac{16e^6}{3m_e^2c^3}\left(\frac{2\pi}{3}\right)^{1/2}\langle g_{\mathrm{ff}}\rangle Z^2n_in_e\left(\frac{m_e}{kT}\right)^{1/2}\exp\left(-\frac{h\nu}{kT}\right) \tag{5.17}$$

Finally, with $\mathrm{d}\omega = 2\pi\mathrm{d}\nu$, we have the **emissivity**

$$\boxed{\varepsilon_{\mathrm{ff}}(\nu) \equiv \frac{\mathrm{d}E}{\mathrm{d}\nu\,\mathrm{d}V\,\mathrm{d}t} = \frac{32\pi e^6}{3m_e^2c^3}\left(\frac{2\pi}{3}\right)^{1/2}\langle g_{\mathrm{ff}}\rangle Z^2n_in_e\left(\frac{m_e}{kT}\right)^{1/2}\exp\left(-\frac{h\nu}{kT}\right)} \tag{5.18}$$

And dividing by 4π gives us the **spontaneous emission coefficient**

$$\boxed{j_{\mathrm{ff}}(\nu) \equiv \frac{\mathrm{d}E}{\mathrm{d}\nu\,\mathrm{d}V\,\mathrm{d}t\,\mathrm{d}\Omega} = \frac{8e^6}{3m_e^2c^3}\left(\frac{2\pi}{3}\right)^{1/2}\langle g_{\mathrm{ff}}\rangle Z^2n_in_e\left(\frac{m_e}{kT}\right)^{1/2}\exp\left(-\frac{h\nu}{kT}\right)} \tag{5.19}$$

Looking at $\varepsilon_{\mathrm{ff}}(\nu)$, we notice that for frequencies $h\nu \ll kT$, it is essentially "flat" with respect to ν.

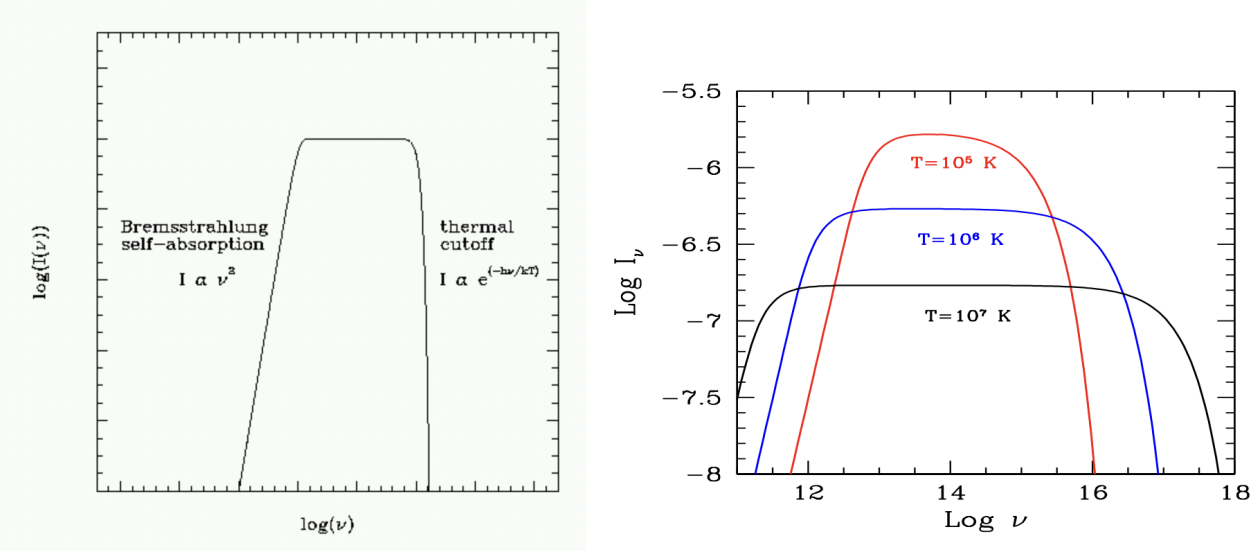


Figure 5.3: Prototypical bremsstrahlung spectra[91]. The right panel shows how the spectrum changes at different temperatures: $\varepsilon_\nu \propto T^{-1/2} e^{-h\nu/kT}$, so as T increases the flat part drops ($T^{-1/2}$) and the drop-off at the high frequency end increases to higher frequencies ($e^{-h\nu/kT}$)[92].

If the medium is optically thin, then the bremsstrahlung spectrum will look like a flat line until a cutoff frequency at $h\nu \sim kT$ where it will quickly drop off. However, in an optically thick medium the absorption becomes important. For a detailed look at the absorption, see §5.Q76. In brief, the absorption coefficient $\alpha_{\rm ff}(\nu) \propto \nu^{-2}$, so there is also a drop-off at the low frequency end of the thermal bremsstrahlung spectrum. See Figure 5.3 for an example of what this might look like.

Also note that the cause of the cutoff at high frequencies here is distinct from what we find in §5.Q86 by considering the classical case of monoenergetic electrons. In the latter, we don't get frequencies higher than $\nu \sim v/b$ because the timescale of the interactions between the electron and ion is too long (shorter timescale = longer emitted frequencies since $t = 1/\nu$). However, in the former case, the cutoff is because we cannot create photons with energies this high because the thermal distribution of electron speeds doesn't reach energies this high.

What astrophysical sources exhibit bremsstrahlung?

- Hot ICM in galaxy clusters ($T \gtrsim 10^7$ K; see §7.Q101)
- Coronal gas / HIM in the ISM of galaxies ($T \sim 10^6$ K; see §5.Q72)
- The solar corona ($T \sim 10^6$ K)
- H II regions (optical to radio emission)

For more information, see Refs. 64,93,65,94,95.

75. What is synchrotron radiation? Draw a typical synchrotron spectrum. From what kind of astrophysical objects is such radiation observed? How can the synchrotron spectrum be used to constrain the age of the source?

TL;DR: Synchrotron radiation is non-thermal radiation that arises from the acceleration of relativistic charged particles around a magnetic field. This results in the particles following a characteristic spiral trajectory as seen in Figure 5.4. The spectrum we derive from this radiation is a power law $F(\nu) \propto \nu^{-s}$ where the index s is related to the power law index of the energy distribution of electrons p by $s = (p-1)/2$. At low frequencies, we are in an optically thick regime where self-absorption becomes important, so we have a turnover where the spectrum goes like $\nu^{5/2}$. Ages can be estimated from the observed spectra because more energetic electrons lose energy faster ($\mathrm{d}E/\mathrm{d}t \propto E^2$), so there is a knee at high energies that gets steeper over time.

There is a lot to talk about for synchrotron radiation, so let's break this town by topic. **Note**: much like with the previous question on bremsstrahlung, probably most of the following derivations are unnecessary to know, but they are interesting, and they show where the proportionalities (which you should know) come from. Most of them are taken from Rybicki & Lightman chapter 6[93].

I. The total synchrotron power radiated by an electron:

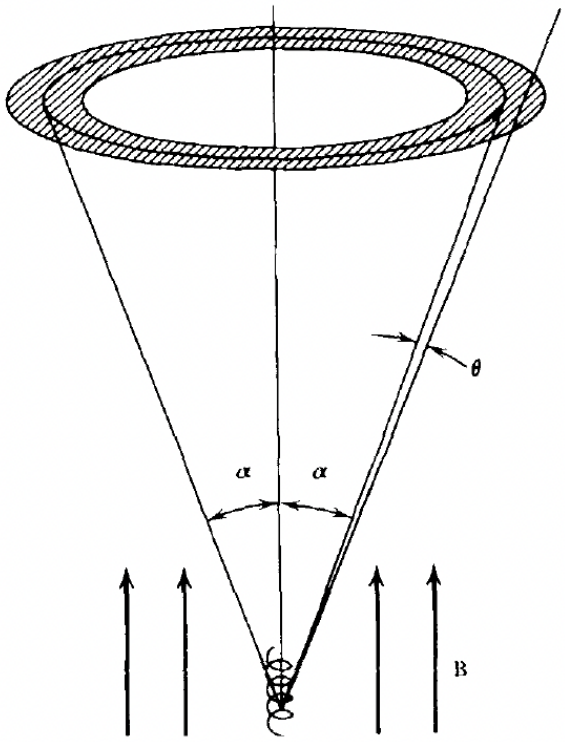


Figure 5.4: Diagram showing how we observe synchrotron radiation. The particle follows the spiral patch through the vertical magnetic field at the bottom. As it spirals, it emits radiation in a small beam that traces out the shaded circle at the top of the figure at a pitch angle α.

To derive what the spectrum looks like, we start with the power that is emitted by a relativistic particle

spiraling in a magnetic field. The relativistic equation of motion is

$$F^\mu = \frac{\mathrm{d}P^\mu}{\mathrm{d}\tau} = \frac{e}{c} F^\mu_{\ \nu} U^\nu \tag{5.20}$$

where F^μ is the 4-force, P^μ is the 4-momentum, U^ν is the 4-velocity, and $F^\mu_{\ \nu}$ is the electromagnetic field tensor. Recall the definitions from special relativity (SR)

$$F^{\mu\nu} = \begin{pmatrix} 0 & E^x & E^y & E^z \\ -E^x & 0 & B^z & -B^y \\ -E^y & -B^z & 0 & B^x \\ -E^z & B^y & -B^x & 0 \end{pmatrix} \tag{5.21}$$

and

$$U^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\tau} = \gamma(u)(c, \mathbf{v}) \ \ , \ \ P^\mu = mU^\mu \tag{5.22}$$

The $\mu = 0$ component gives

$$m\frac{\mathrm{d}(c\gamma)}{\mathrm{d}\tau} = \frac{e}{c}\gamma \mathbf{E}\cdot\mathbf{v} \ \ \longrightarrow \ \ \frac{\mathrm{d}}{\mathrm{d}t}(\gamma mc^2) = e\mathbf{E}\cdot\mathbf{v} \tag{5.23}$$

And the spatial components give

$$\frac{\mathrm{d}}{\mathrm{d}t}(\gamma m\mathbf{v}) = e\left[\mathbf{E} + \frac{1}{c}\mathbf{v}\times\mathbf{B}\right] \tag{5.24}$$

In the case of synchrotron radiation, $\mathbf{E} = 0$. So we have

$$\dot{\gamma} = 0 \ \ , \ \ m\gamma\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = \frac{e}{c}\mathbf{v}\times\mathbf{B} \tag{5.25}$$

Clearly, the component of $\mathbf{v}$ parallel to $\mathbf{B}$ will cause the cross product to vanish, so we can split this up to obtain

$$\frac{\mathrm{d}v_\parallel}{\mathrm{d}t} = 0 \ \ , \ \ \frac{\mathrm{d}\mathbf{v}_\perp}{\mathrm{d}t} = \frac{e}{\gamma mc}\mathbf{v}_\perp\times\mathbf{B} \tag{5.26}$$

Notice that this is the same as the classical result for cyclotron radiation if we just substitute $m \to \gamma m$. The cross product $\mathbf{v}_\perp \times \mathbf{B}$ will be $\perp$ to $\mathbf{v}_\perp$, so clearly this describes circular motion. The frequency is

$$\omega_B = \frac{v_\perp}{r} \ \ \text{and} \ \ a_\perp = \frac{v_\perp^2}{r} \ \ \text{so} \ \ \omega_B = \frac{a_\perp}{v_\perp} = \frac{eB}{\gamma mc} \tag{5.27}$$

The power emitted by a particle is Lorentz invariant and can be computed with the Larmor formula, which in Lorentz covariant form reads

$$P = \frac{2}{3}\frac{e^2}{c^3} a_\mu a^\mu \tag{5.28}$$

Thus we can choose any frame we want to measure the acceleration in. If S' is the instantaneous rest frame of the particle, we can relate this to the acceleration in the observed frame S by

$$a'_\parallel = \gamma^3 a_\parallel \ \ , \ \ a'_\perp = \gamma^2 a_\perp \tag{5.29}$$

Such that

$$P = \frac{2}{3}\frac{e^2}{c^3}\gamma^4(a_\perp^2 + \gamma^2 a_\parallel^2) \tag{5.30}$$

Using (5.26), we find

$$P = \frac{2}{3}\frac{e^4\gamma^2}{m^2c^5}v_\perp^2 B^2 = \frac{2}{3}r_e^2\gamma^2 c\beta_\perp^2 B^2 \tag{5.31}$$

where $r_e = e^2/mc^2$ is the classical electron radius and $\beta_\perp = v_\perp/c$. Let's assume an isotropic distribution of velocities. Then, the average of $\beta_\perp^2$ over pitch angle α is

$$\langle\beta_\perp^2\rangle = \frac{\beta^2}{4\pi}\int \mathrm{d}\Omega \sin^2\alpha = \frac{\beta^2}{2}\int_0^\pi \mathrm{d}\alpha \sin^3\alpha = \frac{2}{3}\beta^2 \tag{5.32}$$

Finally, we have

$$\boxed{P = \frac{4}{3}\sigma_T c\beta^2\gamma^2 u_B} \tag{5.33}$$

where $\sigma_T = 8\pi r_e^2/3$ is the Thomson cross-section and $u_B = B^2/8\pi$ is the magnetic energy density. This is the **average power** radiated by a charged particle moving in a **uniform magnetic field**.

II. The relativistic beaming effect:

In general, the power from (5.33) will be radiated in all directions, but for an electron moving relativistically, the **relativistic beaming effect** causes all of this radiation to look like it's coming from a small cone in the direction the particle travels within some observer frame. As the electron gyrates, this beam circles around the shaded region in Figure 5.4, passing our line of sight once every rotation and thus creating short pulses with a small Δt, meaning the spectrum we see from these pulses is necessarily spread out in frequency space (i.e. Heisenberg).

Recall the form of the Lorentz transformation:

$$\mathrm{d}x = \gamma(\mathrm{d}x' + \beta c\mathrm{d}t') \;\;,\;\; c\mathrm{d}t = \gamma(c\mathrm{d}t' + \beta\mathrm{d}x') \;\;,\;\; \mathrm{d}y = \mathrm{d}y' \;\;,\;\; \mathrm{d}z = \mathrm{d}z' \tag{5.34}$$

From which we can derive the velocity addition formulae:

$$u_x = \frac{u_x' + v}{1 + vu_x'/c^2} \;\;,\;\; u_y = \frac{u_y'}{\gamma(1 + vu_x'/c^2)} \tag{5.35}$$

Say the electron emits a photon at an angle θ' in its rest frame and θ in the observed frame. Then we can relate these angles by

$$\tan\theta = \frac{u_y}{u_x} = \frac{u_y'}{\gamma(u_x' + v)} = \frac{u'\sin\theta'}{\gamma(u'\cos\theta' + v)} \tag{5.36}$$

For a photon, $u' = c$, and the maximum angle it could be emitted at is $\theta' = \pi/2$, so

$$\tan\theta = \frac{\sin\theta'}{\gamma(\cos\theta' + \beta)} \longrightarrow \boxed{\theta = \frac{1}{\gamma}} \tag{5.37}$$

Thus, all of the radiation from the electron will be emitted in the direction of propagataion within a cone that sweeps out an angle of $\Delta\theta = 2/\gamma$. See Figure 5.5 for a demonstration of this effect visually.

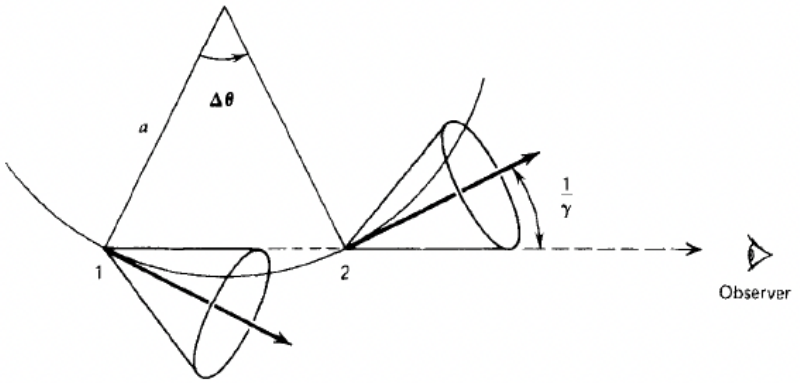


Figure 5.5: The relativistic beaming effect[93].

How long is the pulse duration $\Delta t'$ for each gyration? It will just be the length of the arc over which the cone of radiation falls along our line of sight (from point 1 to 2 in the Figure), over the speed of the electron. If the radius of curvature is r (labeled a in the Figure), then:

$$\Delta t' = \frac{r\Delta\theta}{v} = \frac{2r}{\gamma v} \tag{5.38}$$

We can get r from the frequency of gyration:

$$\mathbf{r} \times \boldsymbol{\omega} = \mathbf{v} \quad \longrightarrow \quad r \sin\alpha\omega_B = v \quad \longrightarrow \quad r = \frac{v}{\omega_B \sin\alpha} \tag{5.39}$$

We must also consider the arrival time of the radiation from points 1 and 2, which will take a difference of $r\Delta\theta/c$, so our time difference in the observed frame is

$$\Delta t = \Delta t'\left(1 - \frac{v}{c}\right) = \frac{2}{\gamma\omega_B \sin\alpha}\left(1 - \frac{v}{c}\right) \approx \frac{1}{\gamma^3\omega_B \sin\alpha} \tag{5.40}$$

where in the last step we approximated $1 - v/c \approx 1/2\gamma^2$. The width of the pulses is smaller than the gyration period by a factor of γ^3. Recall that $E = \gamma m_e c^2$. Thus, for each electron, the primary frequency it will emit at will be

$$\boxed{\nu \approx \gamma^3 \omega_B \sin\alpha \propto \gamma^2 B \propto E^2 B} \tag{5.41}$$

III. The power emitted per unit frequency for a distribution of electrons:

Now let's generalize to considering a distribution of electrons with energies given by a power law with index p:

$$N(E)\mathrm{d}E \propto E^{-p}\mathrm{d}E \tag{5.42}$$

Using equation (5.41), we have

$$\mathrm{d}\nu \propto 2E\mathrm{d}E \quad \longrightarrow \quad \mathrm{d}E \propto \nu^{-1/2}\mathrm{d}\nu \tag{5.43}$$

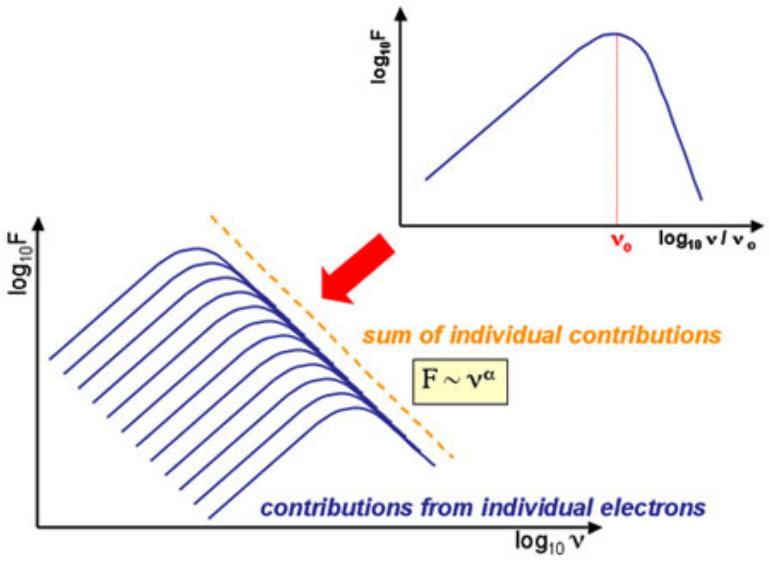


Figure 5.6: A typical power-law synchrotron spectrum[96].

Our rate of radiation should be proportional to both the number of electrons and the rate of emission per electron. Using (5.33) we see that $\mathrm{d}E/\mathrm{d}t \propto \gamma^2 \propto E^2$, so multiplying by the number of electrons $N(E)\mathrm{d}E$ and using (5.41) and (5.43) to put everything in terms of ν gives

$$F(\nu)\mathrm{d}\nu \propto E^2 E^{-p}\mathrm{d}E \longrightarrow \boxed{F(\nu)\mathrm{d}\nu \propto \nu^{-(p-1)/2}\mathrm{d}\nu} \tag{5.44}$$

In other words, the spectrum we see is a **power law** of the form $F(\nu) \propto \nu^{-s}$ with a **spectral index** $s = (p-1)/2$. Typically, $p \sim 2.6$ and $s \sim 0.8$.

IV. Synchrotron self-absorption:

Synchrotron **self-absorption** is also important in the optically thick regime, which happens at lower frequencies. Here, we cannot just use Kirchhoff's law since the electrons are not a thermal distribution but rather a power-law distribution. Thus, we have to start from the generic formula for the emissivity and the absorption coefficient (see §5.Q88), summing over all possible energies E_1 and E_2 that differ by $h\nu$:

$$j_\nu = \frac{h\nu}{4\pi}\sum_{E_1} n(E_2)A_{21}\phi_\nu \tag{5.45}$$

$$\alpha_\nu = \frac{h\nu}{c}\sum_{E_1}\sum_{E_2}[n(E_1)B_{12} - n(E_2)B_{21}]\phi_\nu \tag{5.46}$$

For the true absorption term, we have

$$\frac{h\nu}{c}\sum_{E_1}\sum_{E_2} n(E_1)B_{12}\phi_\nu = \frac{g_1}{g_2}\frac{c^2}{8\pi\nu^2}\sum_{E_1}\sum_{E_2} n(E_1)A_{21}\phi_\nu = \frac{c^2}{2h\nu^3}\sum_{E_2}\frac{n(E_1)}{n(E_2)}j_\nu(E_2) \tag{5.47}$$

Where the degeneracies are set to 1 since these are elementary states. Because of this, we have $B_{21} = B_{12}$. We can replace $n(E_1) = n(E_2 - h\nu)$ since the ϕ_ν within j_ν enforces that $E_2 - E_1 = h\nu$.

Putting it all together, we have

$$\alpha_\nu = \frac{c^2}{2h\nu^3} \sum_{E_2} \frac{n(E_2 - h\nu) - n(E_2)}{n(E_2)} j_\nu(E_2) \tag{5.48}$$

Now we take the limit for small $h\nu$, for which we can write the part inside the summation as a derivative and the summation itself becomes an integral. For simplicity, we relabel $n(E_2) \to N(E)$:

$$\alpha_\nu = \frac{c^2}{2h\nu^3} \int \mathrm{d}E\, E^2 \frac{h\nu}{N(E)} \frac{N(E - h\nu)/(E - h\nu)^2 - N(E)/E^2}{h\nu} j_\nu(E) \tag{5.49}$$

$$\alpha_\nu = -\frac{c^2}{2\nu^2} \int \mathrm{d}E\, \frac{E^2}{N(E)} \frac{\partial}{\partial E} \left(\frac{N(E)}{E^2} \right) j_\nu(E) \tag{5.50}$$

Evaluating the derivative,

$$-E^2 \frac{\partial}{\partial E} \left(\frac{N(E)}{E^2} \right) = (p+2) C E^{-(p+1)} = \frac{(p+2)N(E)}{E} \tag{5.51}$$

Finally, we have

$$\alpha_\nu = \frac{(p+2)c^2}{2\nu^2} \int \mathrm{d}E\, \frac{j_\nu(E)}{E} \propto \nu^{-2} E^{-p} \quad \longrightarrow \quad \boxed{\alpha_\nu \propto \nu^{-(p+4)/2}} \tag{5.52}$$

Therefore, the source function has a proportionality $S_\nu = j_\nu / \alpha_\nu \propto \nu^{5/2}$, which is different from the ν^2 from a thermal distribution. Thus, at low frequencies we are in the optically thick region where the spectrum looks like $\nu^{5/2}$, and at high frequencies we are in the optically thin region where it goes like ν^{-s}. See an example in Figure 5.6. Also see Refs. 93,24.

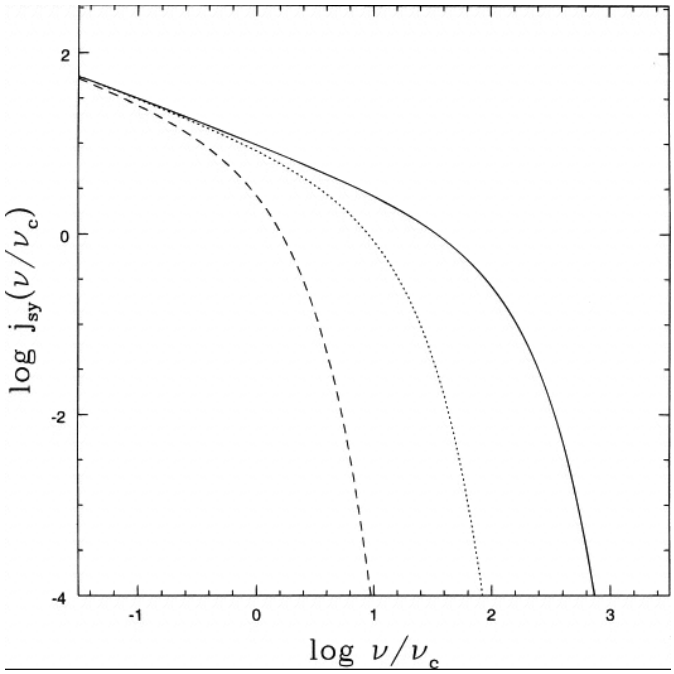


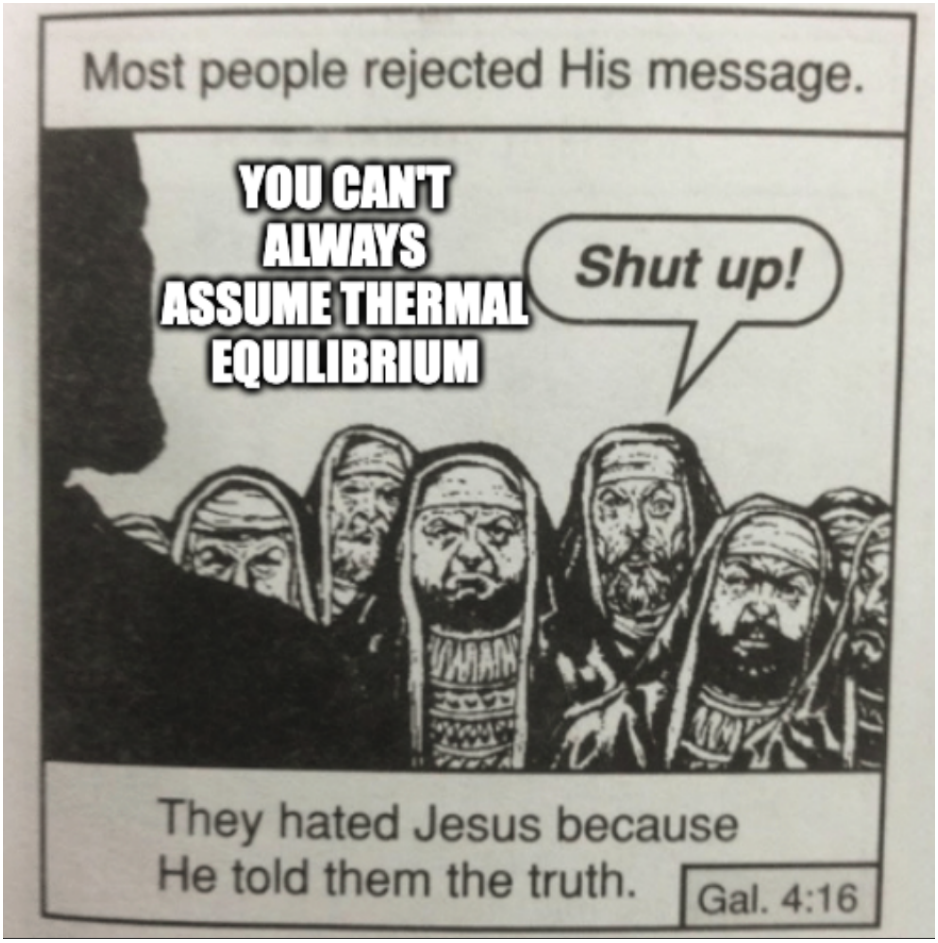
Figure 5.7: A synchrotron spectrum with a distinct knee that can be used to determine the age[97].

V. What astrophysical sources exhibit synchrotron?

- AGN radio jets
- XRB radio jets
- Pulsar wind nebulae
- Galactic ISM
- Sunspots
- GRBs

VI. How can we estimate ages?

Recall that the energy emitted per electron is related to the energy squared, $\mathrm{d}E/\mathrm{d}t \propto E^2$. Therefore, higher energy electrons lose energy faster than lower energy electrons. This causes a break in the power law distribution at high energies, called a "knee". The older a source is, the more prominent this effect becomes, allowing us to estimate the age from the location of the break. See Figure 5.7.



76. What is meant by "free-free" absorption? How is this different from electron scattering?

Free-free absorption is the inverse process to **bremsstrahlung emission** ("free-free emission"; see §5.Q74). This arises when electrons are able to absorb photons by interacting with a nearby ion. Note: an isolated free electron cannot absorb a photon on its own because it cannot conserve both energy and momentum. By having a nearby ion, the photon can distribute its momentum evenly and be absorbed by the electron. We can see why this is the case by considering a hypothetical scenario where a free electron absorbs a photon. The initial electron momentum and energy are p_i and E_i, the final electron momentum and energy are p_f and E_f, and the photon energy and momentum are p_ν and E_ν:

$$p_i + p_\nu = p_f \tag{5.53}$$

$$E_e + E_\nu = E_f \quad \longrightarrow \quad \sqrt{(p_i c)^2 + (m_e c^2)^2} + p_\nu c = \sqrt{(p_f c)^2 + (m_e c^2)^2} \tag{5.54}$$

Squaring the energy equation and plugging in the photon momentum from the momentum equation, we get

$$c^2(p_f - p_i)^2 - c^2(p_f^2 - p_i^2) + 2c(p_f - p_i)\sqrt{(p_i c)^2 + (m_e c^2)^2} = 0 \tag{5.55}$$

This equation has two solutions, both of which produce contradictory results

$$p_f - p_i = 0 \tag{5.56}$$

$$\sqrt{(p_i c)^2 + (m_e c^2)^2} - p_i c = 0 \tag{5.57}$$

The first one implies the photon has no momentum, $p_\nu = 0$, and the second one implies the electron has no rest mass, $m_e = 0$. Therefore, we conclude that an isolated electron cannot absorb a photon while conserving both energy and momentum.

Since the energy levels of bremsstrahlung emission are populated by a thermal distribution (the speeds of the electrons), we can assume the system should be in thermal equilibrium, in which case the emission and absorption of photons should balance each other by satisfying Kirchhoff's law (see §12.11):

$$j_{\text{ff}}(\nu) = \alpha_{\text{ff}}(\nu) B_\nu(T) \tag{5.58}$$

where $B_\nu(T)$ is the Planck function and $j_{\text{ff}}(\nu)$ is the bremsstrahlung emissivity that we derived in §5.Q74. Solving for α_{ff} gives

$$\boxed{\alpha_{\text{ff}}(\nu) = \frac{4e^6}{3m_e^2 hc\,\nu^3}\left(\frac{2\pi}{3}\right)^{1/2}\langle g_{\text{ff}}\rangle Z^2 n_i n_e \left(\frac{m_e}{kT}\right)^{1/2}\left[1 - \exp\left(-\frac{h\nu}{kT}\right)\right]} \tag{5.59}$$

We can see that $\alpha_{\text{ff}}(\nu) \propto \nu^{-3}T^{-1/2}[1 - \exp(-h\nu/kT)]$. In the Rayleigh-Jeans limit ($h\nu \ll kT$), the frequency dependence reduces to ν^{-2}. This is what causes the ν^{-2} drop-off on the low frequency end of the bremsstrahlung spectrum (§5.Q74).

Recall that opacity is just the absorption coefficient times the density: $\kappa_{\text{ff}}(\nu) = \rho\,\alpha_{\text{ff}}(\nu)$. We can define the **Rosseland mean opacity** by doing an average weighted by the derivative of the Planck function

(see motivation in §3.Q30):

$$\frac{1}{\kappa_{\mathrm{R,ff}}} = \frac{\int_0^\infty \mathrm{d}\nu\, \kappa_{\mathrm{ff}}^{-1}(\nu) \dfrac{\partial B_\nu(\nu,T)}{\partial T}}{\int_0^\infty \mathrm{d}\nu\, \dfrac{\partial B_\nu(\nu,T)}{\partial T}} \tag{5.60}$$

First let's look at the denominator, which can be quickly solved by switching the order of the frequency integral and temperature derivative (which is possible because they are independent variables) and using the Stefan-Boltzmann law:

$$\frac{\partial}{\partial T}\int_0^\infty \mathrm{d}\nu B_\nu(\nu,T) = \frac{\partial}{\partial T}\int_0^\infty \mathrm{d}\nu \frac{F_\nu}{\pi} = \frac{\partial}{\partial T}\left(\frac{1}{\pi}\sigma_{\mathrm{SB}}T^4\right) = \frac{4\sigma_{\mathrm{SB}}}{\pi}T^3 \tag{5.61}$$

Now for the numerator, the derivative of the Planck function is

$$\frac{\partial B_\nu(\nu,T)}{\partial T} = \frac{\partial}{\partial T}\left(\frac{2h\nu^3}{c^2}\frac{1}{e^{h\nu/kT}-1}\right) = \frac{2h^2\nu^4}{kT^2c^2}\frac{e^{h\nu/kT}}{(e^{h\nu/kT}-1)^2} \tag{5.62}$$

$$= \frac{2k}{c^2}\left(\frac{kT}{h}\right)^2 x^4 \frac{e^x}{(e^x-1)^2} \tag{5.63}$$

where in the last line I defined $x \equiv h\nu/kT$.

Also writing the free-free opacity in terms of x:

$$\kappa_{\mathrm{ff}}(x) = \frac{4e^6}{3m_e^{3/2}h^{3/2}c}\left(\frac{2\pi}{3}\right)^{1/2}\langle g_{\mathrm{ff}}\rangle Z^2 n_i n_e \rho \left(\frac{kT}{h}\right)^{-7/2} x^{-3}(1-e^{-x}) \tag{5.64}$$

Ignoring irrelevant constants so we can just look at the proportionality, we have (remembering to substitute $\mathrm{d}\nu = \mathrm{d}x(kT/h)$):

$$\frac{1}{\kappa_{\mathrm{R,ff}}} \propto \frac{\rho^{-1}T^{7/2}T^2T}{T^3}\int_0^\infty \mathrm{d}x\, x^7(e^x-1)^{-1} \propto \rho^{-1}T^{7/2} \tag{5.65}$$

This results in a **Kramer's opacity law**:

$$\boxed{\kappa_{\mathrm{R,ff}} \propto \rho T^{-7/2}} \tag{5.66}$$

This shows that the opacity becomes most important at lower temperatures (and quite steeply so), when the particles are (on average) moving slower and thus having stronger interactions. The density proportionality simply shows that particles are more likely to interact when they are more densely packed. Generally, free-free absorption is negligible in the ISM above $\nu \gtrsim 10$ GHz, but it can become important in dense H II regions at $\nu \lesssim 1$ GHz.

Now let's compare this to **electron scattering**. This process is conceptually different. Here, we have a charged particle like an electron that interacts with a photon, but there is not a nearby particle that it can exchange momentum with to allow it to absorb the photon. However, it can still scatter the photon into another direction. An electron feels the oscillating electric field of a photon and starts oscillating in response, redirecting the indcident radiation in another direction (see Figure 5.8).

Contributions to scattering from more massive particles can largely be neglected because their higher

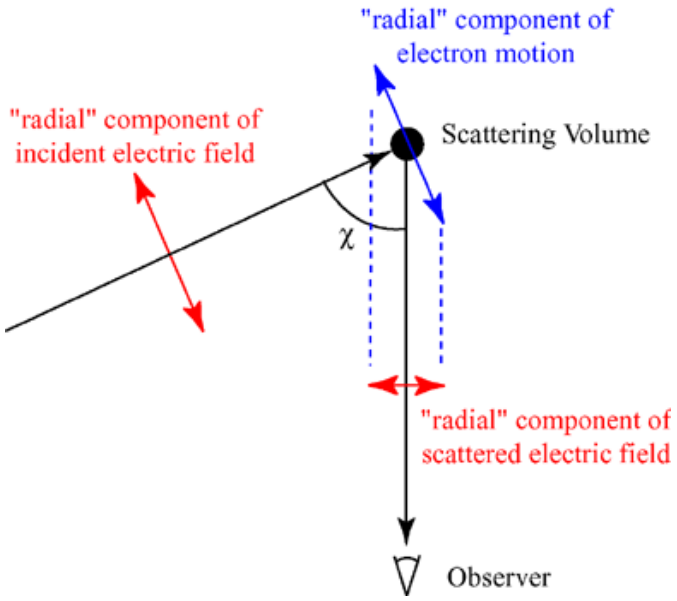


Figure 5.8: Geometry for Thomson scattering

masses cause them to be more resistant to the electric fields. In the classical limit (where the photon energy is much less than the rest energy of the electron, $h\nu \ll m_e c^2$, and the electron itself is also non-relativistic, $v \ll c$) this is called **Thomson scattering**. The opacity is defined as the cross section per unit mass, so in this case we have

$$\kappa_{es} = \frac{n_e \sigma_T}{\rho} \tag{5.67}$$

Where σ_T is the Thomson cross section, which is defined as

$$\boxed{\sigma_T = \frac{8\pi}{3}\left(\frac{e^2}{m_e c^2}\right)^2} \tag{5.68}$$

Notice how this cross-section, and by extension the opacity, is **independent of frequency**, unlike free-free absorption.

If we consider the relativistic case, we get **Compton scattering**. Here, the electron actually recoils by a non-negligible amount. This means that the light loses energy and thus changes frequencies. To account for this, we have to treat the light as individual photons, thus this is a quantum effect. The amount the wavelength changes is given by

$$\lambda' - \lambda = \frac{h}{m_e c}(1 - \cos\theta) \tag{5.69}$$

where λ is the wavelength before scattering, λ' is the wavelength after scattering, and θ is the angle the photon is scattered at. In this case the cross section and opacity **do** depend on the frequency in a complicated way.

77. What is the Saha equation and how is it used in stellar structure calculations?

The **Saha equation** describes the distribution of an atomic species among its various stages of ionization. To derive it, we can start from the Boltzmann law:

$$n_i \propto g_i e^{-E_i/kT} \tag{5.70}$$

That is, the number density of atoms n_i in an excited state i is exponentially smaller than lower energy states by a factor E_i/kT (g_i is the statistical weight). Then, we can compare the number in two states

$$\frac{n_2}{n_1} = \frac{g_2}{g_1} e^{-(E_2-E_1)/kT} \tag{5.71}$$

Let's look at the example of two ions of the same element that differ by one electron. Then, the difference in energies between these two states is $E_{i+1} - E_i = \chi_i + m_e v^2/2$ where χ_i is the ionization potential for state i (the energy needed to ionize i from the ground state). The statistical weight of the lower level is just $g_1 = g_i$, but for the upper level we also have to consider all the possible energy states that the electron might be put into, i.e. $g_2 = g_{i+1} g_e$. Putting this together, we have

$$\frac{\mathrm{d}n_{i+1}}{n_i} = \frac{g_{i+1} g_e}{g_i} \exp\left(- \frac{\chi_i + \frac{1}{2} m_e v^2}{kT} \right) \tag{5.72}$$

Looking at g_e in phase space, we have cells with a volume of h^3, within which each electron has 2 spin states, so our statistical weight becomes

$$g_e = \frac{2}{h^3} \mathrm{d}^3\mathbf{x}\, \mathrm{d}^3\mathbf{p} \tag{5.73}$$

We consider a volume element that contains one electron, and we assume the electrons have an isotropic velocity distribution, so

$$\mathrm{d}^3\mathbf{x} = \frac{1}{n_e} \quad , \quad \mathrm{d}^3\mathbf{p} = 4\pi p^2 \mathrm{d}p = 4\pi m_e^3 v^2 \mathrm{d}v \tag{5.74}$$

Thus

$$\frac{n_e \mathrm{d}n_{i+1}}{n_i} = \frac{8\pi m_e^3}{h^3} \frac{g_{i+1}}{g_i} \exp\left(- \frac{\chi_i + \frac{1}{2} m_e v^2}{kT} \right) v^2 \, \mathrm{d}v \tag{5.75}$$

Making a substitution $x = \sqrt{m_e v^2 / 2kT}$ and integrating

$$\frac{n_e n_{i+1}}{n_i} = \frac{8\pi m_e^3}{h^3} \frac{g_{i+1}}{g_i} e^{-\chi_i/kT} \left(\frac{2kT}{m_e} \right)^{3/2} \int_0^\infty \mathrm{d}x \, x^2 e^{-x^2} \tag{5.76}$$

And the integral is $\sqrt{\pi}/4$. Simplifying, we obtain the **Saha equation**[98,93]:

$$\boxed{\frac{n_e n_{i+1}}{n_i} = \frac{2 g_{i+1}}{g_i} \left(\frac{2\pi m_e kT}{h^2} \right)^{3/2} e^{-\chi_i/kT}} \tag{5.77}$$

We can generalize further by writing in terms of the partition function Z rather than the statistical weights:

$$\frac{n_e n_{i+1}}{n_i} = \frac{2Z_{i+1}}{Z_i}\left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} e^{-\chi_i/kT} \tag{5.78}$$

This is the Saha equation. Note: this is only valid in **thermal equilibrium** and when the **pressure is not too high** (otherwise you start getting pressure ionization and electron degeneracy pressure).

If we are only concerned with the first excited state ($i = 0$), we can make some simplifying assumptions and rewrite this in terms of the ionization fraction $x \equiv n_1/n$ where $n = n_1 + n_0$. We make the assumption that the total charge is neutral ($n_e = n_1$) and that the total particle number n is conserved. Then,

$$\frac{n_1}{n_0} = \frac{xn}{(1-x)n} = \frac{x}{1-x} \quad , \quad n_e = xn \tag{5.79}$$

So we have

$$\boxed{\frac{x^2}{1-x} = \frac{1}{n}\frac{2g_1}{g_0}\left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} e^{-\chi/kT}} \tag{5.80}$$

How is it used in stellar structure calculations?[39]

- The Saha equation tells us how the relative populations of different ionization levels are dependent on the temperature T of the star.
- It shows us how the gas inside a star differs from an ideal gas—each ionization event absorbs energy and changes the composition of the gas (decreasing the chemical potential μ). Therefore, adding heat to the gas will give some energy to increasing the ionization and some energy to increasing the temperature.
- It can be used to calculate the opacities associated with bound-bound and bound-free absorption in different elements (since these depend on the number densities of the different ionization states). For Hydrogen in particular, this explains the differences in strength of the Balmer lines in stars of different temperatures.

78. Explain, from a statistical mechanics point of view, why the Balmer lines are most prominent in A stars with an effective temperature of $\approx 10^4$ K.

There are two different processes going on that are important to consider:

1. What fraction of H atoms are in the neutral state (necessary to produce the Balmer lines)?
2. What fraction of neutral H atoms are in the $n = 2$ excited state?

To answer the first question, we can use the **Saha equation** (see §5.Q77). If we make the simplifying assumption that all neutral H atoms are in the ground state, then the degeneracies $g_1 = 1$ for ionized H and $g_0 = 2$ for neutral H (2 electron spins), so we have

$$\frac{n_{\rm HII}}{n_{\rm HI}} = \frac{1}{n_e}\left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} e^{-\chi/kT} \tag{5.81}$$

We can then solve for the neutral fraction $n_{\rm HI}/n$:

$$\frac{n}{n_{\rm HI}} = 1 + \frac{n_{\rm HII}}{n_{\rm HI}} = 1 + \frac{1}{n_e}\left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} e^{-\chi/kT} \tag{5.82}$$

$$\frac{n_{\rm HI}}{n} = \left[1 + \frac{kT}{P_e}\left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} e^{-\chi/kT}\right]^{-1} \tag{5.83}$$

where I used $P_e = n_e kT$ as the pressure from the electrons. Also recall that the ionization potential $\chi = 13.6$ eV. Plotting this as a function of temperature shows that the neutral fraction quickly transitions from fully neutral to fully ionized at around $T \simeq 10^4$ K (see graph in Figure 5.9).

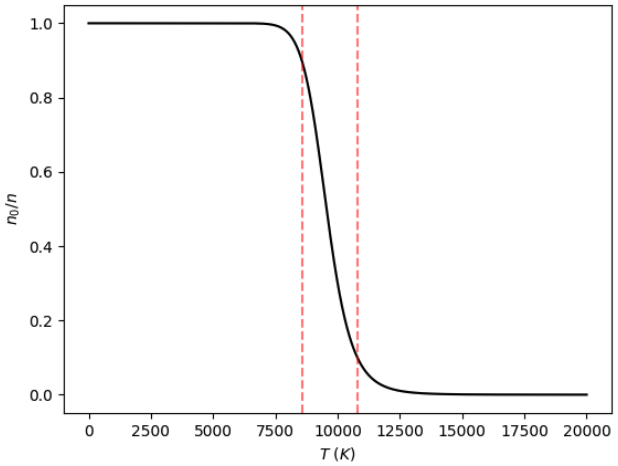


Figure 5.9: The neutral fraction of Hydrogen as a function of temperature.

Now we move on to the second question. To answer this we can use simple Boltzmann statistics. Denoting the $(m-1)^{\rm th}$ excited state with $n_{0,m}$ and the $m = 1$ ground state, as above, with n_0, we have:

$$\frac{n_{0,m}}{n_0} = \frac{g_{0,m}}{g_0} e^{-(E_{0,m}-E_0)/kT} \tag{5.84}$$

Where the energies are $E_m = -13.6\,{\rm eV}/m^2$ and the degeneracies are $g_m = 2m^2$. Then, the fraction

of all H atoms in the $m = 2$ excited state is

$$\frac{n_{0,2}}{n_{\mathrm{HI}}} = \frac{n_{0,2}}{\sum_m n_{0,m}} = \left[\sum_{m=1}^{\infty} \frac{g_{0,m}}{g_{0,2}} e^{-(E_{0,m}-E_{0,2})/kT}\right]^{-1} = \left[\sum_{m=1}^{\infty} \frac{m^2}{4} e^{13.6\,\mathrm{eV}(1/m^2-1/2^2)/kT}\right]^{-1} \tag{5.85}$$

This is well approximated by just taking the $m = 1$ and $m = 2$ terms, which dominate:

$$\frac{n_{0,2}}{n_{\mathrm{HI}}} \approx \frac{1}{1 + n_0/n_{0,2}} = \frac{1}{1 + \frac{1}{4} e^{13.6\,\mathrm{eV}(3/4)/kT}} \tag{5.86}$$

This shows us that a significant fraction of $m = 2$ state atoms do not start appearing until temperatures $\gtrsim 5000$ K, while they continue climbing for higher temperatures (see Figure 5.10).

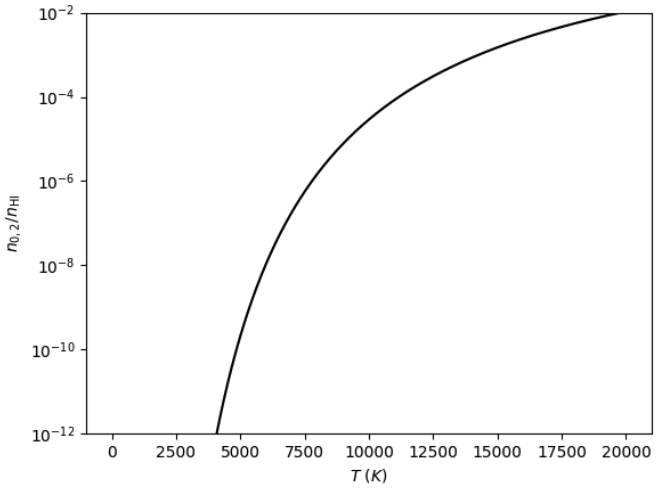


Figure 5.10: The fraction of neutral Hydrogen in the $m = 2$ excited state vs. the ground state and other excited states, as a function of temperature. Note the logarithmic y axis.

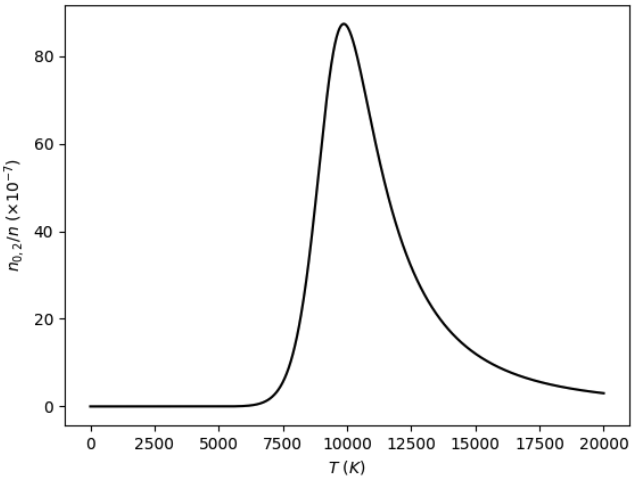


Figure 5.11: The fraction of H atoms in the $m = 2$ excited state compared to *all* H atoms (neutral and ionized).

So, to answer our original question, we need to consider the fraction of H atoms in the $m = 2$ excited state compared to *all* H atoms (neutral and ionized) in the star. What this means is multiplying

(5.86) by (5.83). We can imagine what multiplying these functions will do: the $m = 2$ excited state in neutral Hydrogen doesn't start popping up until higher temperatures, whereas the number of neutral H atoms starts dropping at even higher temperatures due to ionization. So, there should be some sweet spot where we have a modest amount of excitations without too many ionizations. It is at this point that the Balmer lines will be the strongest. It turns out that this happens right around $T \sim 10^4$ K, as seen in Figure 5.11. Analytically:

$$\frac{n_{0,2}}{n} \approx \left\{ \left[1 + \frac{1}{4} e^{13.6\,\text{eV}(3/4)/kT} \right] \times \left[1 + \frac{kT}{P_e} \left(\frac{2\pi m_e kT}{h^2} \right)^{3/2} e^{-13.6\,\text{eV}/kT} \right] \right\}^{-1} \tag{5.87}$$

This is why the Balmer lines are the most prominent in A-type stars, which have effective temperatures around $\sim 10^4$ K[22].

79. Explain how interstellar dust grains can result in linear polarization of transmitted light. How is the direction of polarization related to the average direction of the interstellar magnetic field (as projected on the plane of the sky)? What major discovery in 2014 was derailed by this phenomenon?

Polarization of starlight was noticed to correlate with distance and direction on the sky rather than with stellar properties—thus, it was attributed to the ISM. Further study revealed that the continuum itself is polarized (i.e. the polarization is not confined to particular atomic, ionic, or molecular lines), so the cause has been further constrained to dust grains.

The mechanism where dust grains polarize light is called **dichroic extinction**. First, the dust grains themselves are non-spherical. They are also spinning with some angular momentum vector **J**. The grains tend to become aligned such that their "short" axes are parallel to the local interstellar magnetic field. Because of this alignment, they tend to preferentially absorb incident radiation along the direction of their long axis, i.e. the direction perpendicular to the magnetic field. This causes the outgoing light to become polarized in the direction **parallel** to the local interstellar magnetic field. See Figure 5.12 for a schematic.

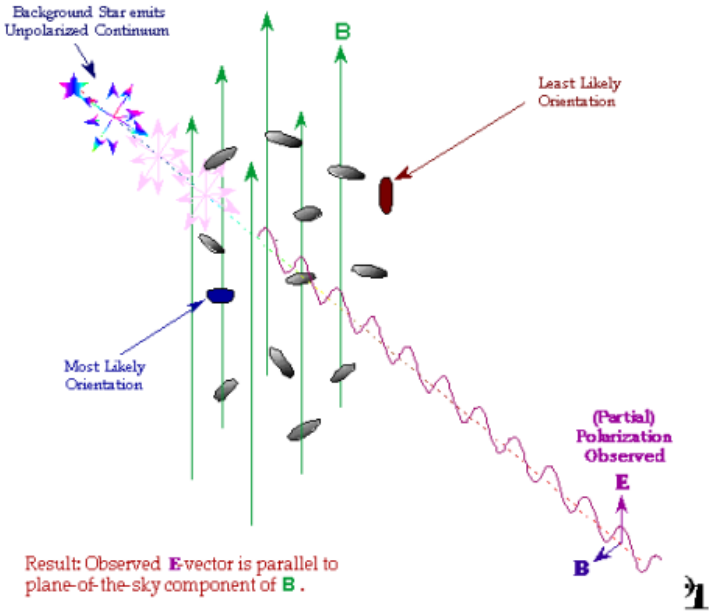


Figure 5.12: A schematic of how dust grains cause polarization of starlight to become parallel to the local interstellar magnetic field[99].

How do the grains become aligned with the local interstellar magnetic field? This occurs in two steps:

- The nonspherical grain body becomes aligned with the grain's angular momentum **J**
- The grain angular momentum **J** becomes aligned with the magnetic field **B**

Alignment of J with the grain short axis:

Looking at a simplified case of an oblate spheroid, we have a moment of inertia tensor with eigenvalues $I_1 \geqslant I_2 = I_3$ corresponding to the three principal axes, with I_1 being the "short axis." Let's label

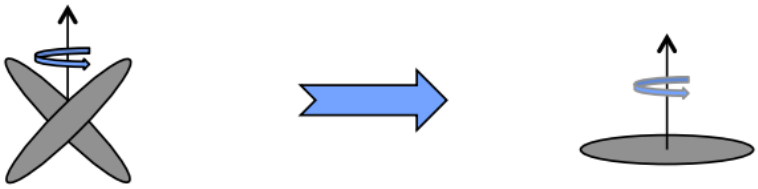


Figure 5.13: Diagram of how the grain angular momentum **J** becomes aligned with its short axis[99].

the angle between this short axis and the axis of rotation to be θ. Then, the rotational kinetic energy is

$$E_{\rm rot} = \frac{J^2}{2I_1} + \frac{J^2(I_1 - I_2)}{2I_1 I_2}\sin^2\theta \tag{5.88}$$

As one can see, this tends to be minimized when the angular momentum **J** is aligned with the short axis, so $\theta \to 0$. Conceptually, if they are not aligned with each other, the grain will "tumble" and several mechanisms will contribute to the dissipation of energy, causing θ to decrease (see Figure 5.13).

Alignment of J with B:

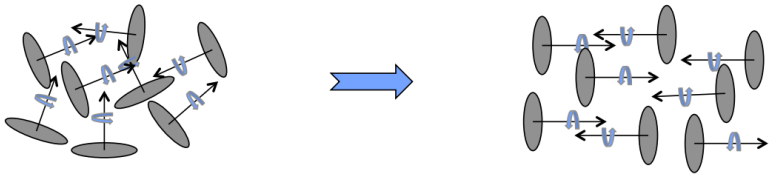


Figure 5.14: Diagram of how many grains' angular momenta **J** become aligned in the same direction with an external magnetic field **B**[99].

Due to the spinning, grains have an effective magnetic dipole moment $\boldsymbol{\mu}$ that is aligned with the instantaneous angular momentum **J**. This moment may arise either from a net charge that creates a current, or by the **Barnett effect** (which results from electrons preferentially changing their spins to be aligned with **J**). Note that this is a **paramagnetic** effect—i.e. the magnetization of the dust grain is not permanent (ferromagnetic) but only exists in response to the external field. In a similar manner to the above argument, the magnetic moment will want to align itself with an external magnetic field to minimize the magnetic dipole energy (see Figure 5.14):

$$E_{\rm dip} = -\boldsymbol{\mu}\cdot\mathbf{B} \tag{5.89}$$

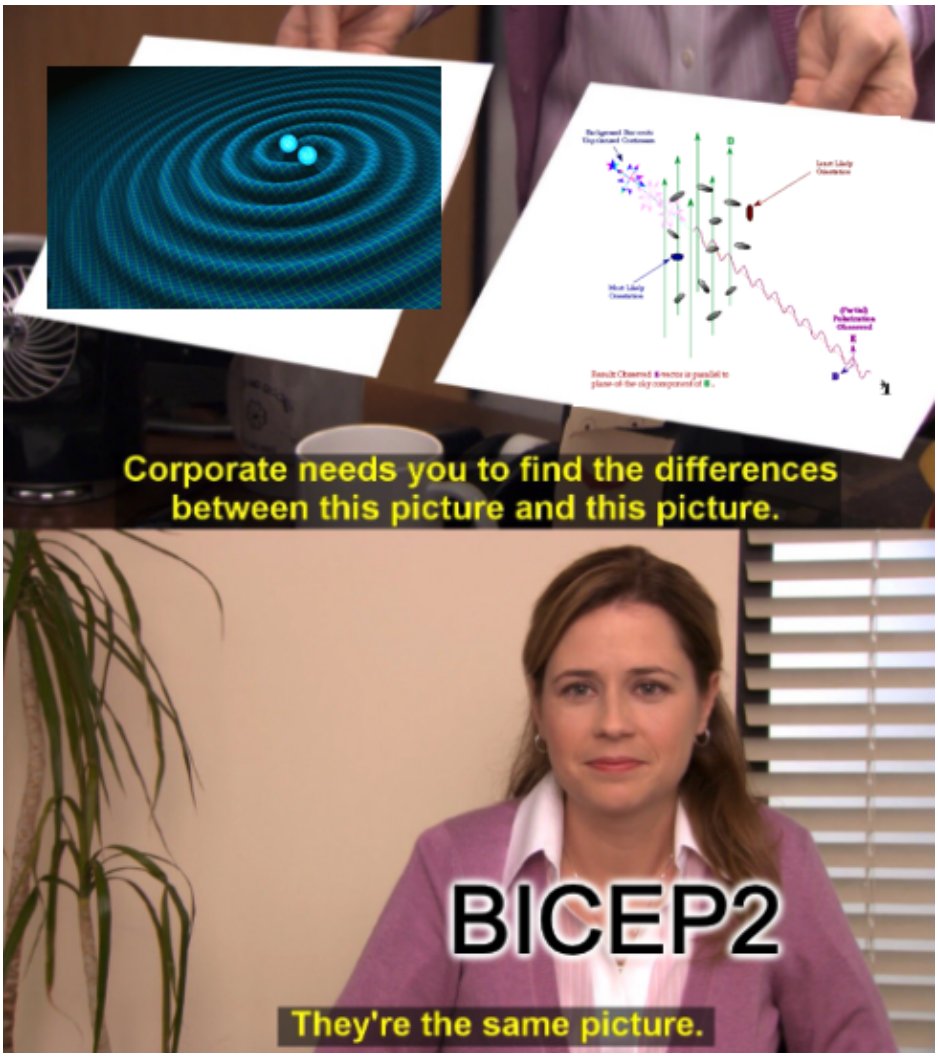
More generally, **J** will be precessing around **B** with precession rates as fast as $\lesssim 1$ yr. The external field produces a torque $\boldsymbol{\tau}_{\rm dip} = \boldsymbol{\mu}\times\mathbf{B}$ that reduces the component of **J** that is perpendicular to **B**, causing it to gradually become aligned with the field. This process tends to happen on timescales of $\tau \sim 10^6$ yr, which is short compared to the lifetimes of interstellar dust and clouds. However, there is a complication due to the fact that this mechanism is expected to be more efficient for smaller grain sizes, whereas we observe that only larger dust grains are substantially oriented to the magnetic field[64,65]. This is still an active area of research.

What major discovery in 2014 was derailed by this phenomenon? Initially, the BICEP2 team

claimed to have observed B-mode polarization in the CMB, which if true would be evidence of primordial gravitational wave signals due to inflation. This is because gravitational waves stretch space in one direction while, at the same time, compressing space in the other direction.

Note: when people talk about the "E"-mode and "B"-mode polarizations, these **do not** directly relate to the oscillations of the electric and magnetic fields. Instead, they are simply choosing two components to decompose the polarization into where one has 0 curl (E) and one has 0 divergence (B), and the names are chosen in analogy to electrostatics where the E field has no curl and the B field has no divergence. This is useful because gravitational wave signals are expected to be purely B-mode

However, it was later discovered that this polarization could be entirely explained by dichroic polarization from dust grains in the ISM. The dust grains are heated up when they absorb incident radiation, which then causes them to re-emit light that is polarized along their long axis direction (i.e. **perpendicular** to the interstellar magnetic field and also perpendicular to the direction of absorption). The average dust temperature in the Milky Way is ~ 20 K, which corresponds to a peak blackbody wavelength of ~ 150 μm. Thus, there is a significant overlap with the CMB at ~ 1 mm[100].



80. Explain the physics of 21 cm radio emissions from neutral hydrogen atoms.

The 21 cm emission line originates from atomic (neutral) hydrogen, and it arises from the spin-flip transition where the spins of the proton and electron go from aligned to anti-aligned. This is known as the **hyperfine splitting** of the 1s electronic ground state (see Figure 5.15)[64].

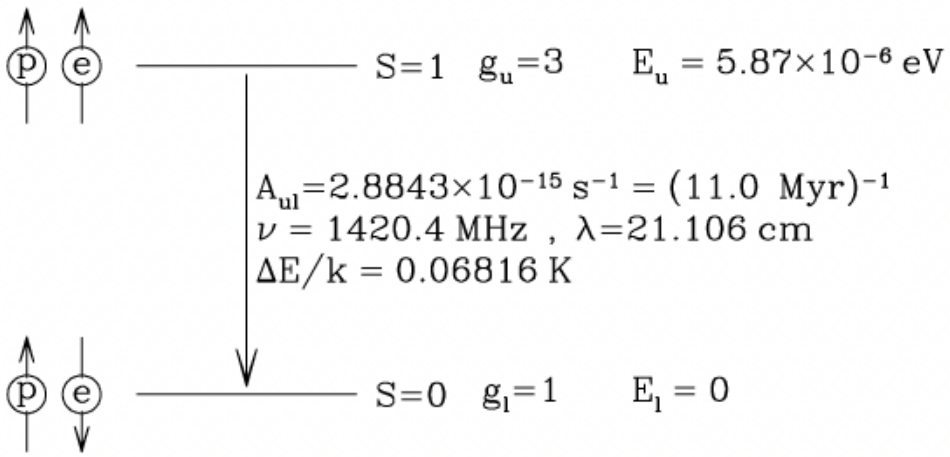


Figure 5.15: Hyperfine splitting of the 1s ground state of atomic hydrogen[64].

The transition itself is rare, happening on timescales of ~ 11 Myr, but it is observable if the quantity of H is large enough. The anti-aligned state has the lower energy, but the energy difference is small, so the temperature needed to excite atoms into the aligned-spin state is also quite small:

$$T = \frac{h\nu}{k} = \frac{hc}{\lambda k} = 0.0682\,\mathrm{K} \tag{5.90}$$

This is much smaller than the CMB temperature at ~ 2.7 K, so we can reasonably approximate the ratio of H atoms in the upper vs. lower energy states to be

$$\frac{n_u}{n_\ell} = \frac{g_u}{g_\ell} e^{-h\nu/kT_{\mathrm{spin}}} = 3e^{-0.0682\,\mathrm{K}/T_{\mathrm{spin}}} \approx 3 \tag{5.91}$$

where the **spin temperature** $T_{\mathrm{spin}} \gg 0.0682\,\mathrm{K}$. Thus, in all practical applications, the upper level contains 3/4 of the H I and the lower level contains 1/4. We can therefore say that the H I 21-cm emissivity, for all practical purposes, is independent of the spin temperature. We can use the generic formula for isotropic line emissivity from equation (5.153), which states

$$j_\nu = \frac{1}{4\pi} n_u A_{u\ell} h\nu \phi_\nu \tag{5.92}$$

given a number density n_u in the upper energy level, Einstein A coefficient $A_{u\ell}$ (see §5.Q88), and normalized line profile ϕ_ν (such that $\int \mathrm{d}\nu\, \phi_\nu = 1$). The velocity distribution of the H I gas determines ϕ_ν, and $\phi_\nu \mathrm{d}\nu$ is the probability that a photon will be emitted in the frequency interval $(\nu, \nu + \mathrm{d}\nu)$. Thus, using $n_u = (3/4) n_{\mathrm{HI}}$:

$$j_\nu \approx \frac{3}{16\pi} A_{u\ell} h\nu n_{\mathrm{HI}} \phi_\nu \tag{5.93}$$

We can also find the absorption coefficient. Again, the generic formula, from equation (5.157), is

$$\alpha_\nu = \frac{h\nu}{c}(n_\ell B_{\ell u} - n_u B_{u\ell})\phi_\nu \tag{5.94}$$

This can be rewritten in terms of the excitation temperature and the Einstein A coefficient using the relationship in equation (5.152)

$$\alpha_\nu = n_\ell \frac{h\nu}{c}\frac{g_u}{g_\ell}\frac{c^3}{8\pi h\nu^3}A_{u\ell}\phi_\nu\left[1 - \frac{n_u/n_\ell}{g_u/g_\ell}\right] = n_\ell \frac{g_u}{g_\ell}\frac{c^2}{8\pi\nu^2}A_{u\ell}\phi_\nu\left[1 - e^{-h\nu/kT_{\rm exc}}\right] \tag{5.95}$$

Where in the second step we used Boltzmann statistics. $T_{\rm exc}$ becomes the spin temperature, which as above we said $kT_{\rm spin} \gg h\nu$. Therefore, we can do a first-order Taylor expansion of the exponential to get

$$\alpha_\nu \approx n_\ell \frac{g_u}{g_\ell}\frac{A_{u\ell}}{8\pi}\lambda^2\phi_\nu\left(\frac{h\nu}{kT_{\rm spin}}\right) \approx \frac{3}{32\pi}A_{u\ell}n_{\rm HI}\frac{hc\lambda}{kT_{\rm spin}}\phi_\nu \tag{5.96}$$

We find a proportionality $\alpha_\nu \propto 1/T_{\rm spin}$. From this, we can back out an optical depth ($\tau_\nu = \int {\rm d}s\,\alpha_\nu$) and therefore a column density ($N_{\rm HI} = \int {\rm d}s\, n_{\rm HI}$). For example, if we assume a Gaussian velocity distribution with dispersion σ_V, then

$$\phi_\nu = \frac{1}{\sqrt{2\pi}\sigma_V}\frac{c}{\nu}e^{-u^2/2\sigma_V^2} \tag{5.97}$$

and

$$\alpha_\nu = \frac{3}{32\pi}\frac{A_{u\ell}\lambda^2}{\sqrt{2\pi}\sigma_V}\frac{hc}{kT_{\rm spin}}n_{\rm HI}e^{-u^2/2\sigma_V^2} \quad\longrightarrow\quad \tau_\nu = \frac{3}{32\pi}\frac{A_{u\ell}\lambda^2}{\sqrt{2\pi}\sigma_V}\frac{hc}{kT_{\rm spin}}e^{-u^2/\sigma_V^2}\int {\rm d}s\, n_{\rm HI} \tag{5.98}$$

$$\therefore\ \tau_\nu = 2.190\frac{N_{\rm HI}}{10^{21}\,{\rm cm^{-2}}}\frac{100\,{\rm K}}{T_{\rm spin}}\frac{{\rm km\,s^{-1}}}{\sigma_V}e^{-u^2/2\sigma_V^2} \quad (\therefore\ \tau_\nu \propto N_{\rm HI}) \tag{5.99}$$

Typical values in the ISM are $N_{\rm HI} \gtrsim 10^{21}$ cm^{-2}, $T_{\rm spin} \approx 100$ K, and σ_V on the order of km s^{-1}. Once the spin temperature is known, we can use it to back out the true n_u/n_ℓ ratio, since by definition[64]

$$\boxed{T_{\rm spin} = \frac{E_{u\ell}/k}{\ln\left(\frac{n_\ell/g_\ell}{n_u/g_u}\right)}} \tag{5.100}$$

This line has a few nice properties:

- it is in a wavelength regime that our atmosphere is mostly transparent to, so it is easy to observe from the ground with radio telescopes.
- Most of the Galaxy is optically thin in this wavelength band, so we can use H I 21 cm emission to map out neutral Hydrogen in the Galaxy and construct rotation curves (i.e. see §7.Q103).
- It can be used to map out neutral Hydrogen in the universe during the "dark ages" between recombination ($z = 1100$) and reionization ($z \sim 6-20$)
- It is a valuable tracer of the WNM and CNM since it originates from cold gas

81. Qualitatively describe the "equipartition energy" of a synchrotron source? What can we learn about cosmic rays from this?

A synchrotron source contains energy both in the magnetic fields (u_B) and in the relativistic particles themselves (u_E). The **equipartition** principle says that the lowest energy configuration of the system is when the energy is distributed evenly between each degree of freedom. In this case, our "degrees of freedom" can roughly be thought of as the magnetic fields and the particles.

We can find when this happens. First, the magnetic energy density is

$$\boxed{u_B = \frac{B^2}{8\pi}} \tag{5.101}$$

Similarly, the energy density in the relativistic electrons, assuming a power law distribution with index p, is

$$u_E = \int_{E_{\min}}^{E_{\max}} \mathrm{d}E\, EN(E) \propto \int_{E_{\min}}^{E_{\max}} \mathrm{d}E\, E^{1-p} \propto E^{2-p}\Big|_{E_{\min}}^{E_{\max}} \tag{5.102}$$

Now, what are the limits of integration, $E_{\min}$ and $E_{\max}$? Well, recall in §5.Q75, equation (5.41), we derived that the primary frequency a synchrotron electron with energy E will emit at is proportional to E^2B. Therefore, at some minimum and maximum frequency we have

$$E_{\min} \propto \sqrt{\nu_{\min}/B} \;\;,\;\; E_{\max} \propto \sqrt{\nu_{\max}/B} \tag{5.103}$$

In both cases, $E \propto B^{-1/2}$. Thus, we can write the electron's energy density in terms of the magnetic field:

$$u_E \propto E_{\min/\max}^{2-p} \propto B^{p/2-1} \tag{5.104}$$

We need to do some more work, since in general we don't know p. Consider the luminsoity L:

$$L = \int_{\nu_{\min}}^{\nu_{\max}} \mathrm{d}\nu\, L_\nu = \int_{E_{\min}}^{E_{\max}} \mathrm{d}E\, \frac{\mathrm{d}E}{\mathrm{d}t} N(E) \tag{5.105}$$

Recall once again from §5.Q75 we found the power emitted by a synchrotron particle (derived from the Larmor formula), see equation (5.33). We can see by inspection that $P = \mathrm{d}E/\mathrm{d}t \propto \gamma^2 u_B$, and $\gamma \propto E$, so the power $\mathrm{d}E/\mathrm{d}t \propto E^2B^2$. Thus

$$L \propto \int_{E_{\min}}^{E_{\max}} \mathrm{d}E\, E^{2-p}B^2 \propto E^{3-p}B^2\Big|_{E_{\min}}^{E_{\max}} \propto B^{p/2+1/2} \tag{5.106}$$

Therefore, if we look at the ratio of u_E to L, we see

$$\frac{u_E}{L} \propto \frac{B^{p/2-1}}{B^{p/2+1/2}} = B^{-3/2} \;\longrightarrow\; \boxed{u_E \propto B^{-3/2}} \tag{5.107}$$

Since u_B and u_E have such different dependences on B, the total energy density $u = u_B + u_E$ has a pretty sharp minimum at some point where $u_E \approx u_B$, as seen in Figure 5.16.

To find the more exact point where the minimum occurs, we can find when the derivative of u vanishes

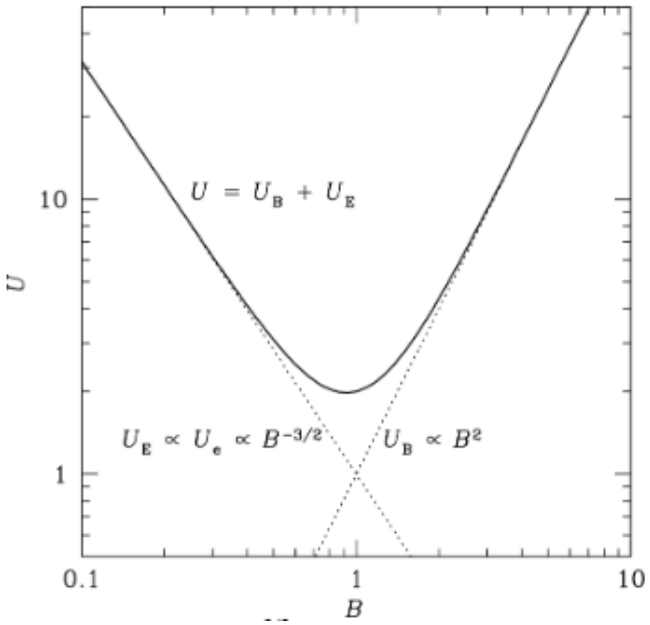


Figure 5.16: Plot of the total energy density of a synchrotron source as a function of the magnetic field B[101].

$$\frac{\mathrm{d}u}{\mathrm{d}B} = \frac{\mathrm{d}u_B}{\mathrm{d}B} + \frac{\mathrm{d}u_E}{\mathrm{d}B} = 0 \tag{5.108}$$

Since we only have proportionalities, we can work with the logarithmic derivatives:

$$\frac{1}{u_B}\frac{\mathrm{d}u_B}{\mathrm{d}B} = \frac{2B}{B^2} = \frac{2}{B} \tag{5.109}$$

$$\frac{1}{u_E}\frac{\mathrm{d}u_E}{\mathrm{d}B} = \frac{(-3/2)B^{-5/2}}{B^{-3/2}} = -\frac{3}{2B} \tag{5.110}$$

Therefore,

$$\frac{\mathrm{d}u}{\mathrm{d}B} = \frac{2u_B}{B} - \frac{3u_E}{2B} = 0 \tag{5.111}$$

Eliminating B, we're left with

$$\boxed{\frac{u_E}{u_B} = \frac{4}{3}} \tag{5.112}$$

Evidently, the state with the minimum energy doesn't have exactly equal energy densities, but rather the energy in the particles is a factor of 4/3 higher than the energy in the fields. However, this is close enough to 1 that we often approximately refer to it as an **equipartition** of energy[101].

Currently it is not known whether most radio sources are actually in equipartition or not, but most astronomers assume so, for a few reasons[101]:

1. It is physically plausible—systems with interacting components often tend towards equipartition
2. Extragalactic radio sources with high L and large volumes V end up with massive total energy requirements, $E = UV$, so minimizing the energy becomes extra important
3. It is useful to assume so because it allows for us to estimate the magnetic field strength and the relativistic energies of the particles, since we know exactly how much of the observed luminosity is due to the magnetic field vs. the particles.

What can we learn about comsic rays from this? Within the synchrotron source there are also **cosmic rays**—heavy particles like protons and ions that emit negligible synchrotron power but still contribute to the particle energy. If the ion-to-electron energy ratio is η, then the full particle energy including these cosmic rays is $u_{E,\text{tot}} = (1 + \eta)u_E$. This changes nothing about the above analysis or conclusions, since the $1+\eta$ factor would just be absorbed into the definition of u_E. Thus, by assuming equipartition, we can decompose the amount of energy in the fields and in the particles (including the cosmic rays). For reference, typical cosmic rays have $\eta \approx 40$, so $u_{E,\text{tot}}$ will be dominated by the cosmic ray energy.

82. What kinds of sources are detected at TeV energies? What is the "cosmic gamma ray horizon"?

Note: 1 TeV $\approx$ 1.602 erg.

Astrophysical TeV Sources:

- **AGN/Blazars**: Jets from AGN can accelerate electrons to relativistic speeds, and the strong magnetic fields cause them to spiral and produce synchrotron radiation (see §5.Q75). The synchrotron photons can then be scattered up to TeV energies through inverse Compton scattering. This is most commonly seen in blazars where the jets are pointed directly at us.
- **GRBs**: This works essentially by the same mechanism as with AGN/Blazars. Synchrotron photons are inverse Compton scattered up to TeV energies.
- **SNe remnants**: Supernovae produce shocks that can efficiently accelerate particles (thought to be the main source of cosmic rays). Once again, synchrotron radiation can be inverse Compton scattered to TeV energies.
- **Starburst galaxies**: Contain many young stars that have short lifetimes and produce lots of SNe, which can then produce TeVs as described in the last bullet.

The "cosmic gamma ray horizon":

The Extragalactic Background Light (EBL) is a soup of photons that permeates the space between galaxies that have been created by stars and AGN and reprocessed by dust. An energetic gamma-ray photon traveling through the EBL has a probability of interacting with another photon. If the combined energies of the two photons are large enough, they can annihilate with each other and create a pair of fermions—most likely an e^- and an e^+ due to their small masses. The Feynman diagrams for this process are shown in Figure 5.17.

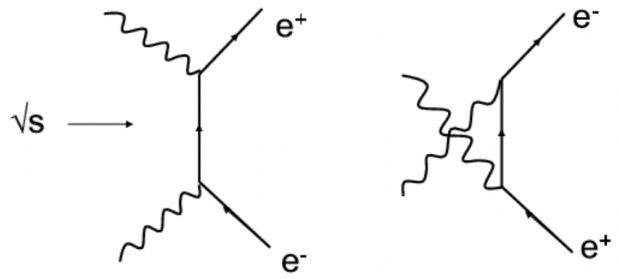


Figure 5.17: Feynman diagrams for pair production: $\gamma + \gamma \rightarrow e^- + e^+$

For this process to happen, the combined photon energies must be $h\nu_1 + h\nu_2 > 2m_e c^2$. Typical EBL photons are in the UV/optical/IR, which have energies $\ll m_e c^2$, so an extraordinarily energetic gamma-ray photon is required for pair production ($\gtrsim$ 1 GeV). Any excess energy is then given to the pair of e^- and e^+ as kinetic energy. The probability of this happening is also dependent on redshift, since the average star formation rate and AGN activity in the universe varies over cosmic time (see the "Madau plot" in §7.Q124), which necessarily varies the amount of EBL. The net effect is that high-energy sources will be **attenuated** because a fraction of their photons will be lost to pair production.

Due to these effects, the EBL has an optical depth τ that attenuates the incident radiation field I_{in} by

$$I_{\text{out}} = I_{\text{in}} e^{-\tau(E,z)} \tag{5.113}$$

Where $\tau(E,z)$ depends on both the gamma-ray photon's energy E and on the redshift of the source z. Then, the **cosmic gamma ray horizon** (CGRH) is defined as the **energy** at which this optical depth becomes 1 for a given redshift. In other words, only a fraction $1/e \approx 36.8\%$ of photons with energies equal to the CGRH will reach us (and even less for photons with higher energies). The CGRH energy evolves with redshift and is shown in Figure 5.18.

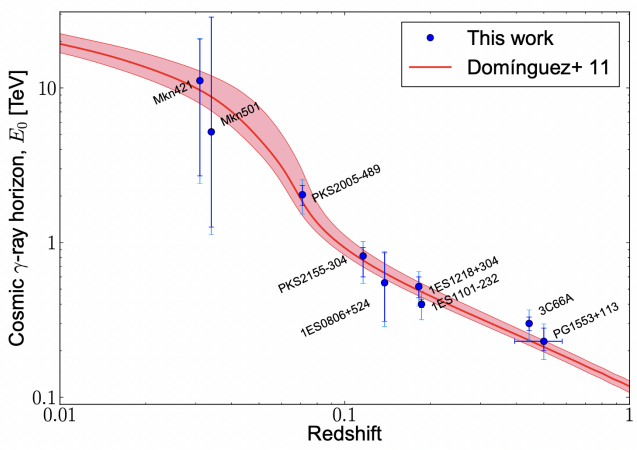


Figure 5.18: The evolution of the CGRH with redshift[102].

Thus, in the local universe, current constraints put the CGRH at ~ 20 TeV, and this drops following a polynomial curve such that it reaches ~ 0.1 TeV at a redshift of ~ 1. Again, to clarify what this means: from sources in the local universe, we can expect to see most gamma-ray photons with energies $\lesssim 20$ TeV, but as we look at sources further away, the highest energy photons start getting more likely to have been absorbed before reaching us, so this maximum energy drops down to ~ 0.1 TeV by a reshift of 1[102].

Note: we could equivalently look at this from the opposite perspective and think of the CGRH as the **source redshift** at which $\tau = 1$, such that the CGRH is now a function of the gamma-ray energy E instead of the redshift. This amounts to flipping the abscissa and ordinate of Figure 5.18. Thus, with this frame of mind, gamma-ray photons with energies of ~ 20 TeV have a CGRH of $z \lesssim 0.01$—they are only visible from sources in the local universe within $\sim 50 h_{70}^{-1}$ Mpc (comoving distance). Whereas, if we look at ~ 0.1 TeV energies, they have a CGRH of $z \approx 1$ (or $\sim 3 h_{70}^{-1}$ Gpc), allowing them to be seen at much further distances[103].

83. Why is the gas in the interstellar medium largely transparent at visible wavelengths? How does the transparency of the ISM depend on wavelength in the optical and near-IR, and why?

The combined effects of **absorption** and **scattering** are collectively referred to as **extinction**. The dependence of the extinction on wavelength is complex and relies on contributions from many different physical processes[22,65,64]. Let's go through them one at a time and explore how important they are in the context of the ISM.

- **Bound-bound absorption**: This is when a photon gets absorbed by an atom or molecule, causing an electron to change energy levels (but not ionize). This attenuates a narrow region of the continuum called an **absorption line**, which can become broadened due to Heisenberg uncertainty, pressure broadnening, and Doppler shifts. It is the opposite of bound-bound emission which creates emission lines. The strength of these is largely dependent on the individual transition.
- **Bound-free absorption**: This is when a photon with enough energy gets absorbed by an electron which kicks it out of an atom, causing it to become ionized. This results in broad, sawtooth-like features rather than lines because a specific photon energy is not required, it just has to be above some threshold (called the ionization potential or χ). So at these cutoffs there are large jumps in attenuation (e.g. the 4000 Å break) followed by gradual declines that are $\propto \nu^{-3}$. See Figure 5.19 for an example. A particularly important case here is the ionization of Hydrogen which has $\chi = 13.6$ eV from the ground state. Therefore, photons above this energy ($\lambda < 912$ Å) are strongly absorbed by this mechanism.
  - Consider an H II region around a bright star that produces lots of Lyman continuum (UV) photons with energies > 13.6 eV. Most of the atomic Hydrogen in H II regions is neutral. Therefore, these Lyman photons are basically guaranteed to be absorbed before escaping the H II region. When this happens, the H atom becomes ionized and a free electron is created. The electron will then recombine with some other H ion. Now, it can decay into the $n = 1$ ground state, producing another Lyman photon, or it can decay into the $n = 2$ state and produce a Balmer photon. In the former case, the Lyman photon will just be reabsorbed by another H atom. In the latter case, however, the Balmer photon (in the optical) will be able to escape the H II region, since it is not energetic enough to be reabsorbed by a ground state H atom, and there are much fewer H atoms in excited states that could potentially be ionized by a Balmer photon (but not 0, so this does lead to a slight jump in absorption at ~ 3600 Å known as the Balmer break). Therefore, these regions are **optically thick** to UV radiation and **optically thin** to optical radiation.
- **Free-free absorption**: The inverse of bremsstrahlung radiation, which arises when an electron interacts with another nearby ion which allows it to absorb a photon. See §5.Q76 for a detailed description of this process and how the opacity depends on the frequency. The absorption coefficient has a dependence of approximately $\alpha_\nu \propto \nu^{-2} T^{-3/2} n^2$ in the Rayleigh-Jeans limit.
  - In dense H II regions with $T \sim 10^4$ K and $n \sim 10^4$ cm^{-3}, the optical depth τ becomes 1 at ~ 141 GHz, well into the radio regime. Thus, at optical wavelengths, free-free absorption is negligible. The less dense phases that comprise most of the volume of the ISM will have an even smaller effect due to the n^2 dependence.
- **Thomson scattering**: In this process, an electron scatters a photon into a different direction

but does not absorb it. This leads to a net loss in flux since the photon can be scattered in any random direction, and our line of sight will only occupy a small fraction of that. Again, see §5.Q76 for a detailed description, but this process leads to a rather simple opacity that is independent of frequency, $\alpha = n_e \sigma_T$.

– Since the Thomson cross section σ_T is so small ($\sim 10^{-25}$ cm^2), electron scattering is typically not an important contributor to the total opacity. It will only become important at very high temperatures when the other opacity sources tend to decrease, but since the hot phases of the ISM with $T \gtrsim 10^{5.5}$ K have such low densities $n \sim 0.004$ cm^{-3}, the opacity from electron scattering is still low.

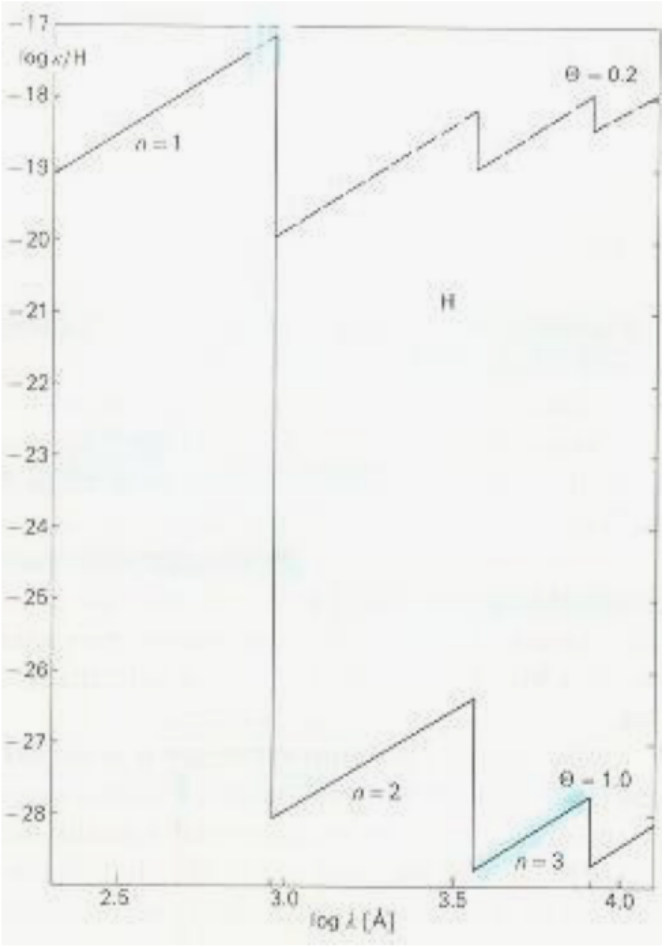


Figure 5.19: A plot showing the characteristic sawtooth shape of the bound-free opacity as a function of wavelength.

To review, **in all of the above cases, in the context of the gas in the ISM, the extinction in the optical regime is basically negligible**. We saw that bound-free absorption starts becoming important in the UV (with a slight bump in the blue-optical for the Balmer break), free-free absorption becomes important in the radio, and Thomson scattering is almost always negligible.

But wait! This all comes with one big caveat: something we have not considered here is **extinction from dust**. Dust is far and away the primary cause of extinction from the ISM in the optical and UV regimes (below the Lyman limit). The **extinction curve** plotted as a function of inverse wavelength

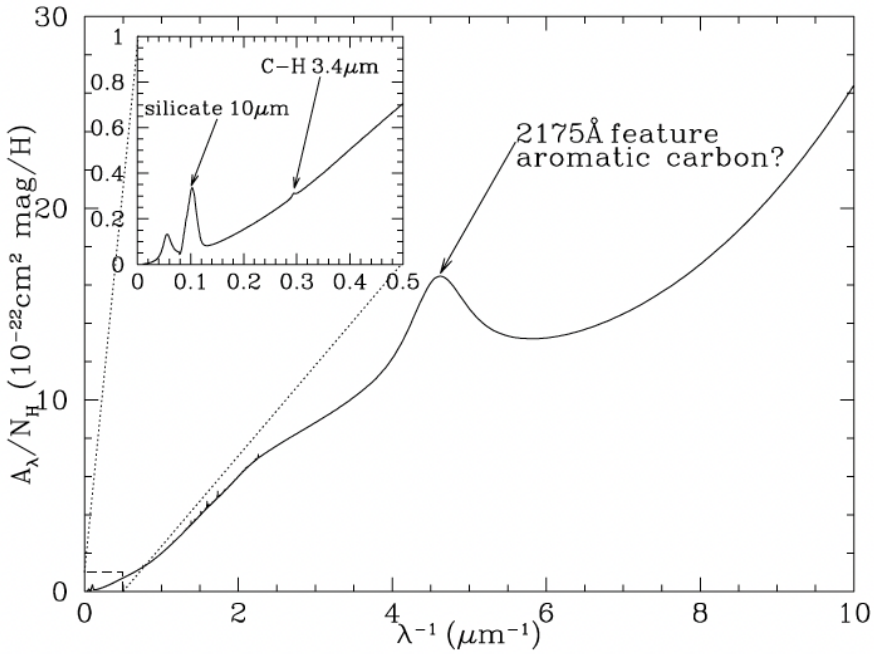


Figure 5.20: Extinction curve A_λ plotted as a function of inverse wavelength λ^{-1}. The inset plot zooms in to show the region with $\lambda > 2$ μm[64].

(λ^{-1}) is shown in Figure 5.20. Astronomers typically quantify the extinction curve as

$$\boxed{A_\lambda = -2.5 \log_{10}\left(\frac{F_\lambda}{F_\lambda^0}\right)} \tag{5.114}$$

where A_λ is in magnitudes, F_λ is the observed flux of an object, and F_λ^0 is the flux the object would have had with no extinction. This can be related to the optical depth by

$$A_\lambda = -2.5 \log_{10}(e^{-\tau_\lambda}) \approx 1.086 \tau_\lambda \tag{5.115}$$

Examining the plot, we can see that extinction from dust is *mostly* unimportant in the infrared ($\lambda^{-1} \lesssim 1$ μm^{-1}), but increases steeply through the optical ($1 \lesssim \lambda^{-1} \lesssim 3$ μm^{-1}), and continues increasing even more steeply through the UV ($\lambda^{-1} \gtrsim 3$ μm^{-1}), with a distinct "bump" feature at 2175 Å. There are a few observations we can make about this curve.

- Observed curves can vary in shape from one line of sight to another. We therefore characterize the slope of the extinction curve in visible wavelengths by the dimensionless parameter

$$\boxed{R_V \equiv \frac{A_V}{A_B - A_V} \equiv \frac{A_V}{E(B-V)}} \tag{5.116}$$

  where A_B and A_V are the extinctions measured in the B(4405 Å) and V(5470 Å) bands, and $E(B-V) \equiv A_B - A_V$ is the "reddening." Typical sight lines in the Milky Way have an $R_V \approx 3.1$ (using a Cardelli extinction law)[104], whereas starburst galaxies seem to have an average $R_V \approx 4.05$ (using a Calzetti extinction law)[105].

- If the dust grains were large compared to the wavelength, the extinction cross section would be independent of wavelength since it would just be due to the geometry of the dust grains. In this case, $R_V \to \infty$. However, we notice that the extinction continues to rise even for the shortest UV wavelengths where we can measure it. Therefore, there must be dust grains with size parameters $2\pi a/(\lambda = 1000\,\text{Å}) \lesssim 1$, i.e. dust grains with sizes $a \lesssim 150\,\text{Å}$. The actual extinction from these grains can then be calculated analytically with **Mie theory**, which assumes the grains are homogeneous spheres with an isotropic dielectric function.
- The dust grains and gas in the ISM appear to be well mixed. Therefore, we can use the reddening to estimate the column density of Hydrogen (or vice versa). $N_H/E(B-V) = 5.8 \times 10^{21}\,\text{cm}^{-2}\,\text{mag}^{-1}$.

Now we look at some of the prominent features individually[65].

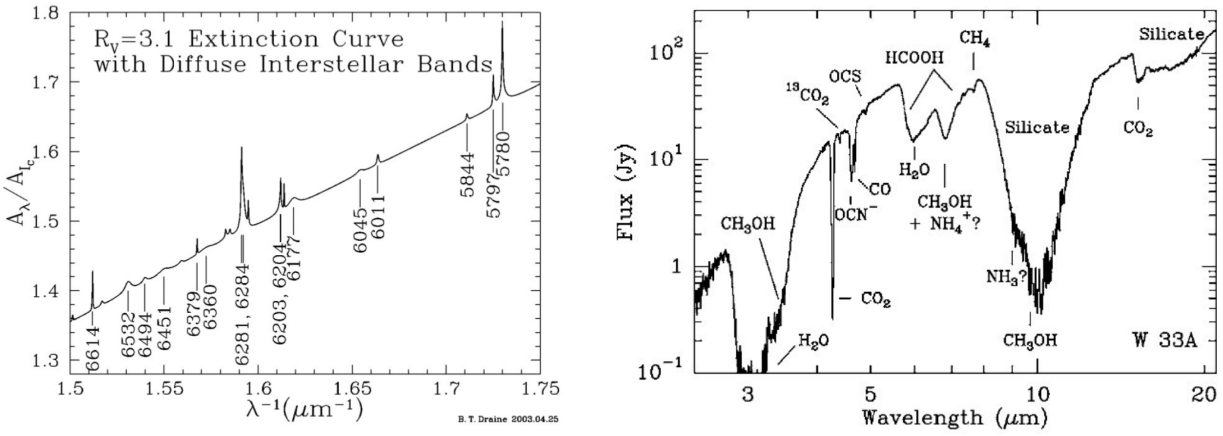


Figure 5.21: *Left*: A zoomed-in part of the extinction curve showing the DIBs[65]. *Right*: An infrared spectrum of the protostar W 33A that is embedded in a molecular cloud. The spectrum shows many absorption features from ices[65,106].

The 2175 Å feature: This feature is thought to be caused by $\pi \to \pi^*$ transitions (i.e. from delocalized electrons) in graphitic material, which could be graphite itself, or it could be Polycyclic Aromatic Hydrocarbon (PAH) molecules. These have a distinct hexagonal structure, and due to this complex molecular structure, vibrational and rotational energy transitions generally lead to much wider emission and absorption bands than those seen from atomic transitions.

Diffuse Interstellar Bands (DIBs): Between 0.38–0.868 μm there are over 100 observed absorption features of varying widths and strengths. They were discovered in 1922, yet over 100 years later we are still unaware of what causes them! They are too broad to be small molecules. See a zoomed-in part of the extinction curve that shows these features in the left panel of Figure 5.21.

The silicate features: These are barely visible on the full plot, but become apparent on the inset that is zoomed in to the region $\lambda^{-1} < 0.5\ \mu\text{m}^{-1}$. They are at ~ 10 and ~ 18 μm and are thought to be caused by silicate dust grains consisting of Si, O, Fe, Mg, and other metals. The 10 μm feature corresponds to the stretching of the Si-O bonds, and the 18 μm feature corresponds to the bending of the Si-O bonds about the Si vertex (see Figure 5.22). Since these features are so broad and lack substructure, they are likely from amorphous grains rather than crystalline grains (meaning there is no long-range order, the structure varies smoothly throughout the grain).

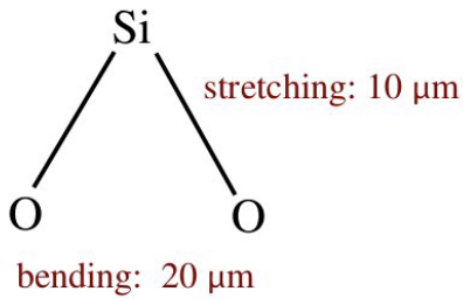


Figure 5.22: Depiction of the bending and stretching of SiO to produce the observed infrared absorption features[65].

The CH feature: Again, this feature is only barely visible on the inset next to the silicate features. At 3.4 μm, this is caused by the stretching of CH bonds in aliphatic (i.e. chain-like) hydrocarbons.

Ice features: There are many more absorption features due to ices that can be observed through sight lines passing through molecular clouds. These include, but are not limited to, H_2O, CH_4, NH_3, and CO_2. See an example in the right panel of Figure 5.21.

The amogus feature: This sussy feature is only visible if we zoom in extremely close to the peak of the 10 μm silicate feature (see Figure 5.23). This feature is thought to arise from isolated and relatively large amogus-shaped dust grains which have likely been ejected from their fellow dust crews' clouds for being too sussy. The composition of these sussy bakas is not well understood. The fact that their absorption band lies on top of the silicate features suggests that they are some combination of Si, O, and other metals, but it could be that they are impostors hiding some icy compositions instead, which would explain why they have been ejected[4].

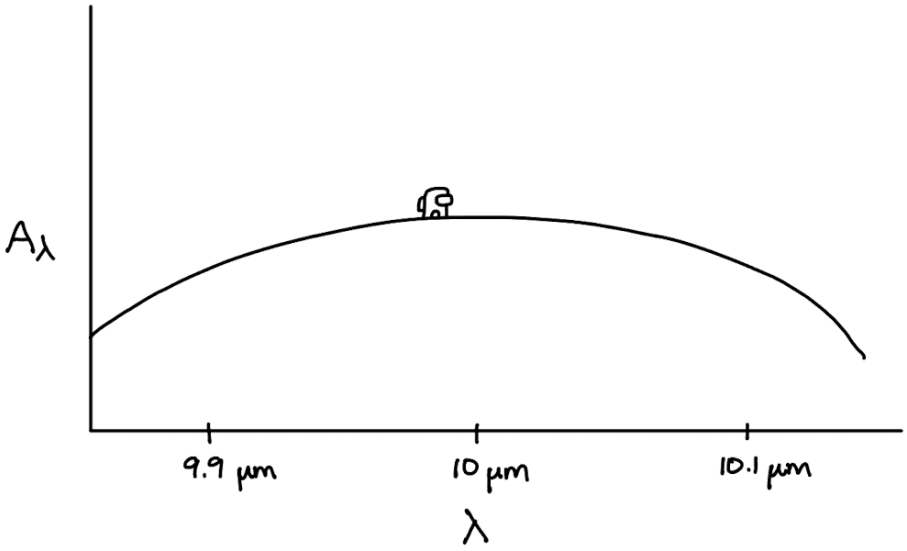


Figure 5.23: The amogus feature at 10 μm.

[4]See the database of all the amogi spectral features here

84. Sketch a typical cooling function $\Lambda(T)$ for diffuse interstellar gas and identify its prominent features, spanning from IR to X-ray. Overplot a hypothetical heating curve and show how to identify points of thermal equilibrium and their stability.

Cooling[64,107,108,109]:

Cooling can be quantified in terms of the radiative cooling rate per unit volume, $[\mathscr{C}] = \mathrm{erg\,s^{-1}\,cm^{-3}}$. A number of different processes contribute to cooling at different temperatures. In the high-temperature regime (above $\sim 10^7$ K), we have

- **Bremsstrahlung emission**: Free-free interactions between electrons and ions produce an emissivity given by equation (7.18). Integrating this over frequency gives us the contribution to cooling, which we see has the proportionality $\mathscr{C}_{\mathrm{ff}} \propto n_e^2 T^{1/2}$.

In the lower temperature regime, we have

- **Collisional excitation**: An electron collides with an atom, exciting one of its electrons (but not ionizing it). The excited electron then emits a photon and returns to a lower energy level. The cooling rate from these excitations depends on how likely they are to interact, so it's proportional to the densities of both the electrons and the atoms they are exciting (most likely H), so $\mathscr{C}_{\mathrm{exc}} \propto n_e n_H$.
- **Collisional ionization**: An electron collides with an atom, removing one of its electrons and creating an ion. This removes an energy equal to the ionization threshold from the gas. This does not directly result in any radiative cooling, but it is a necessary process to produce ions that can then recombine with electrons.
- **Recombination**: An electron recombines with an atom, emitting a photon. This produces a cooling rate that depends on both the density of electrons and ions in the plasma, $\mathscr{C}_i \propto n_e n_i$.

(note the above three processes also have strong temperature proportionalities). It is often assumed that the plasma is in **Collisional Ionization Equilibrium** (CIE), which is a steady state where collisional ionization, charge exchange, and recombination are the only processes that alter the ionization balance. In other words, CIE assumes that there is no incident ionizing radiation. Under this assumption, the ionization fraction of each element depends **only** on the temperature, and not the density of the gas.

Notice that all of the above examples of $\mathscr{C}$ are proportional to two number densities. At temperatures $\gtrsim 10^4$ K, the ionization of Hydrogen provides enough free electrons that collisional excitation is dominated by electron collisions. And at low enough densities, every collisional excitation is followed by a radiative decay. Therefore, all of the cooling processes have a common factor of $n_e n_H$. Thus it is often customary to introduce the **cooling function**

$$\Lambda(T, Z) \equiv \frac{\mathscr{C}}{n_e n_H} \quad , \quad [\Lambda] = \mathrm{erg\,s^{-1}\,cm^3} \tag{5.117}$$

which removes the dependence on density. Also note that since collisional excitation, ionization, and recombination will look differently for different metals, Λ also strongly depends on the metallicity Z. The cooling function in CGS units ($\mathrm{erg\,s^{-1}\,cm^3}$) is plotted in Figure 5.24 for a number of different metallicities.

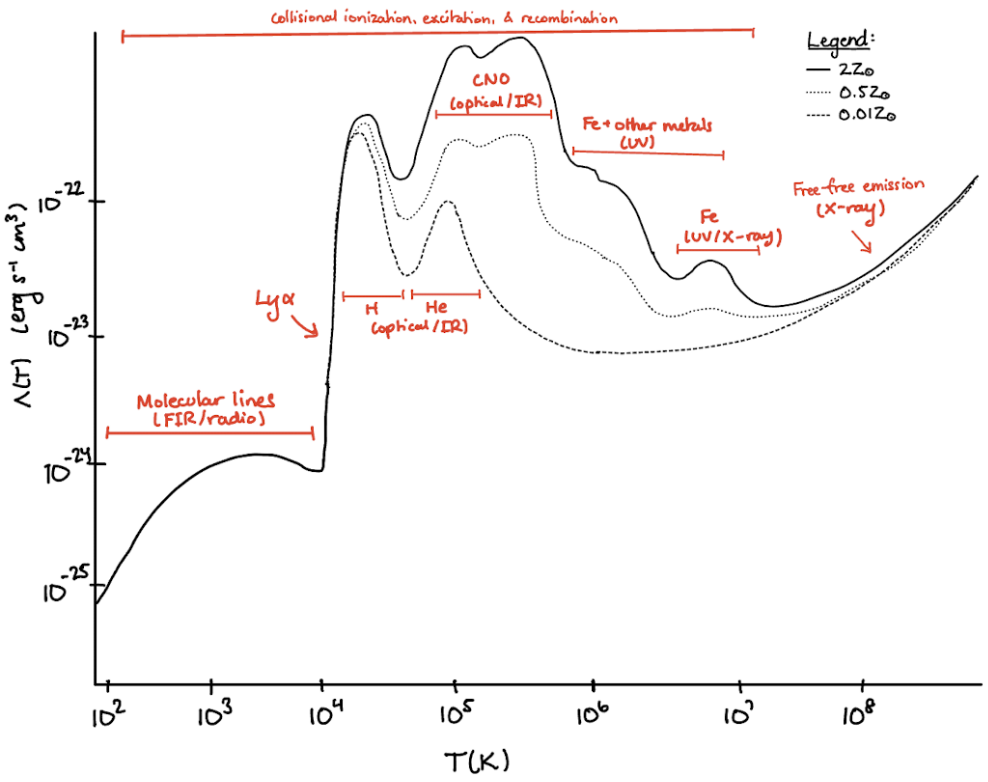


Figure 5.24: The cooling function Λ plotted against the temperature T, for a few select metallicities.

We can decompose this into contributions from different lines, which looks like Figure 5.25.

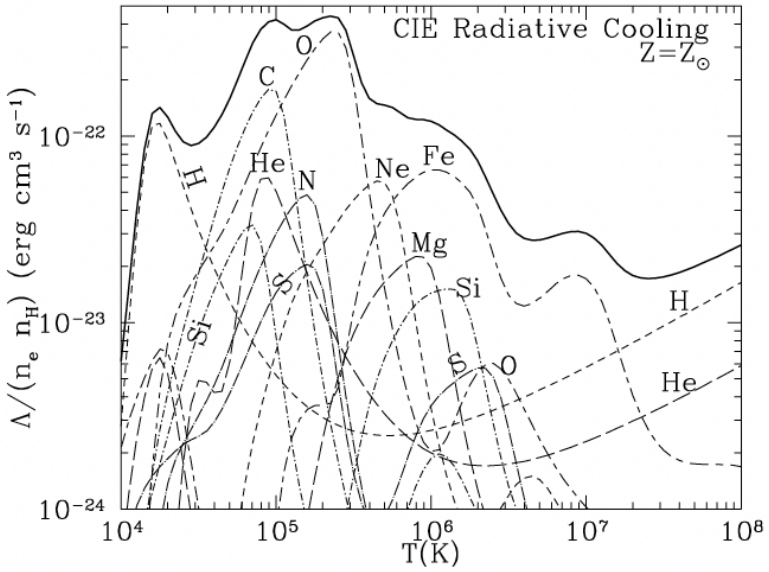


Figure 5.25: Decomposing of a $Z = Z_\odot$ cooling curve into contributions from individual species. Notice that each species is fairly localized at a specific temperature[64]. Note: ignore the axis label, Draine defines Λ differently than us (he essentially has Λ for what we are calling $\mathscr{C}$).

Heating[108,109]:

Heating can be quantified in a similar way to cooling, with the heating rate per unit volume that has the same units as $\mathscr{C}$, i.e. $[\mathscr{H}] = \mathrm{erg\,s^{-1}\,cm^{-3}}$. Once again, there are a number of sources that contribute to heating in different conditions:

- **Photoelectrons ejected from dust grains**: High-energy photons can also knock electrons out of dust grains, pumping more kinetic energy into the gas. Most likely these are from small grains like PAHs. The heating rate here is proportional to the gas number density, since the gas and dust are well mixed, $\mathscr{H}_{\mathrm{pe}} \propto n$.
- **Cosmic rays**: These high-energy (1–10 MeV) protons can collide with atoms and eject electrons, injecting energy much like the previous two mechanisms. As such, the proportionality here is the same, $\mathscr{H}_{\mathrm{CR}} \propto n$.
- **Photoionization**: Aside from collisional ionization, an atom may also be ionized if it absorbs a photon with enough energy. It is unlikely that the photon has the *exact* amount of energy needed to ionize the atom, so whatever extra energy it has will be transformed into the kinetic energy of the electron, thus heating the gas by an amount $h\nu - \chi$. This is offset by recombination, but since recombination rates are generally higher for lower-energy electrons, there is still a net heating that occurs from photoionization. This heating rate is proportional to the number density of ions that can be ionized by the incident radiation and the cross-section for photoionization. Therefore, $\mathscr{H}_{\mathrm{phot}} \propto n_i$.

Again, in a similar fashion to what we did for the cooling curve, we can define a density-independent version of the heating rate since all of these mechanisms have the same scaling with density (linear this time instead of quadratic). This is the **heating function**

$$\Gamma(T) \equiv \frac{\mathscr{H}}{n} \quad , \quad [\Gamma] = \mathrm{erg\,s^{-1}} \tag{5.118}$$

In general, the first two mechanisms dominate the heating of the diffuse ISM, while photoionization is relatively unimportant. This is because the primary source of photoionizing photons for Hydrogen is in the UV above the Lyman limit, which as we learned in §5.Q83 is highly attenuated due to being mostly trapped within dense H II regions around O and B stars. Thus, outside of these regions, the rest of the ISM has relatively little UV photon flux for photoionization.

The balance of heating and cooling[65,108]:

The cosmic ray heating rate is independent of temperature, so for simplicity let's just consider a cosmic-ray heating function $\Gamma_{\mathrm{CR}} = \mathrm{const}$ (in reality, the photoelectron heating rate is just as important, and has a complicated dependence on temperature and ionization fraction). For this constant Γ_{CR}, we can plot it on top of our cooling function Λ, but to keep it in the same units we will have to look at Γ/n. Since for the ISM, n decreases as T increases, maintaining a constant pressure, we will have a positive slope ($n^{-1} \propto T$). See what this looks like in the left panel of Figure 5.26.

Now, we can imagine subtracting these two function from each other to get a net heating/cooling rate. This is describes by the **generalized loss function**

$$\boxed{\mathscr{L}(T) \equiv n^2\Lambda(T) - n\Gamma(T)} \tag{5.119}$$

where if $\mathscr{L} > 0$ we have net cooling, if $\mathscr{L} < 0$ we have net heating, and if $\mathscr{L} = 0$ we have equilibrium. Plotting this for our example makes it obvious which points have thermal equilibrium, we just find

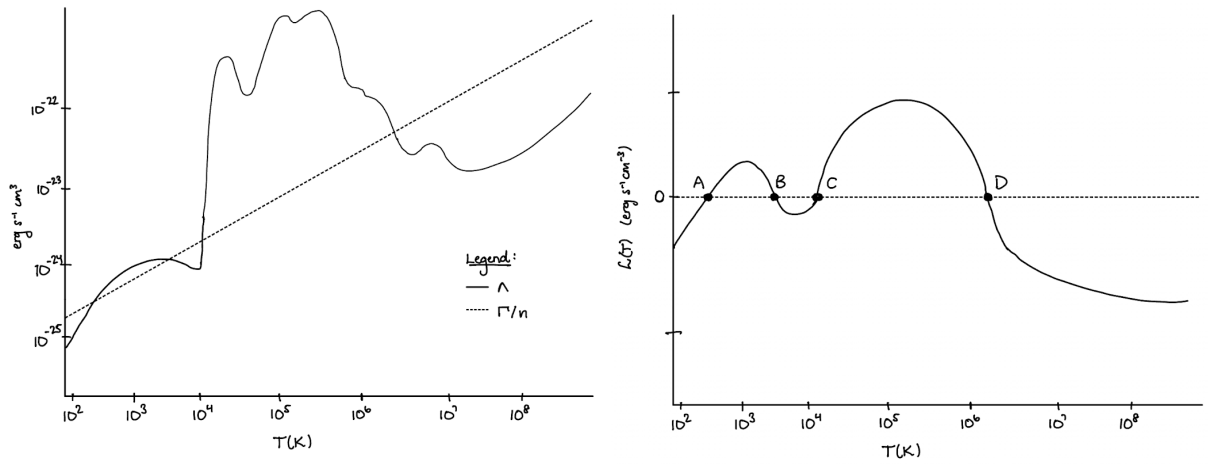


Figure 5.26: *Left*: A hypothetical heating rate from cosmic rays plotted over our CIE cooling function. *Right*: The loss function obtained by subtracting the cooling and heating functions from the left panel.

where $\mathscr{L} = 0$ (see the right panel of Figure 5.26).

How can we identify the stability of these points? It's fairly straightforward. Imagine we are at the equilibrium point labeled "A" in the figure and we experience a small perturbation in temperature. Here we notice the derivative (at constant pressure) is

$$\left(\frac{\partial \mathscr{L}}{\partial T}\right)_P > 0 \tag{5.120}$$

That is, if our temperature slightly increases, then we start to have net cooling, and if the temperature decreases, we start to have net heating. Perturbations in both directions tend to bring us back to our original point, so this means A is a **stable equilibrium** point. By inspection, so is C. These two points roughly correspond to the CNM and WNM.

On the flip side, points B and D have the opposite relationship:

$$\left(\frac{\partial \mathscr{L}}{\partial T}\right)_P < 0 \tag{5.121}$$

which will tend to cause runaway heating or cooling if we have a slight perturbation in either direction. Therefore, these points are in an **unstable equilibirum**.

Again, this is a simple example with a really unrealistic heating function, but it gets the general point across about how one can use the loss function to find stable and unstable equilibrium points. The points A and C in the figure can roughly be identified as the CNM and WNM phases of the ISM (§5.Q72).

85. What is "brightness temperature"? What is a "Jansky"? How are these different?

Radio astronomers typically describe the specificy intensity of a source I_ν using the **brightness temperature** T_B. This is defined as the temperature of a hypothetical blackbody that has an intensity of I_ν at frequency ν. In other words,

$$I_\nu(\nu) = B_\nu[\nu, T_B(\nu)] = \frac{2h\nu^3}{c^2}\frac{1}{\exp(h\nu/kT_B)-1} \tag{5.122}$$

At first blush this seems ridiculous since the relationship between I_ν and T_B is nonlinear. But in radio astronomy, we are in the Rayleigh-Jeans regime where $h\nu \ll kT_B$, for which the relationship does become linear:

$$\exp\left(\frac{h\nu}{kT_B}\right) \approx 1 + \frac{h\nu}{kT_B} \quad \longrightarrow \quad \boxed{I_\nu \approx \frac{2\nu^2}{c^2}kT_B} \tag{5.123}$$

In contrast, a **Jansky** is a unit of **specific flux** (note: not intensity, i.e. not per solid angle like the brightness temperature; see Appendix §12.11). By definition, it is

$$\boxed{1\,\mathrm{Jy} \equiv 10^{-23}\,\mathrm{erg\,s^{-1}\,cm^{-2}\,Hz^{-1}}} \tag{5.124}$$

This unit is used commonly in radio astronomy since a typical galaxy has a flux on the order of a few Janskys in the radio[65].

86. Make a simple, classical argument to show that the spectrum of radiation from monoenergetic electrons with a speed v impinging on ions at an impact parameter b would be roughly flat up to a frequency $\sim v/b$.

The electrons will be decelerated during their pass-by of the ions (which we assume are positively charged) due to electromagnetic interactions, causing them to emit photons. Note: We ignore the case where the ion is negatively charged, since the electron would be accelerated and gain energy, meaning it does not emit any radiation.

We can start with the **Larmor formula** to find the power radiated from the acceleration:

$$\boxed{P = \frac{dE}{dt} = \frac{2}{3}\frac{e^2}{c^3}\mathbf{a}^2(t)} \tag{5.125}$$

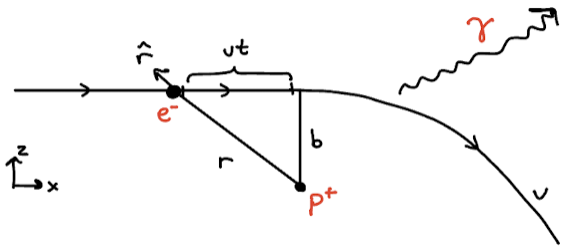


Figure 5.27: Diagram of particle scattering.

Assume an electron traveling in the $+\hat{\mathbf{x}}$ direction at speed v approaches another particle with a charge of Ze at an impact parameter of b. Define the time $t = 0$ to be when the electron is directly above the charged particle. Then the acceleration is

$$\mathbf{F} = -\frac{Ze^2}{r^2}\hat{\mathbf{r}} = m_e\mathbf{a}(t) \quad \longrightarrow \quad \mathbf{a}(t) = -\frac{Ze^2}{m_e r^2}\hat{\mathbf{r}} = -\frac{Ze^2}{m_e(b^2+v^2t^2)}\hat{\mathbf{r}} \tag{5.126}$$

Expanding the unit vector $\hat{\mathbf{r}}$, we have

$$\mathbf{a}(t) = \frac{Ze^2}{m_e(b^2+v^2t^2)^{3/2}}(-vt\hat{\mathbf{x}} - b\hat{\mathbf{z}}) \tag{5.127}$$

See Figure 5.27 for the geometry. From here, we can already arrive at our desired conclusion with a bit of thought: The acceleration $\mathbf{a}(t)$ falls off quickly to 0 when $|t|$ is large because of the t^2 dependence in the denominator. Thus, the acceleration is only substantial during a time interval of $\tau \sim b/v$. Because this time interval is short, the bulk of the radiation will be emitted as a **pulse**. In the frequency domain, this is a **wave packet** consisting of a superposition of many frequencies.

What frequencies will contribute to the packet? Well, the motions of the electrons will not change significantly on any timescales shorter than τ (since during this interval the acceleration is roughly constant at the maximum value), so this corresponds to a **maximum frequency** of $\nu \sim 1/\tau \sim v/b$. However, there are changes in motion on longer timescales since the electromagnetic force is long-range, so there are contributions from frequencies **below** this limit all the way down to essentially $\nu \sim 0$.

With this, in principle, **we're done**. However, for funsies let's continue and see if we can actually get an expression for the spectrum $\mathrm{d}E/\mathrm{d}\omega$ as a function of v.

We can use Fourier analysis to look at our problem in the frequency domain and thus find the spectrum of radiation emitted. There are two properties of the Fourier transform that will be useful for us.

1. Parseval's theorem, which states

$$\int_{-\infty}^{\infty} \mathrm{d}t\, f^2(t) = \int_{-\infty}^{\infty} \mathrm{d}\omega\, \tilde{f}^2(\omega) \tag{5.128}$$

2. If $f(t)$ is real, it follows that

$$\tilde{f}(-\omega) = \tilde{f}^*(\omega) \tag{5.129}$$

Putting these into use, we integrate (5.125) over time to obtain

$$E = \frac{2}{3}\frac{e^2}{c^3}\int_{-\infty}^{\infty} \mathrm{d}t\, \mathbf{a}^2(t) = \frac{2}{3}\frac{e^2}{c^3}\int_{-\infty}^{\infty} \mathrm{d}\omega\, \tilde{\mathbf{a}}^2(\omega) = \frac{4}{3}\frac{e^2}{c^3}\int_{0}^{\infty} \mathrm{d}\omega\, \tilde{\mathbf{a}}^2(\omega) \tag{5.130}$$

Thus, the spectrum is

$$\frac{\mathrm{d}E}{\mathrm{d}\omega} = \frac{4}{3}\frac{e^2}{c^3}\tilde{\mathbf{a}}^2(\omega) \tag{5.131}$$

Evidently, we need to get the Fourier transform of the acceleration, i.e.

$$\tilde{\mathbf{a}}(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{d}t\, e^{-i\omega t}\mathbf{a}(t) \tag{5.132}$$

Using our arguments about the collision time τ, the integral can be well approximated by lowering the limits of integration

$$\tilde{\mathbf{a}}(\omega) \approx \frac{1}{\sqrt{2\pi}}\int_{-\tau/2}^{\tau/2} \mathrm{d}t\, e^{-i\omega t}\mathbf{a}(t) \tag{5.133}$$

In the limit $\omega\tau \gg 1$, the integral oscillates rapidly and averages out to 0, while in the limit $\omega\tau \ll 1$, the exponential goes to 1, so we have

$$\tilde{\mathbf{a}}(\omega) \sim \begin{cases} \Delta v/\sqrt{2\pi} & \omega\tau \ll 1 \\ 0 & \omega\tau \gg 1 \end{cases} \tag{5.134}$$

Also, the x-component of the acceleration is positive for $t < 0$ (as the electron approaches) and negative for $t > 0$ (as the electron flies away), so the change in velocity in the x-direction averages out to $\sim$0. In contrast, the z-component is always negative, so it will add up. Thus we can consider the acceleration to only be important in the z direction. In this case, we can get Δv by integrating the z-component from (5.127). Defining $x \equiv vt/b$:

$$\Delta v = \frac{Ze^2}{m_e b v}\int_{-\infty}^{\infty} \frac{\mathrm{d}x}{(1+x^2)^{3/2}} = \frac{2Ze^2}{m_e b v} \tag{5.135}$$

Plugging back into (5.131)

$$\frac{\mathrm{d}E}{\mathrm{d}\omega} = \begin{cases} \dfrac{8Z^2e^6}{3\pi m_e^2 c^3 b^2 v^2} & \omega\tau \ll 1 \\ 0 & \omega\tau \gg 1 \end{cases} \tag{5.136}$$

Here, notice that $\mathrm{d}E/\mathrm{d}\omega$ approaches a constant value in the low-frequency limit and drops to 0 in the high-frequency limit. The cutoff frequency where we transition between these two limits is $\nu\tau \sim 1 \longrightarrow \nu \sim 1/\tau \sim v/b$, below which point the spectrum is essentially flat. This agrees with the conclusion we reached earlier. The actual frequency dependence is complicated since it depends on the exact form of the Fourier transform in equation (5.132) (which turns out to be a modified Bessel function).

The observed spectrum we actually see will just be this $\mathrm{d}E/\mathrm{d}\omega$ integrated over many different scattering events within some unit time and unit volume, but since in this simple example we are considering all scattering events to have the same v and the same b, this should not introduce any additional dependence on frequency. The spectrum is plotted in Figure 5.28 in both the time and frequency domains. See Refs. 93,94,95,110.

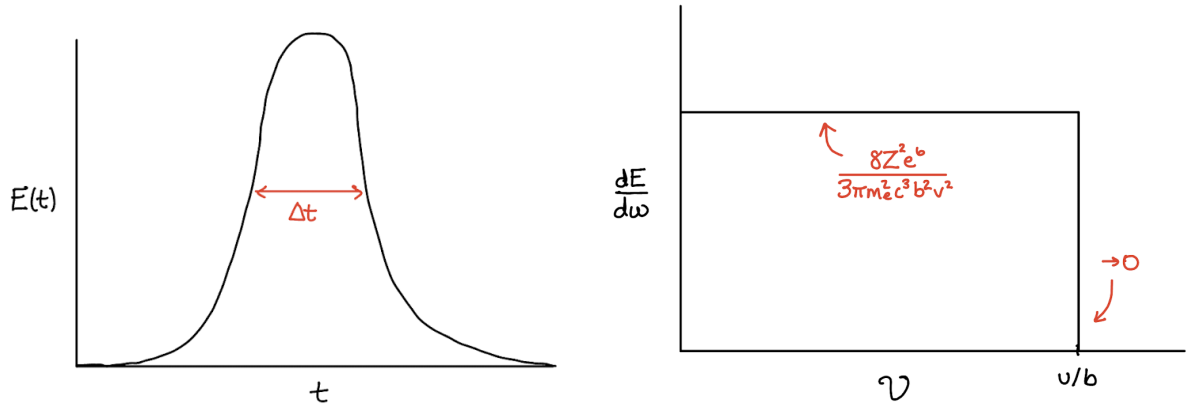


Figure 5.28: *Left*: The pulse of energy emitted by a single scattering event as a function of time. *Right*: The rough behavior of the spectrum for a single scattering as a function of frequency.

87. If a typical interstellar dust grain is 0.2 microns in size, and starlight suffers an extinction of 1 magnitude per kpc, estimate the space density of dust grains.

The extinction, from equation (5.115), is related to the density of the dust grains by

$$A_\lambda \approx 1.086\tau_\lambda = 1.086 \int \mathrm{d}s\, \alpha_\nu = 1.086 n\sigma s \tag{5.137}$$

Where the extinction $A_\lambda = 1$ mag, the path length $s = 1$ kpc, and the cross-section σ is the sum of the cross sections from absorption and scattering from the dust. In this case, the dust grain radii a are large compared to the wavelength ($a \gg \lambda$), so we are in the simple geometric optics limit. Naïvely from everyday experience, we would expect the cross section to just be $\sigma = \pi a^2$, the cross-sectional area of the dust grains. However, **this is wrong**. Why? In this case, the dust grains are small and far away from us, so not only do they block the light that is coming directly towards us from behind the dust grain, but they also produce a non-negligible amount of small-angle scattering at the edges of the dust grain that further reduces the observed flux. The end result is that the actual cross section is **exactly twice** the cross-sectional area of the dust grain,

$$\sigma = 2\pi a^2 \tag{5.138}$$

The factor of 2 can be derived from the $a \gg \lambda$ limit in Mie theory. See a demonstration of this effect in Figure 5.29[64,65].

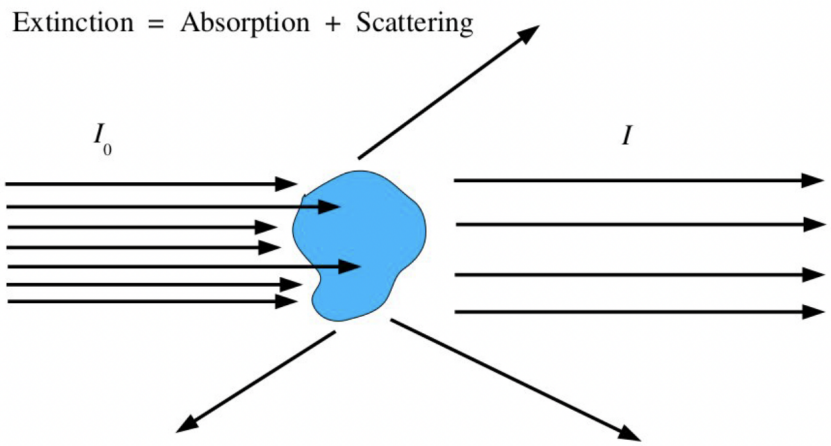


Figure 5.29: A visual demonstration of the extinction of light from a large dust grain[65].

Therefore, solving for n, we have

$$n = \frac{A_\lambda / 1.086}{2\pi a^2 s} \tag{5.139}$$

Ignoring all constants of order unity, we have

$$n \sim \frac{1}{(10^{-4})^2 \times 10^3 \times 10^{18}} \sim 10^{-13}\,\mathrm{cm}^{-3} \tag{5.140}$$

88. What are the Einstein A and B coefficients for a spectral line, and what are the relationships among them?

The are a number of processes involved in the transitions between energy levels that are related to a spectral line. Let's say we have X, which can represent an atom, ion, molecule, dust grain, or anything that has energy levels, and it transitions between a lower level ℓ and upper level u. The number density of X in each level is n_ℓ and n_u.

- **Absorption**: X absorbs an incident photon with just enough energy and moves from the lower level to the upper level

$$X_\ell + h\nu \rightarrow X_u \quad , \quad h\nu = E_u - E_\ell \tag{5.141}$$

  The rate that X will absorb photons will obviously depend on the number density n_ℓ and the number density of incident photons with the right energy. There will also be a proportionality factor: the probability that X will actually absorb a photon per unit time per unit energy density of incident photons. We call this factor the **Einstein B coefficient** $B_{\ell u}$. Thus, we can write the change in the density of the upper and lower levels of X from absorption as

$$\left(\frac{\mathrm{d}n_u}{\mathrm{d}t}\right)_{\mathrm{abs}} = -\left(\frac{\mathrm{d}n_\ell}{\mathrm{d}t}\right)_{\mathrm{abs}} = n_\ell u_\nu B_{\ell u} \tag{5.142}$$

  Here, u_ν is the energy density of photons with the right frequency (erg cm^{-3} Hz^{-1}). Therefore, $[B_{\ell u}] = \mathrm{erg}^{-1}\,\mathrm{cm}^3\,\mathrm{s}^{-1}\,\mathrm{Hz}$.

- **Spontaneous emission**: X moves from the upper level to the lower level and emits a photon with the difference in energies between the two levels

$$X_u \rightarrow X_\ell + h\nu \quad , \quad h\nu = E_u - E_\ell \tag{5.143}$$

  This is a random process, so the probability per unit time that X will undergo this transition is given by the **Einstein A coefficient** $A_{u\ell}$. In other words, we can write the change in the density of X from spontaneous emission as

$$\left(\frac{\mathrm{d}n_\ell}{\mathrm{d}t}\right)_{\mathrm{spon}} = -\left(\frac{\mathrm{d}n_u}{\mathrm{d}t}\right)_{\mathrm{spon}} = n_u A_{u\ell} \tag{5.144}$$

  Thus, $[A_{u\ell}] = \mathrm{s}^{-1}$.

- **Stimulated emission**: This is similar to spontaneous emission, but here an incident photon gives X a little kick that causes it to emit its own photon

$$X_u + h\nu \rightarrow X_\ell + 2h\nu \quad , \quad h\nu = E_u - E_\ell \tag{5.145}$$

  Therefore, this process also depends on the incident radiation of photons, so we use the Einstein B coefficient $B_{u\ell} \neq B_{\ell u}$

$$\left(\frac{\mathrm{d}n_\ell}{\mathrm{d}t}\right)_{\mathrm{stim}} = -\left(\frac{\mathrm{d}n_u}{\mathrm{d}t}\right)_{\mathrm{stim}} = n_u u_\nu B_{u\ell} \tag{5.146}$$

How are these coefficients related to each other?

We can find a relationship between $A_{u\ell}$, $B_{u\ell}$, and $B_{\ell u}$ by considering the case of thermal equilibrium. In this case, the radiation field is blackbody radiation, so we use the Planck function for the incident intensity. Then, the energy density u_ν is

$$(u_\nu)_{\text{LTE}} = \frac{4\pi}{c} B_\nu(T) = \frac{8\pi h\nu^3}{c^3} \frac{1}{\exp(h\nu/kT) - 1} \tag{5.147}$$

And the total change in energy levels is

$$\frac{\mathrm{d}n_u}{\mathrm{d}t} = \left(\frac{\mathrm{d}n_u}{\mathrm{d}t}\right)_{\text{abs}} + \left(\frac{\mathrm{d}n_u}{\mathrm{d}t}\right)_{\text{spon}} + \left(\frac{\mathrm{d}n_u}{\mathrm{d}t}\right)_{\text{stim}} = n_\ell u_\nu B_{\ell u} - n_u(A_{u\ell} + u_\nu B_{u\ell}) \tag{5.148}$$

If the absorbers are in equilibrium with the radiation field, then the relative populations should be given by Boltzmann statistics: $n_u/n_\ell = (g_u/g_\ell)\exp(-(E_u - E_\ell)/kT)$ and $\mathrm{d}n_u/\mathrm{d}t = 0$. Therefore, we can solve for u_ν and compare it to what we would expect for LTE:

$$u_\nu \left(\frac{n_\ell}{n_u} B_{\ell u} - B_{u\ell}\right) = A_{u\ell} \tag{5.149}$$

$$u_\nu = \frac{A_{u\ell}/B_{u\ell}}{(g_\ell B_{\ell u}/g_u B_{u\ell})\exp(h\nu/kT) - 1} \tag{5.150}$$

Comparing this to (5.147), we can immediately see

$$\frac{A_{u\ell}}{B_{u\ell}} = \frac{8\pi h\nu^3}{c^3} \quad , \quad g_\ell B_{\ell u} = g_u B_{u\ell} \tag{5.151}$$

Therefore

$$\boxed{A_{u\ell} = \frac{8\pi h\nu^3}{c^3} B_{u\ell} = \frac{8\pi h\nu^3}{c^3} \frac{g_\ell}{g_u} B_{\ell u}} \tag{5.152}$$

QED[64]. Note that these relationships still hold even outside of thermal equilibrium!

Side note: Relating the Einstein coefficients to emission and absorption:

The spontaneous emission coefficient j_ν can be related to the Einstein coefficients by

$$j_\nu = \left(\frac{\mathrm{d}n_\ell}{\mathrm{d}t}\right)_{\text{spon}} \times \frac{h\nu}{4\pi} \times \phi_\nu \quad \longrightarrow \quad \boxed{j_\nu = \frac{h\nu}{4\pi} n_u A_{u\ell} \phi_\nu} \tag{5.153}$$

where ϕ_ν is the normalized line profile that defines the frequency dependence of j_ν—i.e. $\phi_\nu \mathrm{d}\nu$ is the probability that a photon is emitted within a frequency range $(\nu,\ \nu + \mathrm{d}\nu)$. If we integrate over the frequency ν and the line of sight ℓ, we can get a relationship for the upper level column density $N_u = \int \mathrm{d}\ell\, n_u$ based on the flux of a line

$$N_u = \frac{1}{A_{u\ell}} \frac{F}{h\nu} \frac{4\pi}{\Omega} \tag{5.154}$$

We can do something similar for absorption (including true absorption and stimulated emission) by definition the cross-section $\sigma_{\ell u}(\nu)$. The incident flux of photons is $cu_\nu/h\nu$, each with a cross-section

$\sigma_{\ell u}(\nu)$, thus

$$\left(\frac{\mathrm{d}n_u}{\mathrm{d}t}\right)_{\text{abs}} = n_\ell \int \mathrm{d}\nu\, \sigma_{\ell u}(\nu)\frac{cu_\nu}{h\nu} \quad \text{where} \quad \sigma_{\ell u}(\nu) = \frac{h\nu}{c}B_{\ell u}\phi_\nu \tag{5.155}$$

Where the ϕ_ν here is the same as above—essentially a delta-function that picks out a specific frequency, normalized such that integrating over ϕ_ν gives 1. Now we can define the absorption coefficient α_ν as

$$\left(\frac{\mathrm{d}n_u}{\mathrm{d}t}\right)_{\text{abs}} - \left(\frac{\mathrm{d}n_u}{\mathrm{d}t}\right)_{\text{stim}} = \int \mathrm{d}\nu\, \alpha_\nu \frac{cu_\nu}{h\nu} \quad \longrightarrow \quad \alpha_\nu = n_\ell \sigma_{\ell u}(\nu) - n_u \sigma_{u\ell}(\nu) \tag{5.156}$$

Thus, in terms of the B coefficients,

$$\boxed{\alpha_\nu = \frac{h\nu}{c}(n_\ell B_{\ell u} - n_u B_{u\ell})\phi_\nu} \tag{5.157}$$

Note that the B coefficients can oftentimes be defined in terms of the intensity I_ν instead of the energy density u_ν, which leads to a difference of a factor of $c/4\pi$ compared to our results here.

89. Explain quantitatively why stimulated emission is important and spontaneous emission is usually ignored in the radio domain, whereas the reverse is true in the optical domain. Given a thermal spectrum at some temperature T, at what frequency would the two emission rates be equal?

We can use the relationships derived in §5.Q88 to compare the rate of spontaneous emission to the rate of stimulated emission:

$$\mathscr{F} = \frac{(\mathrm{d}n_\ell/\mathrm{d}t)_{\text{spon}}}{(\mathrm{d}n_\ell/\mathrm{d}t)_{\text{stim}}} = \frac{n_u A_{u\ell}}{n_u u_\nu B_{u\ell}} = \frac{8\pi h\nu^3}{c^3}\frac{1}{u_\nu} = e^{h\nu/kT} - 1 \tag{5.158}$$

Therefore, in the radio domain where $h\nu \ll kT$, we can approximate $\mathscr{F} \approx h\nu/kT \ll 1$. In other words, stimulated emission dominates over spontaneous emission. On the flip side, in the optical $h\nu \gg kT$, so $\mathscr{F} \gg 1$ and spontaneous emission dominates over stimulated emission.

It is easy to find where they are equal by just setting $\mathscr{F} = 1$:

$$e^{h\nu/kT} = 2 \quad\longrightarrow\quad \nu = \frac{kT}{h}\ln 2 \tag{5.159}$$

90. Name five molecules found in the interstellar medium and comment on how they are detected. What limits our ability to directly detect the most common molecules in the ISM?

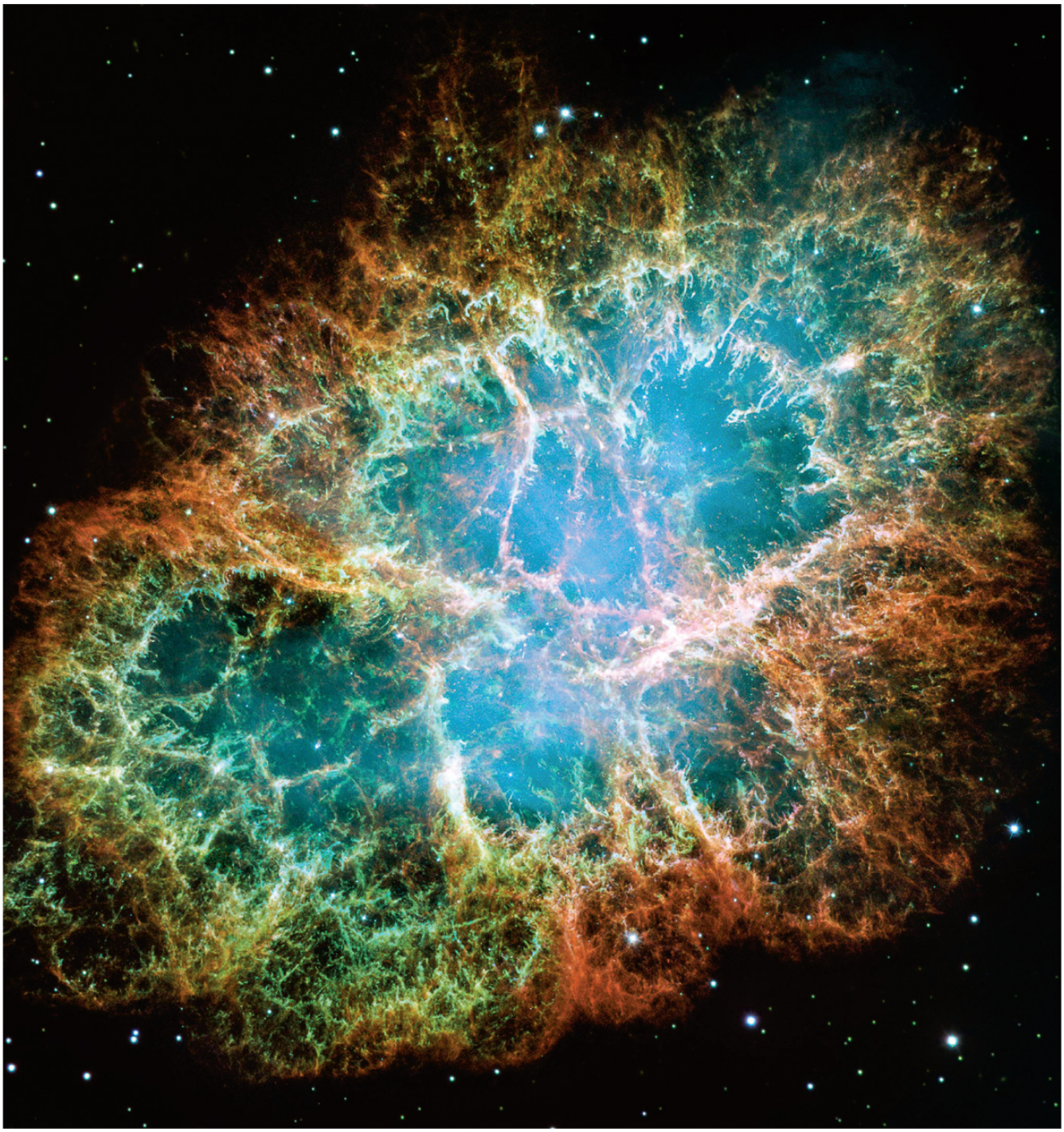
1. **H_2 (molecular hydrogen)**: By far the most abundant molecule in the ISM, found within dense molecular clouds with temperatures $\sim$ 10 K. Unfortunately, oftentimes H_2 is not directly observable. It has many lines from electronic transitions in the FUV, but since molecular clouds are dusty and have large column densities, we cannot see these lines. It also has lines from vibrational and rotational energy transitions in the IR, but the energies of these transitions are high enough that they can't be excited in molecular clouds with temperatures $\sim$ 10 K. Therefore, we have to use tracers from other molecules like CO and infer an H_2-to-CO conversion factor, which is highly uncertain[65].
2. **CO (carbon monoxide)**: The next most abundant molecular after H_2 (with a relative abundance of about 10^{-4}). Its lowest energy transition, CO 1-0, is at 2.6 mm and is easily excitable at $h\nu/k = 5.5$ K. There are still some complications with observing CO at 2.6 mm though. The CMB is bright at 2.6 mm, and the line tends to be optically thick. At low densities, the excitation temperature equals the temperature of the CMB, meaning we cannot see the line at all[65]!
3. **CH (methylidyne radical)**: One of the first interstellar molecules to be identified, because it has a resonance at optical wavelengths. Also has transitions in the radio at $\sim$ 3.33 GHz and absorption bands in the infrared at $\sim$ 3.4 μm.
4. **CN (cyano radical)**: Very similar to CH, this was one of the first identified molecules in the ISM because of its resonances in the optical. Has a rotational transition CN 1-0 at 113 GHz.
5. **OH (hydroxl radical)**: One of the first molecules to be discovered using its radio emission—it has transitions at 1.665 and 1.667 GHz.

Bonus round:

6. **CH^+ (methylidyne ion)**: Like CH and CN, this is resonant at optical wavelengths and was one of the first discovered molecular species in the ISM.
7. **NH_3 (ammonia)**: Detectable in the radio from its inversion transitions at 23.694 and 23.723 GHz (caused by the N atom tunneling back and forth on opposite sides of the plane defined by the 3 H atoms).
8. **H_2O (dihydrogen monoxide[5])**: Has a strong absorption feature in the IR at 3.1 μm from stretching of the H-O bonds, as well as a rotational transition in the radio at 22.2 GHz.
9. **H_2CO (formaldehyde)**: Observed with a rotational transition at 4830 MHz.
10. **HCO^+ (formyl cation)**: Observed in the radio at 89.190 GHz.
11. **HCN (hydrogen cyanide)**: Very common, observed with the HCN 1-0 transition at 88.3 and 88.6 GHz (from two different isotopes).
12. **SiO (silicon monoxide)**: Has a SiO 3-2 transition at 130.246 GHz.
13. **CS (carbon monosulfide)**: Has a CS 3-2 transition at 146.969 GHz.
14. **And many more!** (CH_3OH, HCOOH, CH_3CN, OCS, NH_2CHO, HNC, H_2S, etc.) See this comprehensive list with cool diagrams of each molecule.

[5]Beware of this highly toxic molecule! Everyone who has ingested it has died!

6 Plasma & Gravitational Lensing



91. What is "Faraday rotation"? How is it used in astronomy?

If a linearly polarized electromagnetic wave propagates through a plasma with a static magnetic field B parallel to the direction of motion, this causes the polarization vector (i.e. the direction the electric field oscillates in) to rotate, as shown in Figure 6.1. This is called **Faraday rotation**.

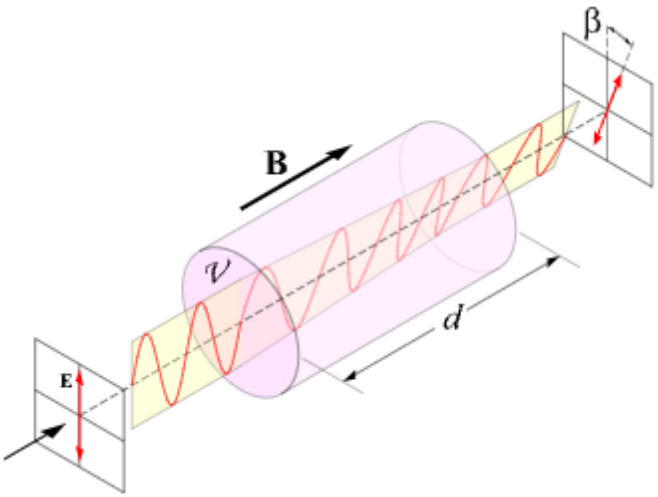


Figure 6.1: The Faraday rotation of a linearly polarized EM wave.

This happens because right circularly polarized waves travel at a different speed than left circularly polarized waves inside the plasma. Since linearly polarized waves can be decomposed into a superposition of equal amounts left and right circular polarization, the different propagation speeds causes a phase shift to build up between the left and right components, which causes the polarization direction to rotate. The dispersion relation for circularly polarized waves in a plasma with a magnetic field is

$$c^2k^2 = \omega^2 - \frac{\omega_p^2}{1 \pm \omega_B/\omega} \tag{6.1}$$

where $\omega_p = (4\pi n_e e^2/m_e)^{1/2}$ is the **plasma frequency** (see §6.Q93) and $\omega_B = eB_\parallel/m_e c$ is the **cyclotron frequency**. This leads to a phase velocity of

$$v_p(\omega) = \frac{\omega}{k(\omega)} \approx c\left[1 + \frac{1}{2}\frac{\omega_p^2}{\omega^2} \mp \frac{1}{2}\frac{\omega_p^2\omega_B}{\omega^3}\right] \tag{6.2}$$

The $+$ is taken for right circular polarization and the $-$ is for left circular polarization. After propagating a distance L, the left and right polarizations differ in phase by $L\Delta k$, so the rotation angle becomes

$$\beta = \frac{1}{2}\int_0^L \mathrm{d}L'\,\Delta k = \int_0^L \mathrm{d}L'\,\frac{\omega_p^2\omega_B}{2c\omega^2} \tag{6.3}$$

$$= \frac{e^3}{2\pi m_e^2 c^2}\frac{1}{\nu^2}\int_0^L \mathrm{d}L'\,n_e B_\parallel \tag{6.4}$$

Therefore, we usually define the **rotation measure** $\mathscr{R}$ such that

$$\beta = \mathscr{R}\lambda^2 \tag{6.5}$$

$$\boxed{\mathscr{R} = \frac{e^3}{2\pi m_e^2 c^4}\int_0^L \mathrm{d}L'\, n_e B_\parallel} \tag{6.6}$$

We typically don't know the polarization direction at the source, so we can't measure β directly, but we can measure the difference between β at two different wavelengths, which allows us to find the rotation measure:

$$\mathscr{R} = \frac{\beta_2 - \beta_1}{\lambda_2^2 - \lambda_1^2} \tag{6.7}$$

Then, if we also have the dispersion measure $\mathscr{D}$ (see §6.Q92), we can get the electron-density-weighted magnetic field

$$\langle B_\parallel \rangle = \frac{\int_0^L \mathrm{d}L'\, n_e B_\parallel}{\int_0^L \mathrm{d}L'\, n_e} = \frac{2\pi m_e^2 c^4}{e^3}\frac{\mathscr{R}}{\mathscr{D}} \tag{6.8}$$

And the actual magnetic field can be measured if we use observations of pulsars at close locations on the sky but at different radial distances, since the difference in the rotation/dispersion measures will purely be due to the difference in $n_e(L)$.

$$B_\parallel = \frac{2\pi m_e^2 c^4}{e^3}\frac{\Delta\mathscr{R}}{\Delta\mathscr{D}} \tag{6.9}$$

We can see why this is more clearly by taking the differential limit:

$$\mathrm{d}\mathscr{R} = \frac{e^3}{2\pi m_e^2 c^4} n_e B_\parallel \mathrm{d}L \tag{6.10}$$

$$\mathrm{d}\mathscr{D} = n_e \mathrm{d}L \tag{6.11}$$

$$\therefore \quad B_\parallel = \frac{2\pi m_e^2 c^4}{e^3}\frac{\mathrm{d}\mathscr{R}}{\mathrm{d}\mathscr{D}} \tag{6.12}$$

See more in Ref. 64.

92. What is the "dispersion measure" for electromagnetic waves in a plasma? How is it used in astronomy?

The dispersion relation for electromagnetic plane waves propagating in a plasma with *no* external magnetic field is

$$c^2k^2 = \omega^2 - \omega_p^2 \tag{6.13}$$

where ω_p is the plasma frequency (see §6.Q93). This leads to a phase velocity and a group velocity of

$$v_p(\omega) = \frac{\omega}{k} = c\left(1 - \frac{\omega_p^2}{\omega^2}\right)^{-1/2} > c \tag{6.14}$$

$$v_g(\omega) = \frac{\mathrm{d}\omega}{\mathrm{d}k} = c\left(1 - \frac{\omega_p^2}{\omega^2}\right)^{1/2} < c \tag{6.15}$$

This causes the arrival time of light to be delayed depending on the frequency ω. Say a pulsar emits a pulse at time $t = 0$ a distance L away from us. The time we receive the pulse will be

$$t = \int_0^L \frac{\mathrm{d}L'}{v_g(\omega)} \approx \int_0^L \frac{\mathrm{d}L'}{c}\left(1 + \frac{1}{2}\frac{\omega_p^2}{\omega^2}\right) \tag{6.16}$$

$$= \frac{L}{c} + \frac{1}{2c\omega^2}\int_0^L \mathrm{d}L'\,\omega_p^2 \tag{6.17}$$

Then, the delay time $\Delta t = t - L/c$ can be written in terms of the **dispersion measure** $\mathscr{D}$

$$\Delta t = \frac{e^2}{2\pi m_e c}\frac{1}{\nu^2}\mathscr{D} \tag{6.18}$$

$$\boxed{\mathscr{D} = \int_0^L \mathrm{d}L'\,n_e} \tag{6.19}$$

Notice that $\Delta t \propto \nu^{-2}$, i.e. lower frequencies have longer delay times. This effect becomes more prominent the closer ω is to ω_p, so it's only really noticeable in the radio. We can measure the difference in arrival times at different frequencies to find the dispersion measure:

$$\mathscr{D} = -\frac{\pi m_e c}{e^2}\nu^3\frac{\mathrm{d}t}{\mathrm{d}\nu} \tag{6.20}$$

Using this, in combination with an independent distance measurement, can be used to get the average electron density along a line of sight:

$$\langle n_e \rangle = \frac{\mathscr{D}}{L} \tag{6.21}$$

The full 3D electron density $n_e(x, y, z)$ can be mapped out by applying this techinque to many Galactic pulsars. Typical models include a 4-arm logarithmic spiral pattern along with a ~ 3.5 kpc ring-like density enhancement around the galactic center. There also tends to be more substructure around the solar region since we can more easily resolve this area (see Figure 6.2)[64].

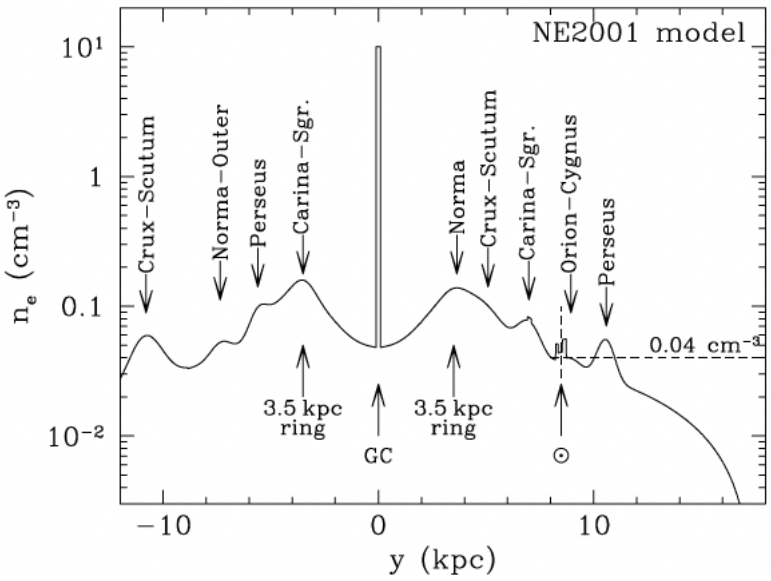


Figure 6.2: The electron density as a function of distance from the galactic center along the galactic midplane ($z = 0$)[64,111].

Fast Radio Bursts (FRBs) are fast and energetic burst events, similar to GRBs (but in the radio obviously). Their origins are still not yet understood. The current leading hypothesis is that they are caused by mergers of compact objects—the bursts only last a few milliseconds, so the size of the objects is limited by the speed of light and cannot be much larger than ∼1000 km. They are also strongly polarized, indicating the presence of a strong magnetic field (i.e. magnetars). Since they are in the radio, we can get their dispersion measures, which cause time delays on the order of a few seconds. They tend to be much larger than the dispersion measures of Galactic pulsars, and they are uniform over the sky, suggesting an extragalactic origin (see Figure 6.3).

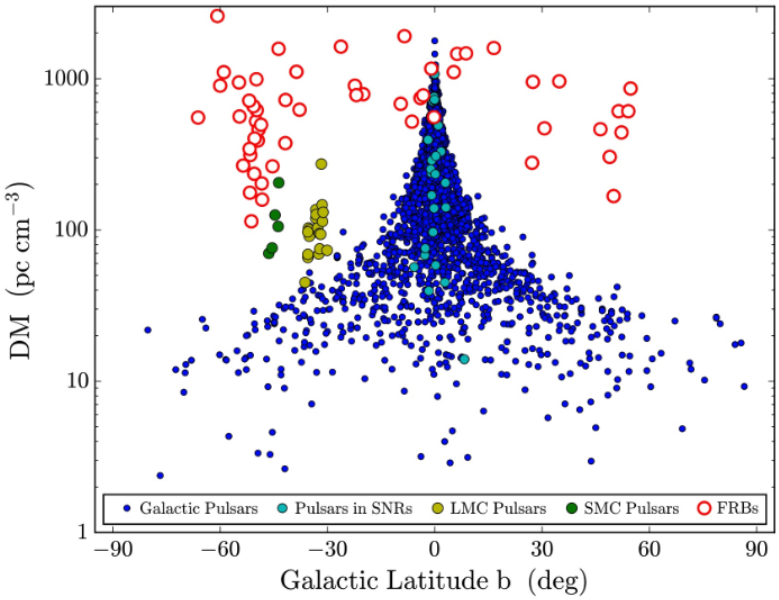


Figure 6.3: The dispersion measures of various astrophysical events, highlighting the contrasts between galactic pulsars and FRBs[112].

93. What is the physical significance of the plasma frequency?

Consider a uniform plasma with electron density n_e. The charge is locally neutral, $\rho = 0$. An electromagnetic plane wave $\mathbf{E}(\mathbf{r}, t)$ with frequency ω propagates in the plasma. Treating the problem classically, the wave induces a force on each electron

$$m_e \ddot{\mathbf{r}} = -e\mathbf{E}(\mathbf{r}, t) \tag{6.22}$$

The magnetic force is much smaller and can be neglected. Since $\mathbf{E}$ is a wave, the solutions for $\mathbf{r}$, the position of the electron, must also be waves with the same frequency:

$$\ddot{\mathbf{r}} = -\omega^2 \mathbf{r} \quad \longrightarrow \quad m_e \omega^2 \mathbf{r} = e\mathbf{E} \tag{6.23}$$

Thus, the velocity of the electron is

$$\mathbf{v} = \frac{e}{m_e \omega^2} \frac{\partial \mathbf{E}}{\partial t} \tag{6.24}$$

(we could've also gotten this by actually doing the integration of (6.22), but this argument is nice). The movement of charged electrons creates a current density

$$\mathbf{J} = -en_e \mathbf{v} = -\frac{e^2 n_e}{m_e \omega^2} \frac{\partial \mathbf{E}}{\partial t} \tag{6.25}$$

And now we can use Maxwell's equations to find the wave equation that describes the propagation of $\mathbf{E}$ through this plasma. Maxwell's equation are

$$\nabla \cdot \mathbf{E} = 4\pi\rho \qquad \text{(Gauss's Law)} \tag{6.26}$$

$$\nabla \cdot \mathbf{B} = 0 \qquad \text{(No magnetic monopoles)} \tag{6.27}$$

$$\nabla \times \mathbf{E} = -\frac{1}{c} \frac{\partial \mathbf{B}}{\partial t} \qquad \text{(Faraday's Law)} \tag{6.28}$$

$$\nabla \times \mathbf{B} = \frac{4\pi}{c} \mathbf{J} + \frac{1}{c} \frac{\partial \mathbf{E}}{\partial t} \qquad \text{(Ampère's Law)} \tag{6.29}$$

We take the curl of Faraday's law and switch the order of the spatial and time derivatives

$$\nabla \times (\nabla \times \mathbf{E}) = -\frac{1}{c} \nabla \times \frac{\partial \mathbf{B}}{\partial t} = -\frac{1}{c} \frac{\partial}{\partial t} \nabla \times \mathbf{B} \tag{6.30}$$

On the LHS, we use a vector identity, and on the RHS, we plug in Ampère's law

$$\nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E} = -\frac{1}{c} \frac{\partial}{\partial t} \left(\frac{4\pi}{c} \mathbf{J} + \frac{1}{c} \frac{\partial \mathbf{E}}{\partial t} \right) \tag{6.31}$$

By Gauss's law with $\rho = 0$, we have $\nabla \cdot \mathbf{E} = 0$, so the first term on the LHS goes to 0. Then we plug in equation (6.25) on the RHS

$$\boxed{\nabla^2 \mathbf{E} = \frac{1}{c^2} \left(1 - \frac{4\pi e^2 n_e}{m_e \omega^2} \right) \frac{\partial^2 \mathbf{E}}{\partial t^2}} \tag{6.32}$$

This is a wave equation, as we should've expected, but it does *not* travel at c. Comparing to a typical wave equation, we see that this wave must have an index of refraction

$$n(\omega) = \left(1 - \frac{\omega_p^2}{\omega^2}\right)^{1/2} \tag{6.33}$$

where I have defined the **plasma frequency** as

$$\boxed{\omega_p^2 \equiv \frac{4\pi e^2 n_e}{m_e}} \tag{6.34}$$

And we have a dispersion relation of

$$\omega = \frac{c}{n}k \quad \longrightarrow \quad \omega^2 n^2 = c^2 k^2 \quad \longrightarrow \quad \boxed{c^2 k^2 = \omega^2 - \omega_p^2} \tag{6.35}$$

The wave has a (phase) velocity of

$$v_p(\omega) = \frac{c}{n} = \frac{c}{\sqrt{1 - \omega_p^2/\omega^2}} \tag{6.36}$$

And a group velocity of

$$v_g(\omega) = \frac{\mathrm{d}\omega}{\mathrm{d}k} = \frac{\mathrm{d}}{\mathrm{d}k}\sqrt{c^2k^2 + \omega_p^2} = \frac{1}{2}(c^2k^2 + \omega_p^2)^{-1/2} \cdot 2c^2 k \tag{6.37}$$

$$= \frac{c^2 k}{\omega} = cn \tag{6.38}$$

$$= c\sqrt{1 - \omega_p^2/\omega^2} \tag{6.39}$$

Physical significance

The plasma frequency ω_p essentially represents how quickly the charges in the plasma can react to external stimuli. Thus, if incident light has a frequency slower than the plasma frequency $\omega < \omega_p$, the plasma can react in time and creates oscillations that damp and reflect the wave (k is imaginary, leading to evanescent waves). Alternatively, if the incident light has a frequency much faster than the plasma frequency ($\omega \gg \omega_p$), the plasma cannot react and the light will pass through essentially unaffected. For intermediate frequencies, the plasma's reaction partially damps the incident waves, but still lets them through, creating a delay in the arrival time dependent on the frequency (since $v_g(\omega) < c$).

94. What are the three types of MHD waves in a magnetized plasma? Are magnetic fields important in the propagation of waves in the interstellar medium? In a star?

Magnetohydrodynamics (MHD) couples Maxwell's equations with hydrodynamics to describe the macroscopic behaviors of highly conducting fluids (like plasmas). The ideal equations of MHD are

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \qquad \text{(charge continuity)} \tag{6.40}$$

$$\frac{\partial \rho}{\partial t}\mathbf{v} + (\rho \mathbf{v} \cdot \nabla)\mathbf{v} = \frac{1}{c}\mathbf{J} \times \mathbf{B} - \nabla P \qquad \text{(momentum continuity, Jeans equation)} \tag{6.41}$$

$$\nabla \times (\mathbf{v} \times \mathbf{E}) = \frac{\partial \mathbf{B}}{\partial t} \qquad \text{(Faraday \& ideal Ohm's law)} \tag{6.42}$$

$$\frac{\partial P}{\partial t} + (\mathbf{v} \cdot \nabla)P = -\gamma \rho \nabla \cdot \mathbf{v} \qquad \text{(adiabatic energy equation)} \tag{6.43}$$

We can derive a dispersion relation similar to how we did in §6.Q93. First we *linearlize* these equations using small perturbations from some equilibrium state, just like we did in §8.Q130 for the Jeans equation. The perturbations will be labeled with "1" subscripts and the backgrounds will be labeled with "0" subscripts. The equations become

$$\frac{\partial \rho_1}{\partial t} = -\mathbf{v}_1 \times \nabla \rho_0 - \rho_0 \nabla \cdot \mathbf{v}_1 \tag{6.44}$$

$$\rho_0 \frac{\partial \mathbf{v}_1}{\partial t} = \frac{1}{4\pi}(\nabla \times \mathbf{B}_1) \times \mathbf{B}_0 - \nabla P_1 \tag{6.45}$$

$$\frac{\partial \mathbf{B}_1}{\partial t} = \nabla \times (\mathbf{v}_1 \times \mathbf{B}_0) \tag{6.46}$$

$$\frac{\partial P_1}{\partial t} = -\mathbf{v}_1 \times \nabla P_0 - \gamma P_0 \nabla \times \mathbf{v}_1 \tag{6.47}$$

Then we define the *displacement vector* $\boldsymbol{\xi}$ as how much the plasma is displaced from its equilibrium:

$$\boldsymbol{\xi}(\mathbf{r}, t) \equiv \int_0^t \mathbf{v}_1(\mathbf{r}, t')\mathrm{d}t' \quad \longrightarrow \quad \frac{\partial \boldsymbol{\xi}}{\partial t} = \mathbf{v}_1(\mathbf{r}, t) \tag{6.48}$$

We can then rewrite the linearlized equations in terms of $\partial \boldsymbol{\xi} / \partial t$ and integrate over time to get ρ_1, $\mathbf{B}_1$, and P_1 in terms of $\boldsymbol{\xi}$. Then the momentum equation becomes

$$\rho_0 \frac{\partial^2 \boldsymbol{\xi}}{\partial t^2} = \hat{\mathbf{F}}[\boldsymbol{\xi}(\mathbf{r}, t)] \tag{6.49}$$

$$\begin{aligned} \hat{\mathbf{F}}(\boldsymbol{\xi}) = \nabla(\boldsymbol{\xi} \cdot \nabla P_0 + \gamma P_0 \nabla \cdot \boldsymbol{\xi}) + \frac{1}{4\pi}(\nabla \times \mathbf{B}_0) \times [\nabla \times (\boldsymbol{\xi} \times \mathbf{B}_0)] \\ + \frac{1}{4\pi}\{[\nabla \times \nabla \times (\boldsymbol{\xi} \times \mathbf{B}_0)] \times \mathbf{B}_0\} \end{aligned} \tag{6.50}$$

where $\hat{\mathbf{F}}[\boldsymbol{\xi}(\mathbf{r}, t)]$ is the *ideal MHD force operator*. Notice the similarities between (6.49) and (6.22). The derivation from here proceeds in a similar way, yielding a dispersion relation of[113,114,115,116]

$$\boxed{(\omega^2 - k_\parallel^2 v_A^2)[\omega^4 - k^2(c_s^2 + v_A^2)\omega^2 + k^2 k_\parallel^2 c_s^2 v_A^2] = 0} \tag{6.51}$$

Here, $k_\parallel = k\cos\theta$ is the component of k in the direction of $\mathbf{B}_0$, $c_s = \sqrt{\partial P/\partial\rho}$ is the sound speed, and

$$v_A = \sqrt{\frac{B_0^2}{4\pi\rho_0}} \tag{6.52}$$

is the **Alfvén velocity**. This dispersion relation has a few different solutions, each of which corresponds to one type of MHD wave.

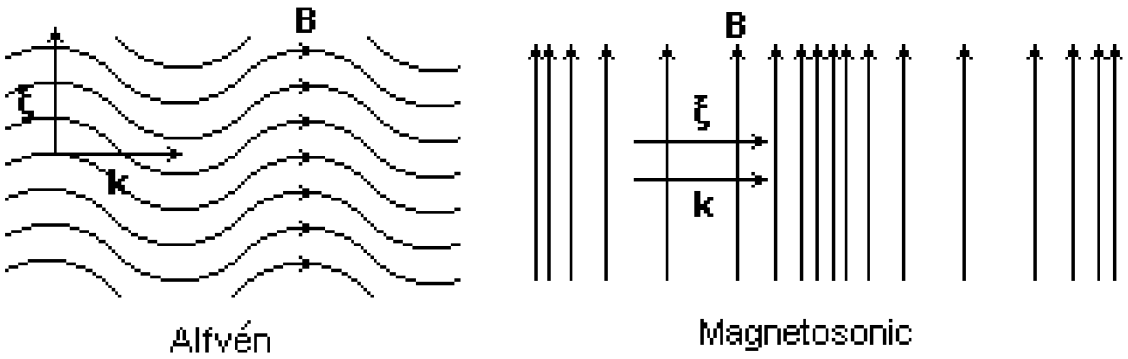


Figure 6.4: Left: Alfvén waves. Right: Magnetostatic waves.

1. Shear Alfvén Waves[115]

These are the waves for which the first term in the dispersion relation goes to 0.

$$\boxed{\omega^2 = k_\parallel^2 v_A^2} \tag{6.53}$$

These waves propagate parallel to $\mathbf{B}_0$ and have a displacement $\boldsymbol{\xi}$ orthogonal to both $\mathbf{B}_0$ and $\mathbf{k}$. Therefore, they are *transverse* waves (just like normal EM waves). The restoring force for these waves is magnetic tension. See the left panel of Figure 6.4.

2. & 3. Slow/Fast Magnetostatic Waves[115]

These are the waves for which the second term in the dispersion relation goes to 0 (it's just a quadratic equation for ω^2).

$$\boxed{\omega^2 = \frac{1}{2}k^2(c_s^2 + v_A^2)\left[1 \pm \sqrt{1 - \frac{4k_\parallel^2 c_s^2 v_A^2}{k^2(c_s^2 + v_A^2)^2}}\right]} \tag{6.54}$$

These waves propagate orthogonal to $\mathbf{B}_0$, and they have a displacement $\boldsymbol{\xi}$ that is *parallel* to $\mathbf{k}$ and *orthogonal* to $\mathbf{B}_0$, making them *longitudianl* waves (like sound waves, except they are within the magnetic field). The restoring force here comes from a combination of magnetic pressure and plasma pressure. See the right panel of Figure 6.4.

The difference between fast and slow magnetostatic waves, other than the obvious speed difference, is that fast waves have magnetic and plasma pressure perturbations *in phase*, while slow waves have

these two *out of phase*.

Astrophysical Importance[115]

- In the ISM:

  MHD waves in the ISM can affect Faraday rotation (see §6.Q91).

- In stars:

  Alfvén waves are the primary mechanism for heating the stellar corona and accelerating stellar winds. Additionally, counter-propagating waves are important in the development of turbulence.

- In shocks and SN remnants:

  Accelerated particles around supernova remnant shock waves generate Alfvén waves, producing “stripe”-like features in X-rays.

95. What is the "ideal Ohm's law" for a plasma? What does "frozen-in" mean?

The familiar version of Ohm's law for a steady current ($\partial/\partial t \to 0$) reads

$$V = IR \ , \ \mathbf{J} = \sigma \mathbf{E} \tag{6.55}$$

where σ is the electrical conductivity. However, if we are in a plasma that is moving relative to some external magnetic field, we also need to add in a term for the magnetic contribution:

$$\mathbf{J} = \sigma\left(\mathbf{E} + \frac{\mathbf{v}}{c} \times \mathbf{B}\right) \tag{6.56}$$

The original form of Ohm's law still works in the *rest frame* of the plasma, because the electric field experienced by the plasma in this frame is different, i.e. $\mathbf{E}' = \gamma(\mathbf{E} + (\mathbf{v}/c) \times \mathbf{B})$.

For an ideal plasma, we assume $\sigma \to \infty$, so to retain a finite current, we need

$$\boxed{\mathbf{E} + \frac{\mathbf{v}}{c} \times \mathbf{B} = 0} \tag{6.57}$$

This is the **ideal Ohm's law**[114].

As a consequence of this, it can be shown that the magnetic flux Φ_B through any surface moving with the plasma (at fluid velocity $\mathbf{v}$) remains constant. At some time t, we have

$$\Phi_B(t) = \int_S \mathbf{B}(t) \cdot \mathrm{d}\mathbf{A}(t) \tag{6.58}$$

A short Δt later, the magnetic field will be slightly different, and the area will also be slightly different (it will have expanded or shrunk by a small ribbon depending upon how the velocity is oriented):

$$\mathbf{B}(t + \Delta t) \approx \mathbf{B}(t) + \frac{\partial \mathbf{B}}{\partial t}\Delta t \tag{6.59}$$

$$\mathrm{d}\mathbf{A}(t + \Delta t) \approx \mathrm{d}\mathbf{A}(t) + \mathbf{v}\Delta t \times \mathrm{d}\boldsymbol{\ell} \tag{6.60}$$

Therefore, our new flux is

$$\Phi_B(t + \Delta t) = \int_S \mathbf{B}(t + \Delta t) \cdot \mathrm{d}\mathbf{A}(t + \Delta t) \approx \int_S \mathbf{B}(t + \Delta t) \cdot \mathrm{d}\mathbf{A}(t) + \int_S \mathbf{B}(t + \Delta t) \cdot (\mathbf{v}\Delta t \times \mathrm{d}\boldsymbol{\ell}) \tag{6.61}$$

Throwing away all higher order terms, this reduces to

$$\Phi_B(t + \Delta t) \approx \int_S \mathbf{B}(t) \cdot \mathrm{d}\mathbf{A}(t) + \Delta t \int_S \frac{\partial \mathbf{B}}{\partial t} \cdot \mathrm{d}\mathbf{A}(t) + \Delta t \int_S \mathbf{B}(t) \cdot (\mathbf{v} \times \mathrm{d}\boldsymbol{\ell}) \tag{6.62}$$

The first term is just $\Phi_B(t)$, and for the last term we can use the vector identity $\mathbf{A} \cdot \mathbf{B} \times \mathbf{C} = \mathbf{A} \times \mathbf{B} \cdot \mathbf{C}$

$$\Phi_B(t + \Delta t) = \Phi_B(t) + \Delta t \int_S \frac{\partial \mathbf{B}}{\partial t} \cdot \mathrm{d}\mathbf{A}(t) + \Delta t \int_S (\mathbf{B}(t) \times \mathbf{v}) \cdot \mathrm{d}\boldsymbol{\ell} \tag{6.63}$$

Now we can use Stokes' theorem on the last term to convert it to an integral over the area $\mathrm{d}\mathbf{A}(t)$.

Then it can be combined with the second term.

$$\Phi_B(t+\Delta t) = \Phi_B(t) + \Delta t \int_S \left[\frac{\partial \mathbf{B}}{\partial t} - \nabla \times (\mathbf{v} \times \mathbf{B})\right] \cdot d\mathbf{A}(t) \tag{6.64}$$

We're almost done. Now we just have to use Faraday's law to convert the derivative of the magnetic field into the curl of the electric field:

$$\Phi_B(t+\Delta t) = \Phi_B(t) + \Delta t \int_S \nabla \times (c\mathbf{E} - \mathbf{v} \times \mathbf{B}) \cdot d\mathbf{A}(t) \tag{6.65}$$

And we see that if the ideal Ohm's law holds, the second term must go to 0, meaning

$$\Phi_B(t+\Delta t) = \Phi_B(t) \longrightarrow \boxed{\frac{\partial \Phi_B}{\partial t} = 0} \tag{6.66}$$

As a consequence of this, along a magnetic field line, the flux coming into one end must equal the flux coming out of the other end—there can be no flux escaping along the sides of the "flux tube." If the tube moves with a fluid velocity $\mathbf{v}(\mathbf{x}, t)$, this must still hold, so the magnetic field lines must move with the fluid to conserve the flux. In other words, the magnetic field lines are **frozen-in** with the fluid flow[117]. This is shown in Figure 6.5.

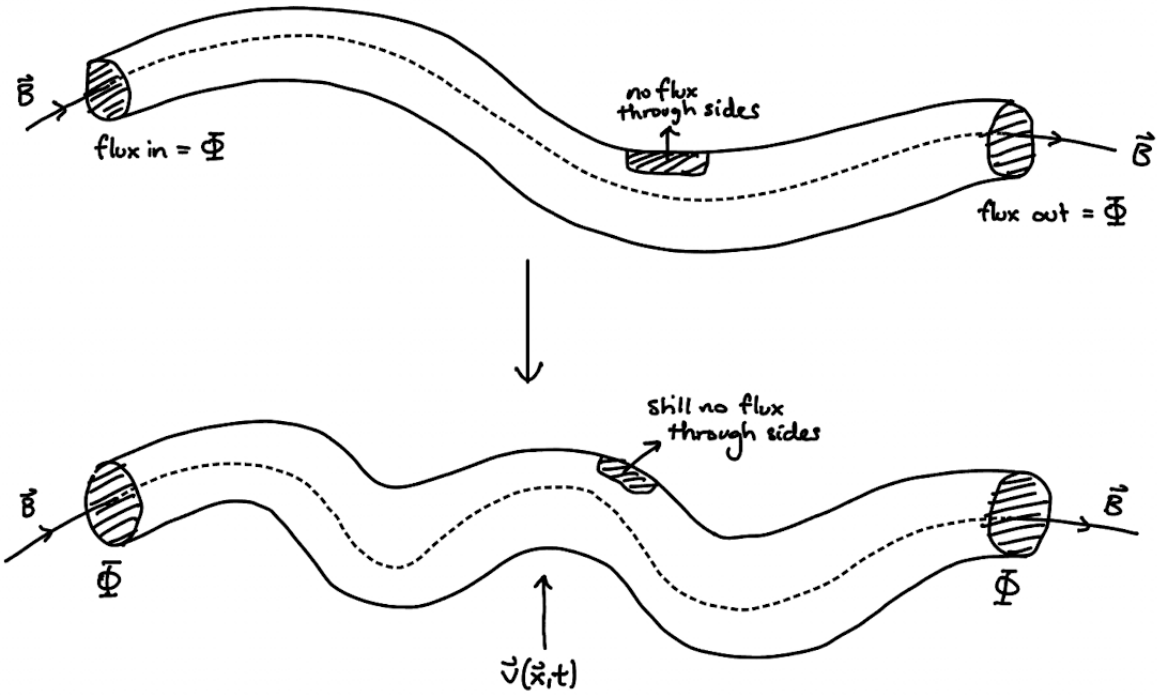


Figure 6.5: The conservation of magnetic flux along a flux tube in a moving plasma. The magnetic field lines have to move with the plasma in order for the flux to continue being conserved. We can take the limit of this flux tube for arbitrarily small radii to enforce this condition rigorously.

96. Qualitatively, how does a gravitational lens work? Explain what needs to be measured to use a gravitational lens system to measure the Hubble constant. What is the current status of these determinations of H_0?

Particles passing through a gravitational field will be deflected due to the gravitational force acting on them. As it turns out, even though light has no mass, light can also be deflected in the presence of gravitational fields. This makes perfect sense if we consider gravity as the curvature of spacetime rather than a force, as it is described in general relativity. Light follows "straight line" paths in curved spacetime called **geodesics**. We perceive objects as being in the direction that their light comes from, so when light is bent by strong gravitational fields, it has the effect of creating images of the object at different locations that are not the true location of the object (and depending on the geometry, sometimes even multiple images of the same object can be created).

Say we have a source a distance D_s from us, and in between us and the source, there is a lens at a distance D_ℓ from us. The distance between the lens and the source is $D_{\ell s} = D_s - D_\ell$. Light rays leave the source and reach the lens plane at some impact parameter ξ, at which point they get deflected into our line of sight by an angle $\hat{\alpha}$. The "true" angle between the lens and the source is β, and the angle between the true location of the source and the image of the source is α. This geometry is all shown in Figure 6.6.

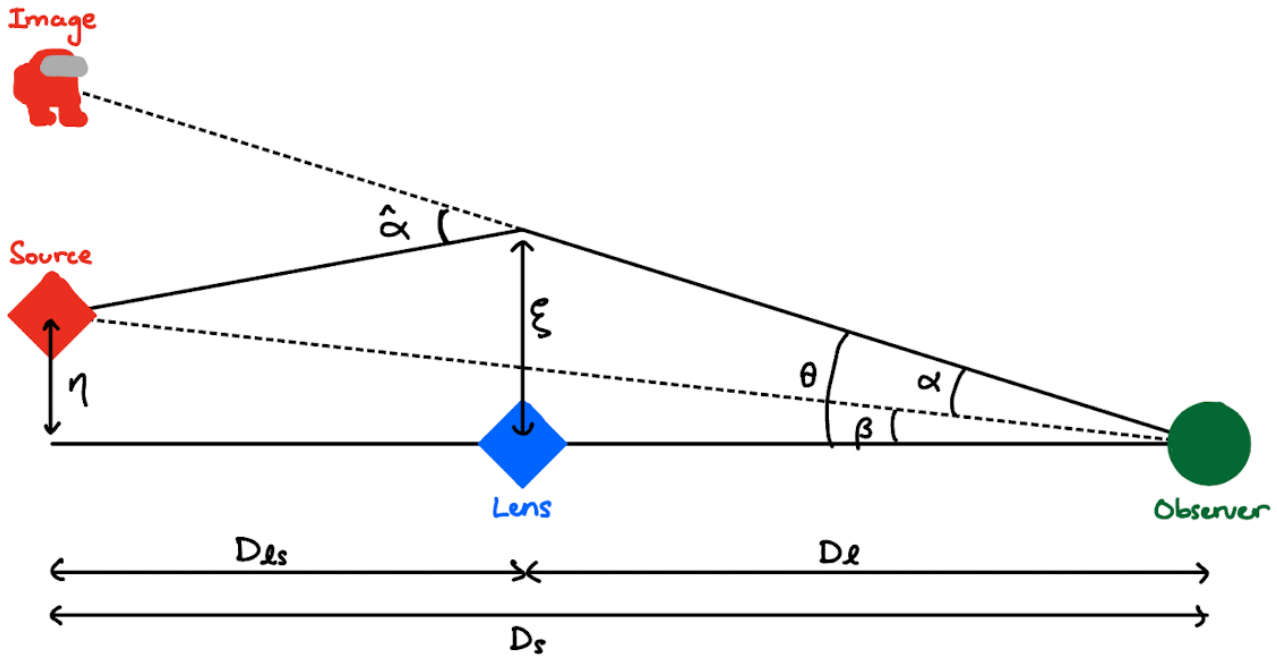


Figure 6.6: The geometry set-up for gravitational lensing due to a point mass.

How much will the light be deflected by? Let's consider a simple example where the lens is a point mass with mass M, and light travels parallel to the lens at a speed c. We'll make a classical argument, which requires that we assume photons have some small nonzero mass, but the result will be independent of this mass. If the parallel distance between the light and the lens is x, then it will feel

an acceleration

$$a = -\frac{GM}{x^2 + \xi^2} \tag{6.67}$$

The component of this acceleration in the perpendicular direction will then be

$$a_\perp = -\frac{GM}{x^2 + \xi^2} \cos\theta = \frac{GM\xi}{(x^2 + \xi^2)^{3/2}} \tag{6.68}$$

Then we can integrate this to obtain the total change in the perpendicular velocity

$$\Delta v_\perp = -\int_0^t \mathrm{d}t' \, \frac{GM\xi}{(x^2 + \xi^2)^{3/2}} \tag{6.69}$$

Assuming $c\mathrm{d}t' \approx \mathrm{d}x$, then

$$\Delta v_\perp = -\frac{1}{c}\int_0^{ct} \mathrm{d}x \, \frac{GM\xi}{(x^2 + \xi^2)^{3/2}} = -\frac{2GM}{c\xi} \tag{6.70}$$

And the deflection angle becomes $\Delta v_\perp / v = \Delta v_\perp / c$:

$$\hat{\alpha} = \frac{2GM}{c^2\xi} \tag{6.71}$$

As it turns out, this is only off by a factor of 2 compared to if we had done a full calculation with GR[118]:

$$\boxed{\hat{\alpha} = \frac{4GM}{c^2\xi}} \tag{6.72}$$

How will we know if a light ray from the source will reach us or not? We can use the lens equation. The deflection angle must satisfy the condition

$$\boldsymbol{\eta} = \frac{D_s}{D_\ell}\boldsymbol{\xi} - D_{\ell s}\hat{\boldsymbol{\alpha}}(\boldsymbol{\xi}) \tag{6.73}$$

Note that I am now treating $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, $\hat{\boldsymbol{\alpha}}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$ all as 2-dimensional vectors, since in principle the source and lens can be oriented in any direction on the sky relative to each other. We can then divide by D_s to obtain

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \frac{D_{\ell s}}{D_s}\hat{\boldsymbol{\alpha}}(D_\ell\boldsymbol{\theta}) \tag{6.74}$$

Then if we define the *reduced deflection angle* as

$$\boxed{\boldsymbol{\alpha}(\boldsymbol{\theta}) = \frac{D_{\ell s}}{D_s}\hat{\boldsymbol{\alpha}}(D_\ell\boldsymbol{\theta})} \tag{6.75}$$

We get the **lens equation**

$$\boxed{\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta})} \tag{6.76}$$

If this equation has multiple solutions $\boldsymbol{\theta}_i$ for the same intrinsic angle $\boldsymbol{\beta}$, we get multiple images! In the case where the lens is a point mass, we can plug in the deflection angle to get

$$\boxed{\boldsymbol{\alpha}(\boldsymbol{\theta}) = \frac{4GM}{c^2}\frac{D_{\ell s}}{D_s D_\ell}\frac{\boldsymbol{\theta}}{|\boldsymbol{\theta}|^2}} \tag{6.77}$$

If the source happens to be directly behind the lens, $\boldsymbol{\eta} = 0$ and $\boldsymbol{\beta} = 0$, so $\boldsymbol{\theta} = \boldsymbol{\alpha}$. The angle that solves the lens equation in this case has a special name, the **Einstein angle** θ_E:

$$\boxed{\theta_E = \sqrt{\frac{4GM}{c^2}\frac{D_{\ell s}}{D_s D_\ell}}} \tag{6.78}$$

We can rewrite the lens equation yet again using this notation

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \theta_E^2 \frac{\boldsymbol{\theta}}{|\boldsymbol{\theta}|^2} \tag{6.79}$$

Notice that so long as the magnitude of $\boldsymbol{\theta}$ is θ_E, it doesn't matter what direction it is, it always solves the lens equation. In other words, if the source is directly behind the lens, the image we get is a *ring* called an **Einstein ring**. This configuration is shown in Figure 6.7. The first person to ever observe such an Einstein ring was Jackie Hewitt here at MIT/MKI!

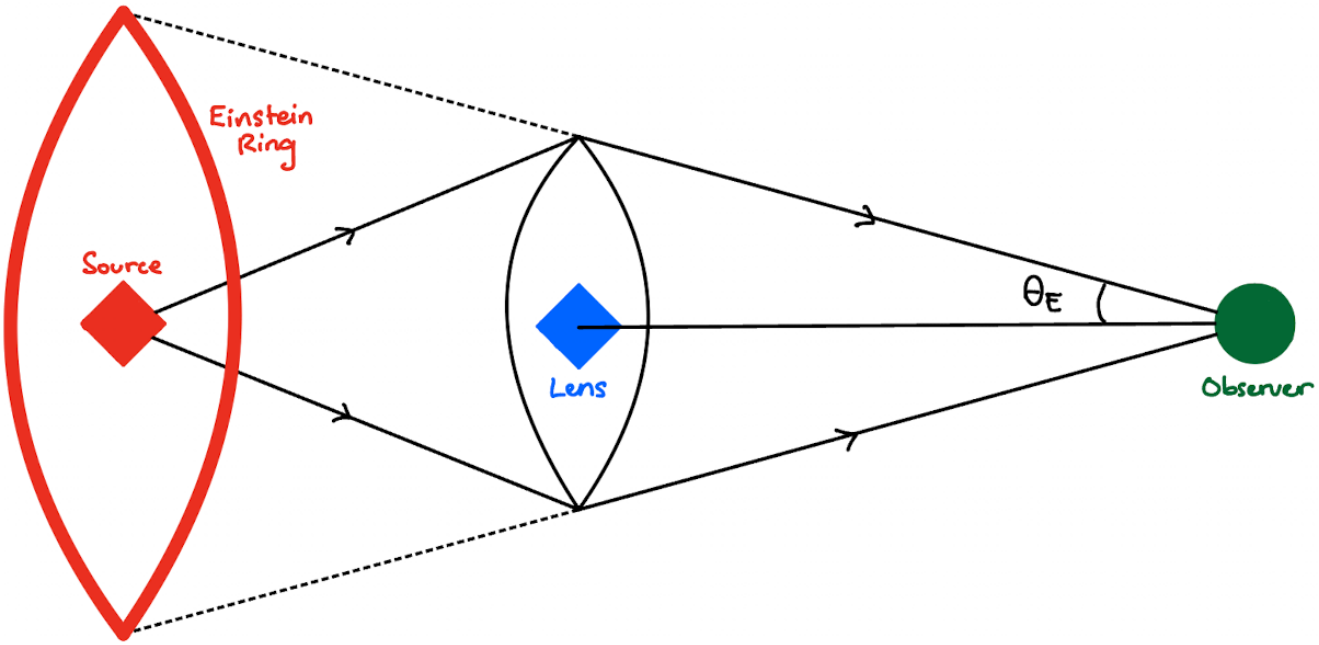


Figure 6.7: Gravitational lensing creating an Einstein ring of a source when it lies directly behind the lens.

Note that, in general, a lensed image of a source can also have its brightness magnified relative to the original source. The light beams are subject to differential deflection, which tends to spread them out. The surface brightness, however, is conserved because the deflection of light does not affect any of the light's emission or absorption. Thefefore, if the unlensed source would subtend a solid angle

Ω_s, but the observed solid angle is Ω, then the magnification factor is

$$\mu = \frac{\Omega}{\Omega_s} \tag{6.80}$$

For sources and images that are much smaller than the characteristic scale of the lens, we can write this as the differential area distortion of the lens mapping

$$\mu = \left| \det\left(\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}} \right) \right|^{-1} \tag{6.81}$$

Or, for a point-mass lens, the magnifications of the two images are

$$\mu_{\pm} = \frac{1}{4}\left(\frac{y}{\sqrt{y^2+4}} + \frac{\sqrt{y^2+4}}{y} \pm 2 \right) \tag{6.82}$$

where $y = |\boldsymbol{\beta}|/\theta_E$ is a dimensionless parameter[119,120].

Measuring the Hubble Constant

For a source that has two or more distinct images, the paths that the light takes to reach us will have different lengths. Thus, the information we receive from the image with the longer path will be delayed by some time relative to the image with the shorter path. See the setup in Figure 6.8.

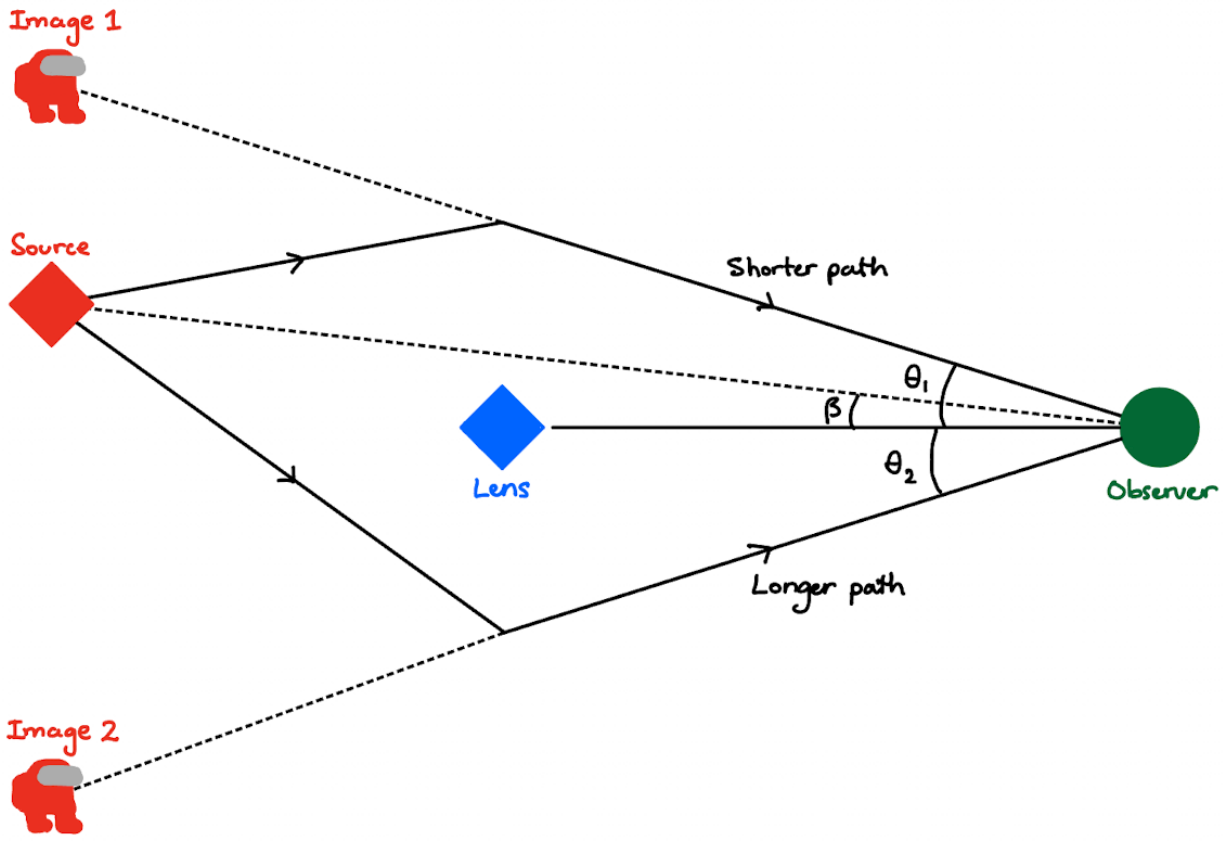


Figure 6.8: Gravitational lensing where two distinct images are created along paths of different lengths.

The time delay will be related to the difference in path lengths. Relative to a straight path that just travels along $D_{\ell s}$ and D_ℓ, a path with an impact parameter ξ will be longer in each segment by

$$\Delta D_{\ell s} = \sqrt{D_{\ell s}^2 + \xi^2} - D_{\ell s} \approx \frac{1}{2}\frac{\xi^2}{D_{\ell s}} \tag{6.83}$$

$$\Delta D_{\ell} = \sqrt{D_{\ell}^2 + \xi^2} - D_{\ell} \approx \frac{1}{2}\frac{\xi^2}{D_{\ell}} \tag{6.84}$$

Then the total difference is

$$\Delta D_s = \frac{1}{2}\xi^2\left(\frac{1}{D_\ell} + \frac{1}{D_{\ell s}}\right) = \frac{1}{2}D_\ell^2(\boldsymbol{\theta} - \boldsymbol{\beta})^2\left(\frac{D_\ell + D_{\ell s}}{D_\ell D_{\ell s}}\right) \tag{6.85}$$

where the $\boldsymbol{\beta}$ is introduced so we measure delays relative to the path that goes straight to the source (rather than straight to the lens). Converting this into a time delay

$$\Delta t = \frac{1+z_\ell}{c}\Delta D_s = \frac{1+z_\ell}{c}\frac{D_\ell D_s}{D_{\ell s}}\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\beta})^2 \tag{6.86}$$

We also have an additional time delay that is introduced due to gravitational time dilation as the photons pass through the gravitational potential of the source (Shapiro time delay; see §6.Q97). This just introduces an additional term, making the total time delay[119]

$$\boxed{\Delta t = \frac{1+z_\ell}{c}\frac{D_\ell D_s}{D_{\ell s}}\left[\frac{(\boldsymbol{\theta} - \boldsymbol{\beta})^2}{2} - \psi(\boldsymbol{\theta})\right]} \tag{6.87}$$

where $\psi(\boldsymbol{\theta})$ is the 2D lensing potential. Often the prefactor out front is given a special name, the **time-delay distance**

$$\boxed{D_{\Delta t} \equiv \frac{1+z_\ell}{c}\frac{D_\ell D_s}{D_{\ell s}}} \tag{6.88}$$

Now, D_ℓ, D_s, and $D_{\ell s}$ are all *angular diameter distances*, which are proportional to H_0^{-1}. Therefore, the time-delay distance is also proportional to H_0^{-1}. In other words, if we measure the time delay Δt, we can estimate the time-delay distance $D_{\Delta t}$ and get an estimate for H_0.

The current state of H_0 measurements using gravitational lensing time delays puts the Hubble constant at[121]

$$H_0 = 73.3^{+1.7}_{-1.8}\,\mathrm{km\,s^{-1}\,Mpc^{-1}} \tag{6.89}$$

In agreement with many other late-universe methods.

97. What is the "Shapiro time delay"? Where was it first measured?

The **Shapiro time delay** is induced by gravitational time dilation when light passes near a gravitational potential well. It can be calcualted from

$$\boxed{\Delta t_g = -\frac{1+z_\ell}{c}\int \frac{2\Phi}{c^2}\mathrm{d}\ell = -\frac{1+z_\ell}{c}\frac{D_\ell D_s}{D_{\ell s}}\psi(\boldsymbol{\theta})} \tag{6.90}$$

where $\psi(\boldsymbol{\theta})$ is the **2D lensing potential**[119]

$$\psi(\boldsymbol{\theta}) = \frac{D_{\ell s}}{D_s D_\ell}\int \frac{2\Phi}{c^2}\mathrm{d}\ell \tag{6.91}$$

This effect was first predicted by Shapiro in 1964 and become of the four classical tests of GR:

1. The perihelion precession of Mercury
2. Deflection of stars by the Sun during a solar eclipse
3. Gravitational redshift
4. Shapiro time delay

(gravitational waves and the Hulse-Taylor pulsar came later). It was first measured in 1966 when Arecibo and Haystack (MIT) send radio pulses to Mercury and Venus when they were on the opposite side of the Sun relative to us. The light pulses should experience a Shapiro time delay as they pass through the gravitational potential of the Sun. The measured values matched the predictions with a delay of 0.2 ms. Measurements of multiple objects at different impact parameters from the Sun were necessary to break the degeneracy between having a more distant object and a less massive lens[122,123].

98. What is gravitational microlensing, what do we learn from it, and how are various experiments studying this phenomenon?

For stars, compact objects, and planets in our galaxy, light is typically only deflected on the order of microarcseconds. This is not resolvable with our current telescopes, but we *can* still resolve the magnification produced by lensing (see §6.Q96). This assumes that we know what the unlensed flux of the source should be, which in general we don't. However, if a lens happens to be moving relative to a source such that it passes in front of it, then the magnification changes over time, and we can compare the magnified flux to the baseline unmagnified flux.

The position of the source can be considered linear during the time intervals we're interested in, so the source passes behind the lens, and the distance $y(t) = |\boldsymbol{\beta}(t)|/\theta_E$ is determined just from the Pythagorean theorem (see Figure 6.9)

$$\boxed{y(t) = \sqrt{y_0^2 + \left(\frac{t - t_0}{t_E}\right)^2}} \tag{6.92}$$

where $t_E = \theta_E/\dot{\theta}$ is a characteristic timescale.

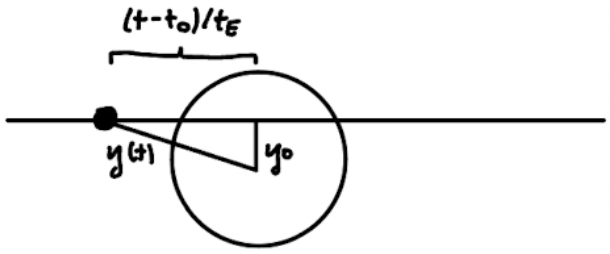


Figure 6.9: The geometry as the source passes behind the lens and changes the dimensionless distance $y(t)$.

The total magnification as a function of position can be obtained by summing up the magnifications of the two images from equation (6.82):

$$\boxed{\mu[y(t)] = \frac{y^2(t) + 2}{y(t)\sqrt{y^2(t) + 4}}} \tag{6.93}$$

This produces a characteristic light curve with a broad peak that depends on the distance of closest approach ($p = y_0$). See Figure 6.10[120].

What can we learn from microlensing events?

We can obtain the mass of the lens, since the amount of magnification depends on it (θ_E depends on M, and y depends on θ_E), however, it is degenerate with the distances D_s, D_ℓ, and $D_{\ell s}$, and the transverse velocity $\dot{\theta}$, so these must be known precisely to get an accurate mass estimate. This can be done with parallax or other distance measuring techniques. We can also measure lensing events in multiple bands to rule out stellar variability, since the lensing effect is achromatic (whereas stellar variability is not).

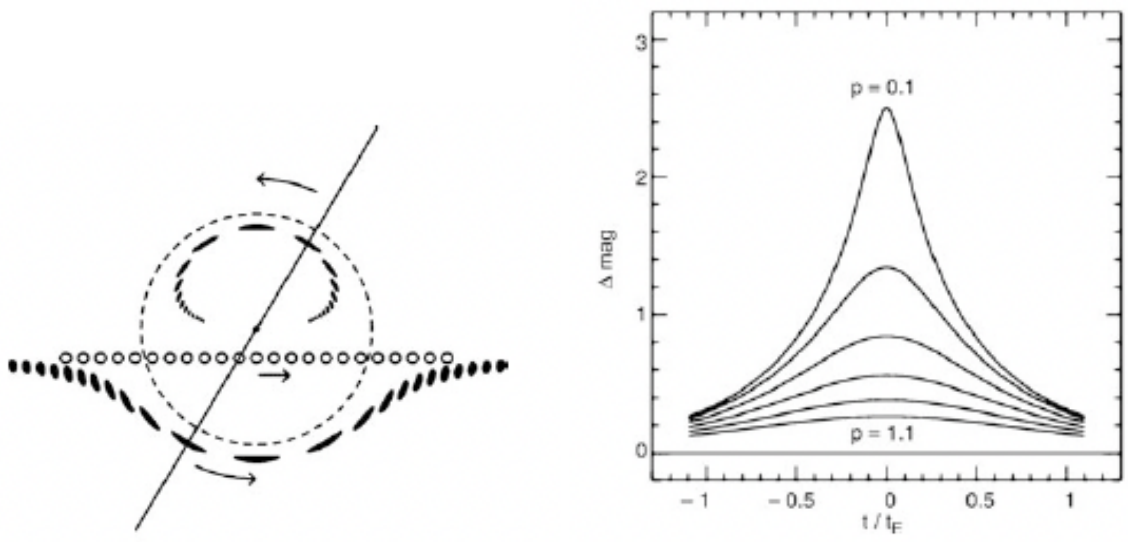


Figure 6.10: The light curve of a microlensing event for various impact parameters p, shown alongside the magnification of the source as a function of the transit.

This technique is useful in uncovering the existence of planets orbiting the lens, as they will induce additional blips in the microlensing light curves (see §2.Q13). Additionally, microlensing can be used to uncover otherwise "dark" objects (i.e. isolated black holes) in the galaxy, since these objects can still act as gravitational lenses for background objects.

How are various experiments studying this phenomenon?[120]

Microlensing events are rare and time sensitive, so they require high cadence surveys of many stars to be constantly monitoring for them. Some examples of these surveys are

- EROS: Searched the SMC, LMC, galactic bulge, and spiral arms from 1993–2002
- MACHO: Searched the galactic bulge and LMC from 1993–1999
- OGLE: Searched the SMC, LMC, and galactic bulge from 1992–present

These surveys were primarily looking for **Massive Compact Halo Objects (MACHOs)**. These are just massive objects in the galactic halo that do not emit a lot of light (i.e. brown dwarfs, white dwarfs, black holes, etc.), and they were thought to be one possible explanation for dark matter. However, these surveys have found that the density of these objects in the galaxy is too low to explain the required amount of dark matter. In particular, the lensing probability in the direction of the Magellanic Clouds suggested that only about 20% of the halo mass consists of MACHOs, with a characteristic mass of $\sim 0.5\ M_\odot$. As a result, this dark matter hypothesis has largely fallen to the wayside in the modern day, being replaced by the **WIMPs (Weakly Interacting Massive Particles)** hypothesis.

Other results from these surveys:

- A couple hundred exoplanet detections have been reported
- The distribution of stars in the galaxy can be measured by analyzing the lensing probability as a function of direction
- Many new variable stars have been newly discovered and accurately monitored

- Proper motions of several million stars have been determined based on 20 years of microlensing surveys
- Some lensing events allow for the measurement of the radius and surface structure of distant stars, since μ depends on the position of the source. For example, if the lens is a binary system, it can no longer be treated as a point source since the star acting as the lens will be moving relative to the source and us, so the light curve now depends on the star's radius.

99. What is the "weak lensing effect"? How does this differ from "strong lensing"? What are each of these phenomena useful for to astronomers?

Strong Lensing

In strong lensing, the source is visibly distorted and may form an arc, ring, or multiple images. This requires a very massive lens and is most often observed in galaxy clusters. An example is shown in Figure 6.11.

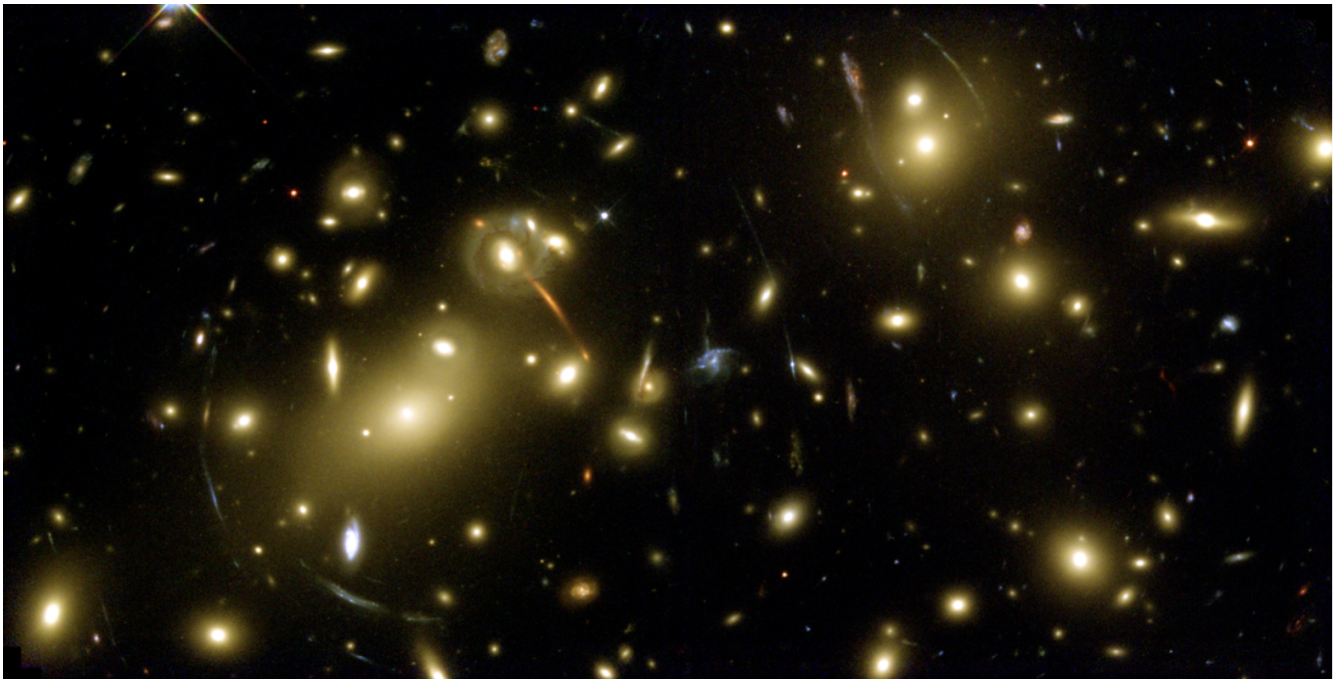


Figure 6.11: An example of strong lensing of background galaxies. The lens in this case is the galaxy cluster Abell 2218.

In §6.Q96, we only considered simple examples where the lens is a point mass. However, in the case of a galaxy cluster, this is a terrible assumption. Instead, we have some extended mass density $\rho(\xi, z)$ where ξ is the 2D vector giving the impact parameter and z is the distance along the line of sight. We can't measure the full 3D mass density, but if we assume the lens plane is relatively localized in comparison to the source and the observer, we can convert this into a 2D mass surface density Σ:

$$\Sigma(\xi) = \int \rho(\xi, z) \mathrm{d}z \tag{6.94}$$

Then, the deflection angle will be given by

$$\boxed{\hat{\boldsymbol{\alpha}}(\xi) = \frac{4G}{c^2} \int \Sigma(\xi') \frac{\xi - \xi'}{|\xi - \xi'|^2} \mathrm{d}^2\xi'} \tag{6.95}$$

We notice that $\hat{\boldsymbol{\alpha}}$ is effectively a convolution of $\Sigma(\xi)$ with the kernel $K(\xi) \propto \xi/|\xi|^2$. Therefore, we

can look at everything in Fourier space to make the analysis slightly easier

$$\tilde{\hat{\boldsymbol{\alpha}}}(\mathbf{k}) \propto \tilde{\Sigma}(\mathbf{k})\tilde{K}(\mathbf{k}) \tag{6.96}$$

We can now quantify the criterion for strong lensing more rigorously. The *critical* surface density Σ_{crit} is defined as

$$\boxed{\Sigma_{\text{crit}} = \frac{c^2}{4\pi G}\frac{D_s}{D_\ell D_{\ell s}}} \tag{6.97}$$

And the **convergence** $\varkappa$ is defined as the surface density relative to the critical surface density[118]

$$\varkappa(\boldsymbol{\theta}) = \frac{\Sigma(\boldsymbol{\theta})}{\Sigma_{\text{crit}}} \tag{6.98}$$

If $\varkappa > 1$, then we expect strong lensing, whereas if $\varkappa < 1$ we do not expect strong lensing.

Strong lensing is useful for a number of applications:

- We can use strong lensing to get the total mass of the lens object, but often it is not as powerful as weak lensing in mapping the spatial distribution of the 2D lensing potential
- High-z sources can be studied because the strong lensing increases the magnification (6.81), making them much brighter than they otherwise would be
- Time delays between multiple images can be used in cosmological studies, like in measuring the Hubble constant (see §6.Q96)

Weak Lensing

In weak lensing, the distortion of the source is...well...much weaker than in strong lensing. For any individual source, this makes it impossible to determine whether it is elongated due to an *intrinsic* elongated shape or due to an *extrinsic* lensing effect. However, weak lensing can still be distinguished by using a statistical analysis of many sources in a field of view. There should be no preferred direction of elongation for intrinsic shape variations, but for lensing distortions, the shapes will all be distorted in a predictable way. The difference between these two scenarios is highlighted in Figure 6.12.

For each galaxy in the image, the first moment of the surface brightness (i.e. the center of luminosity) is

$$\boldsymbol{\theta}_{\text{CM}} = \frac{\int I(\boldsymbol{\theta})\boldsymbol{\theta}\,\mathrm{d}^2\boldsymbol{\theta}}{\int I(\boldsymbol{\theta})\mathrm{d}^2\boldsymbol{\theta}} \tag{6.99}$$

And the shape of the image is quantified by the second moment, determined by the 2x2 quadrupole tensor

$$Q_{ij} = \frac{\int I(\boldsymbol{\theta})(\theta_i - \theta_{\text{CM},i})(\theta_j - \theta_{\text{CM},j})\mathrm{d}^2\boldsymbol{\theta}}{\int I(\boldsymbol{\theta})\mathrm{d}^2\boldsymbol{\theta}} \tag{6.100}$$

Then the *ellipticity* is

$$\boxed{|\epsilon| = \frac{\sqrt{(Q_{11}-Q_{22})^2 + 4Q_{12}^2}}{Q_{11}+Q_{22}+2\sqrt{Q_{11}Q_{22}-Q_{12}}}} \tag{6.101}$$

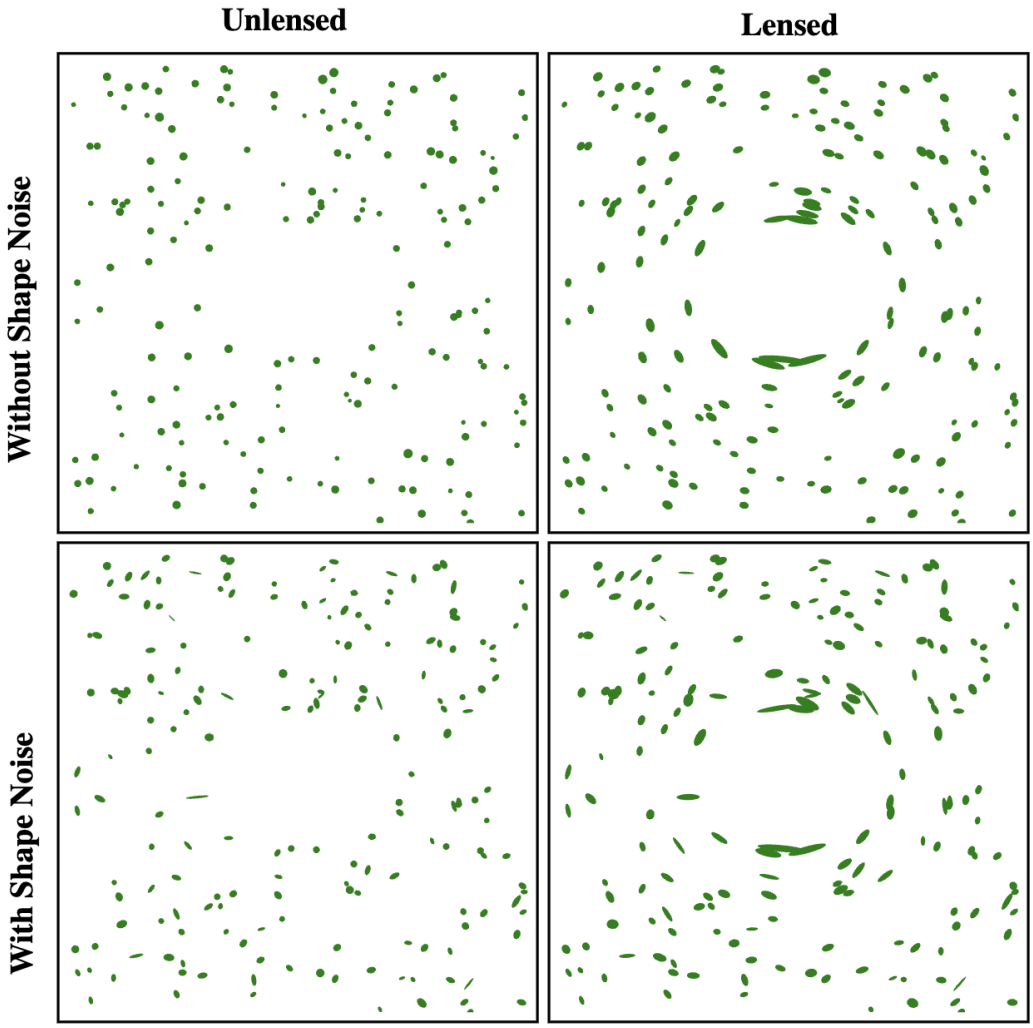


Figure 6.12: The difference between intrinsic shape noise and extrinsic noise generated by gravitational lensing effects.

And the *position angle* φ of the ellipse is[118]

$$\boxed{\tan(2\varphi) = \frac{2Q_{12}}{Q_{11} - Q_{22}}} \tag{6.102}$$

Weak lensing is useful for different applications than strong lensing:

- Can be used to fully map the 2D potential/mass of the lensing object. This has been done, for example, in the Bullet Cluster (§7.Q101).
- The point-spread function (PSF) of the telescope used to observe the galaxies must be well known, since this also has the effect of smearing out point sources in a predictable way.
- Distance measurements for the background galaxies are also required (i.e. with spectroscopic redshifts) to break the degeneracies between the lens mass and the distance.

7 Galaxies

100. Make a sketch of the Galaxy to scale (top and side views) and indicate its various features and properties. How are stars, gas, dust, and dark matter distributed in our Galaxy?

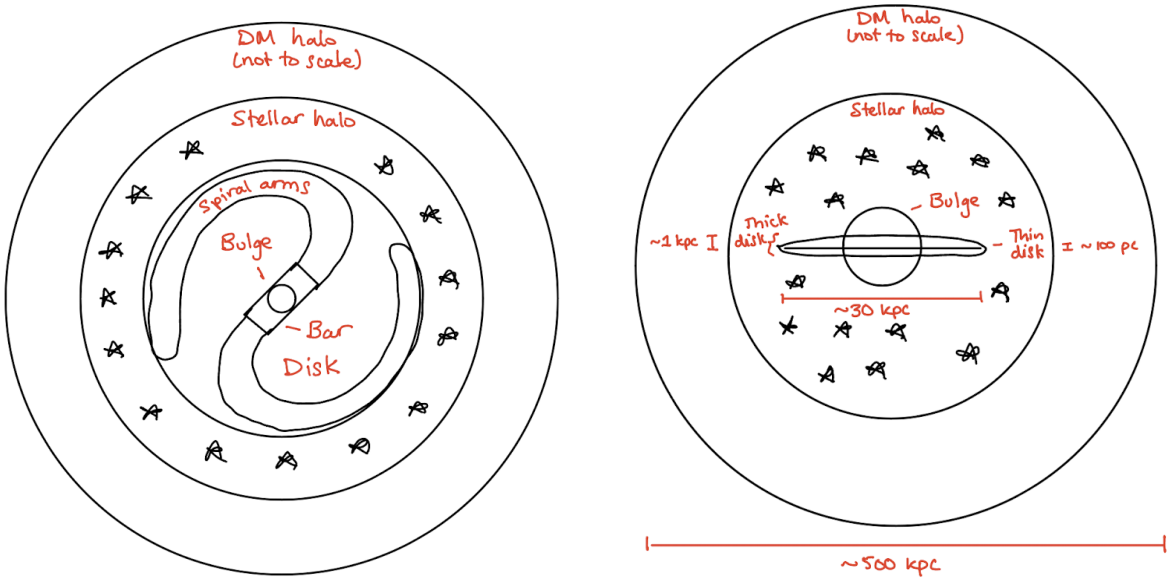


Figure 7.1: Milky Way diagram. *Left*: Top-view, *Right*: Side-view.

The galaxy can be decomposed into a few different regions (see Figure 7.1)[119]:

- **Disk**: Composed of cold/cool gas, dust, and young metal-rich stars. Star formation is actively going on in the spiral arms. The Solar System lies in the thin disk ∼8 kpc from the galactic center. The barred structure is thought to be a recurrent dynamical feature in the disk phase. Elliptical galaxies do not contain a disk and thus are not actively undergoing star formation.
- **Bulge**: Composed of older, metal-poor (Z $\sim 10^{-2}$–$10^{-3} Z_\odot$) stars in dynamically "hot" orbits around the central SMBH Sagittarius A*. In other galaxies, this may be where an AGN resides.
- **Stellar halo**: A spherically symmetric halo around the disk with warm/hot gas and older, metal-poor stars distributed in globular clusters with $\sim 10^6$ stars each. May contain remnants of mergers with satellite galaxies.
- **Dark matter halo**: A spherically symmetric halo even larger than the stellar halo that becomes the dominant feature at large scales. This halo contains just the dark matter.

101. What observational evidence do we have for dark matter in galaxies and clusters of galaxies?

Evidence for dark matter spans a wide variety of categories and shows up prominently from sub-galaxy scales up to the large scale structure of the universe.

1. Galaxy rotation curves[11,119]: We can probe the enclosed mass of a galaxy $M(<r)$ by measuring the circular velocity as a function of radius:

$$m\frac{v^2}{r} = \frac{GM(<r)m}{r^2} \quad \longrightarrow \quad \boxed{v(r) = \sqrt{\frac{GM(<r)}{r}}} \tag{7.1}$$

Measurements of v can be gotten a number of ways. In the Milky Way, H I 21 cm emission (see §7.Q103) is used. In other galaxies, long-slit optical spectroscopy can be used.

Let's try to make a prediction about what the profile will look like based on the brightness profile of the galaxy. We can make a rather naive assumption that the **mass-to-light ratio** M/L of a galaxy is constant with radius. This should make intuitive sense, since having more stars leads to more mass and more light, and any differences in the amount of light produced by a given mass of stars should not be position-dependent as long as we average over large enough areas (it's the same stellar population). Then, we can measure the luminosity of a typical spiral galaxy by considering an exponential surface brightness profile with scale height h:

$$I(r) = I_0 e^{-r/h} \tag{7.2}$$

Note: See §12.11 for an explanation of the units of I here.

We can integrate over the cross-sectional area to get the luminosity:

$$L(r) = \int_0^r \mathrm{d}r' I(r') 2\pi r' = 2\pi I_0 \int_0^r \mathrm{d}r' r' e^{-r'/h} \tag{7.3}$$

Integrating by parts ($\int \mathrm{d}v\, u = uv - \int \mathrm{d}u\, v$):

$$u = r' \quad \longrightarrow \quad \mathrm{d}u = 1 \;\;,\;\; \mathrm{d}v = e^{-r'/h}\mathrm{d}r' \quad \longrightarrow \quad v = -he^{-r'/h} \tag{7.4}$$

$$L(r) = 2\pi I_0 \left[-r'he^{-r'/h}\Big|_0^r + \int_0^r \mathrm{d}r' he^{-r'/h} \right] \tag{7.5}$$

$$= 2\pi I_0 \left[-r'he^{-r'/h} - h^2 e^{-r'/h} \right]_0^r \tag{7.6}$$

$$= 2\pi I_0 [h^2 - he^{-r/h}(r+h)] \tag{7.7}$$

And taking the limit for large r:

$$\lim_{r\to\infty} L(r) = 2\pi I_0 h^2 = \text{const.} \tag{7.8}$$

Thus, the luminosity becomes constant at sufficiently large radii (i.e. when we reach the outskirts of the disk so there are no more stars and the gas density goes way down). Then, since M/L is also constant, the $M(<r)$ implied by the light should also be constant. Again, this makes intuitive sense: once we reach the edge of the disk and the density of stars/gas goes way down, $M(<r)$ should stop

increasing.

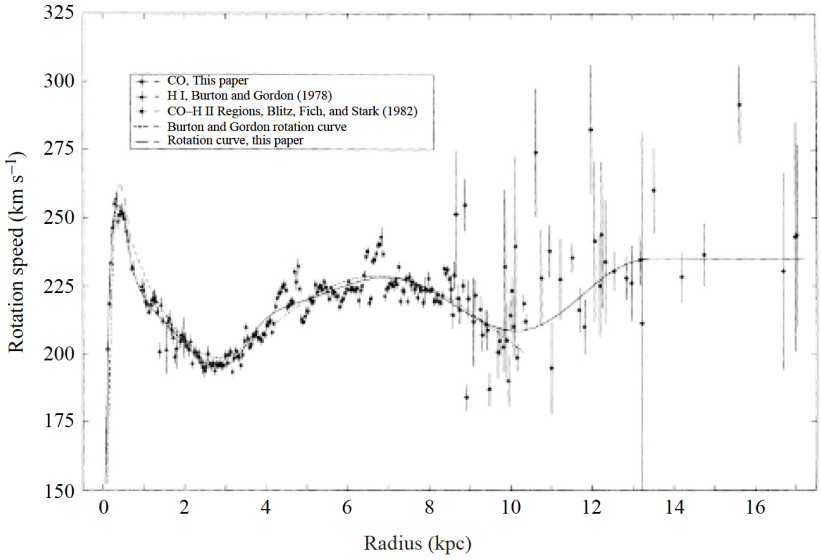


Figure 7.2: The rotation curve of the Milky Way measured out to 17 kpc[11].

Now, we return to the velocity profile. If our assertions are correct, then by ∼15 kpc (the radius of the disk) we should expect $M(< r)$ to become constant, thus the velocities ought to follow an $r^{-1/2}$ Keplerian decline (this also assumes the disk is not self-gravitating). However, it has been observed in the Milky Way (Figure 7.2) and many other galaxies (by **Vera Rubin** in the 1970s) that $v(r)$ remains constant well above this radius! This would imply that instead of $M(< r)$ becoming constant, $M(< r)/r$ becomes constant, i.e. the mass continues increasing as $M(< r) \propto r$. The only explanation is that there is some additional mass that is not in the stars/gas and does not produce any light, which is why it's called **dark matter**.

2. Galaxy cluster velocities[120,124]: By a similar argument to the rotation curves, we can use the **virial theorem** to relate the velocity dispersion of galaxies in a cluster to its mass. We make the assumption that the galaxies in a cluster are **virialized**, i.e. they are in a stable gravitationally bound configuration where the system is not losing/gaining any mass or energy. The virial theorem states:

$$2T + U = 0 \quad \longrightarrow \quad T = -\frac{U}{2} \tag{7.9}$$

We can recast this in terms of the mass and velocity by writing out expressions for T and U, summing over each galaxy in the cluster:

$$T = \frac{1}{2}\sum_i m_i v_i^2 = \frac{1}{2} M_{\text{vir}} \langle v^2 \rangle \quad , \quad U = -\sum_{i<j} \frac{G m_i m_j}{r_{ij}} = -\frac{G M_{\text{vir}}^2}{R} \tag{7.10}$$

Where M_{vir} is the total mass, and $\langle v^2 \rangle$ and R are some characteristic mass-weighted velocity and radius for the whole system:

$$M_{\text{vir}} = \sum_i m_i \quad , \quad \langle v^2 \rangle = \frac{1}{M_{\text{vir}}} \sum_i m_i v_i^2 \quad , \quad R = M_{\text{vir}}^2 \left(\sum_{i<j} \frac{m_i m_j}{r_{ij}} \right)^{-1} \tag{7.11}$$

Which gives

$$M_{\rm vir}\langle v^2\rangle = \frac{GM_{\rm vir}^2}{R} \quad \longrightarrow \quad M_{\rm vir} = \frac{R\langle v^2\rangle}{G} \tag{7.12}$$

Putting this in terms of observables, we measure a projected 1D velocity dispersion in the radial direction (σ) rather than the full 3D dispersion $\langle v^2\rangle$, so we need a correction of $\langle v^2\rangle = \langle v_x^2\rangle + \langle v_y^2\rangle + \langle v_z^2\rangle \approx 3\sigma^2$ if we assume the velocity field is isotropic. Additionally, the characteristic radius R represents the average (mass-weighted) separation between two galaxies in our cluster, whereas we usually want to measure the full radius of the cluster up to some density threshold. We can make a simple assumption that, on average, a galaxy lands at ∼halfway between the center and the outer edge, such that $R \approx R_{\rm vir}/2$. Thus, we end up with:

$$\boxed{M_{\rm vir} \approx \frac{3}{2}\frac{R_{\rm vir}\sigma^2}{G}} \tag{7.13}$$

The first to take advantage of this relationship in galaxy clusters was **Fritz Zwicky** in the 1930s. He used observations of the **Coma cluster** and obtained $\sigma \sim 1000\,{\rm km\,s^{-1}}$. The inferred mass from these dynamics produces a high mass-to-light ratio. Indeed, typical values in clusters are

$$\frac{M}{L} \sim 200h_{70}\left(\frac{M_\odot}{L_\odot}\right) \tag{7.14}$$

which is higher than the typical value in early-type galaxies by a factor of ∼10. This implies that a significant fraction of the mass in galaxy clusters is dark matter. The stars visible in galaxies contribute $\lesssim 5\%$ of the total mass in galaxy clusters!

3. Gravitational lensing[120]: The effects of gravitational lensing can be used to probe the mass of galaxy clusters independently of any kinematics or dynamics, since the deflection angle of light depends on the mass of the lens (see §6.Q96 and §6.Q99 on gravitational lensing). When in the weak lensing regime, the distortion effect on galaxies is of the same order or smaller than natural ellipticities expected in their shapes. However, the intrinsic ellipticities of the galaxies should be randomly oriented, whereas the distortions from the lensing model should have a preferred direction, so these two effects can be separated by averaging over the image ellipticities. This measures the total (dark + luminous) matter in a cluster. Weak lensing of galaxy clusters shows, in agreement with the dynamics, high mass-to-light ratios.

The Bullet Cluster is a unique case where two clusters have recently collided. In principle, the previous two results regarding the rotation curves of galaxies and galaxy velocities in clusters could also be explained by a modification to the law of gravity on large spatial scales, rather than some unexplained dark matter component (i.e. MOND). The Bullet cluster provides us with an unambiguous test for which of these scenarios is correct. The hot X-ray emitting gas in the two clusters contains much more mass than the stars in the galaxies, and it is heavily affected by the collision. However, the galaxies themselves are collisionless and essentially pass straight through each other. We would expect the dark matter component to also be collisionless and trace the galaxies. We can measure the mass distribution of the cluster using weak gravitational lensing and see where the bulk of the mass resides. If it's in the X-ray gas, then the previous results must be due to a modified law of gravity on large scales, whereas if it's in the galaxies, it must be due to some collisionless dark matter component.

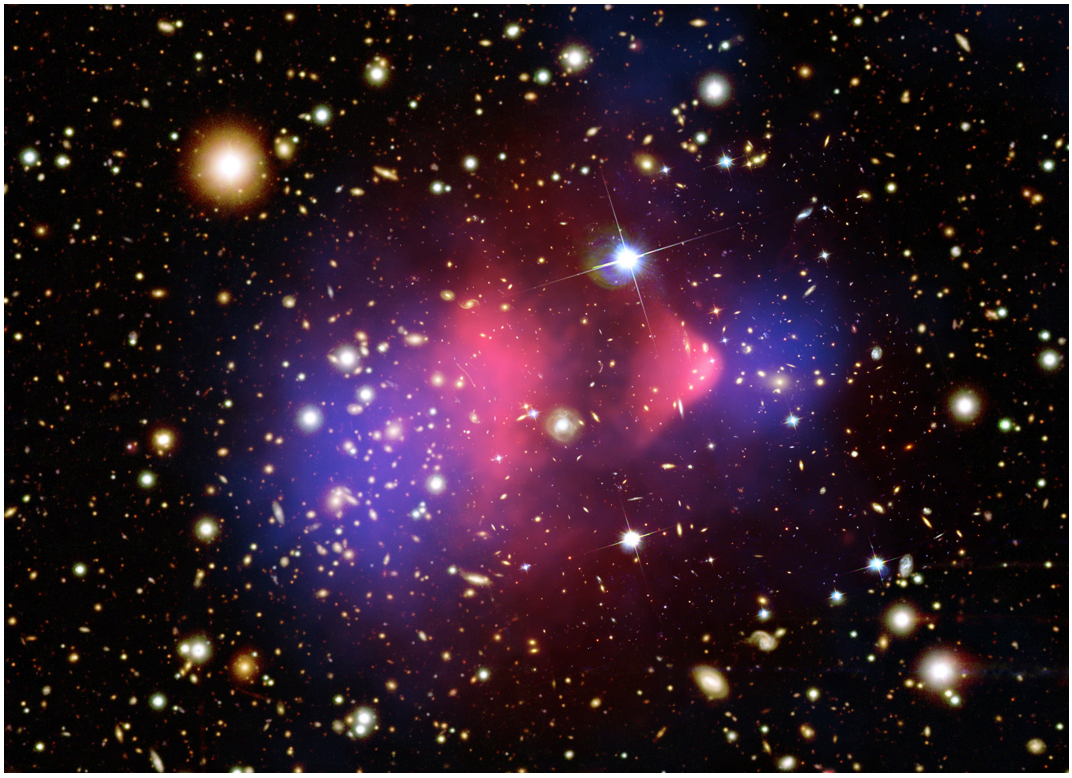


Figure 7.3: Image of the bullet cluster. The X-ray gas is shown in red and the weak lensing mass model is shown in blue. It is clear that the two distributions are separated, with most of the mass residing in the galaxies. Credit: ESA.

We see unambiguously that the bulk of the mass resides in the galaxies, clearly displaced from the X-ray gas, providing unambiguous evidence for dark matter (see Figure 7.3)[120]. Weak lensing can also be employed on a large number of other clusters to reconstruct the mass density as a function of radius, obtaining typical mass-to-light ratios that agree with those found by the dynamics ($\sim 200 h_{70}$ in solar units; equation (7.14)).

4. X-ray emission in clusters[120]: Provided a cluster has not recently merged like the Bullet cluster, the hot intracluster medium (ICM) can be assumed to be in hydrostatic equilibrium, in which case we can use

$$\nabla P = -\rho_g \nabla \Phi \quad \longrightarrow \quad \frac{1}{\rho_g}\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{\mathrm{d}\Phi}{\mathrm{d}r} = -\frac{GM(<r)}{r^2} \tag{7.15}$$

Where ρ_g is the gas density, but $M(<r)$ is the total enclosed mass, not just the gas mass, since it comes from the gravitational potential Φ. Using the ideal gas law:

$$P = nkT = \frac{\rho_g}{\mu m_p} kT \tag{7.16}$$

$$M(<r) = -\frac{kTr}{G\mu m_p}\left(\frac{\mathrm{d}\ln\rho_g}{\mathrm{d}\ln r} + \frac{\mathrm{d}\ln T}{\mathrm{d}\ln r}\right) \tag{7.17}$$

Thus, we can obtain the total mass just by knowing the gas density ρ_g and temperature T of the ICM. The X-rays emitted by the ICM originate from thermal bremsstrahlung radiation, which has as

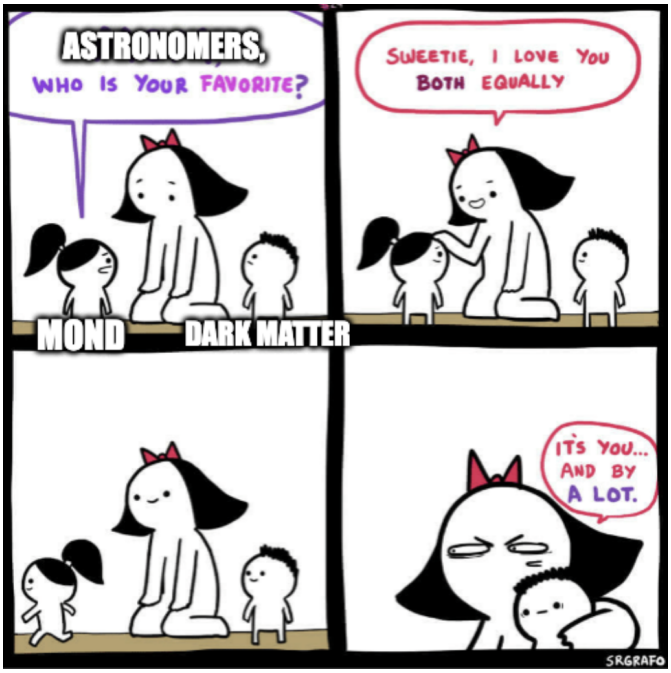
emissivity proportional to both the density and temperature (see §5.Q74 on bremsstrahlung). We can deproject the observed intensity I_ν into the emissivity using[125]

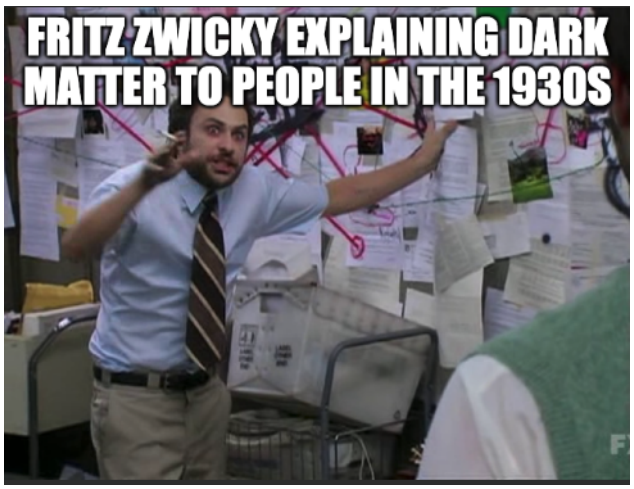
$$I_\nu(b) = 2\int_b^\infty \mathrm{d}r \frac{\varepsilon_\nu^{\mathrm{ff}}(r) r}{\sqrt{r^2 - b^2}} \quad \longrightarrow \quad \varepsilon_\nu^{\mathrm{ff}}(r) = -\frac{1}{\pi r}\frac{\mathrm{d}}{\mathrm{d}r}\int_r^\infty \mathrm{d}b \frac{I_\nu(b) b}{\sqrt{b^2 - r^2}} \tag{7.18}$$

For $h\nu \ll kT$, j_ν^{ff} depends only weakly on T, so one can derive $\rho_g(r)$ by assuming $T(r) = \text{const.}$ (these assumptions lead to the **beta model**). Then, using equation (7.17) to estimate the mass, one obtains mass-to-light ratios that are very similar to those inferred both by the dynamics and the gravitational lensing methods. Thus, by the same argument, ∼5% of the cluster mass is in the stars and ∼15% is in the ICM, so there must be some dark matter component that makes up ∼80% of the mass of the cluster.

5. Cosmic microwave background: The power spectrum of the CMB can also be used to measure the dark matter content of the universe. The relative strength of the odd/even peaks in the spectrum is sensitive to Ω_b (the baryon density), and the overall amplitude of the whole spectrum is sensitive to Ω_m (the total matter density). Thus, the dark matter can be constrained since the dark matter density $\Omega_c = \Omega_m - \Omega_b$. For more details, see §8.Q129.

6. Large scale structure formation: Dark matter plays a crucial role in the formation of structure in the early history of our universe. Since dark matter interacts gravitationally with baryonic matter, but is not opposed by forces like pressure, it allows density perturbations to collapse and begin forming structure (i.e. galaxies) much earlier in the history of the universe than would have been possible if only baryonic matter existed. In other words, without dark matter, the epoch of structure formation would have occurred much later than observed (at about ∼ 1 Gyr after the Big Bang). See §8.Q131 and §8.Q132 for a detailed account of how density perturbations grow in the early universe.





102. What are the typical mass ranges of dwarf galaxies, normal galaxies, and galaxy clusters? Describe a few ways in which we measure the masses of these different types of systems.

Typical masses for each type of system are given below[11,107].

Note: For a description of the Hubble classification system and abbreviations, see §7.Q109.

- **Dwarf galaxies**: $10^7 - 10^9\ M_\odot$
  - **BCD**: $\sim 10^9\ M_\odot$
  - **dE**: $10^7 - 10^9\ M_\odot$
  - **dSph**: $10^7 - 10^8\ M_\odot$
- **Normal galaxies**: $10^9 - 10^{13}\ M_\odot$
  - **Sa/SBa–Sc/SBc**: $10^9 - 10^{12}\ M_\odot$
  - **Sd/SBd–Sm/SBm**: $10^8 - 10^{10}\ M_\odot$
  - **S0/SB0**: $10^{10} - 10^{12}\ M_\odot$
  - **E**: $10^8 - 10^{13}\ M_\odot$
  - **cD**: $10^{13} - 10^{14}\ M_\odot$
  - **Im–Irr**: $10^8 - 10^{10}\ M_\odot$
- **Galaxy groups**: $10^{13} - 10^{14}\ M_\odot$
- **Galaxy clusters**: $10^{14} - 10^{15}\ M_\odot$

Measurement techniques for galaxies[11,119]:

- **Rotation curves**: Use the measured velocities (i.e. from spectra), and use $M(r) = rv_c^2/G$ for spiral galaxies.
- **Virial theorem**: Like above, but use $M \approx 3R\sigma^2/2G$ from the virial theorem for elliptical galaxies.
- **The Tully-Fisher Relation** relates the maximum rotational velocity of a spiral galaxy to its luminosity: $L \propto v_c^4$ (see §7.Q110). Thus, if one measures the luminosity/absolute magnitude of a galaxy (which requires an estimate of the distance), they can use this to estimate the rotational velocity, and then use that to get an estimate of the mass. This is useful in the case that one doesn't have any spectral information and so can't get a direct measurement of v_c, in which case a photometric redshift (or a standard candle) could be used to make an estimate of the distance.
- **The Faber-Jackson Relation** relates the velocity dispersion of an elliptical galaxy to its luminosity: $L \propto \sigma^4$, just like Tully-Fisher for spirals except it's the velocity dispersion rather than the rotational velocity. Thus, one can measure the luminosity/absolute magnitude of a galaxy to get an estimate for σ, and then use that to estimate the mass with the virial theorem.

For galaxy clusters[119,120]:

- **Virial theorem**: Like with galaxies, use the virial theorem to estimate the mass $M \approx 3R\sigma^2/2G$, only here σ is the velocity dispersion of the member galaxies of a cluster rather than the stars of a galaxy.

- **Gravitational lensing**: Use the distortions imparted by strong or weak gravitational lensing to create a lens model that depends on the mass (i.e. for a point mass the angular deflection $\hat{\alpha}(\xi) = 4GM/\xi c^2$).
- **Scaling relations**: There is a scaling relation between the total mass and the mass-weighted X-ray temperature $Y_X = T_X M_{\text{gas}}$. This is because the X-ray temperature T_X specifies the thermal energy per gas particle, which for a cluster in equilibrium, should be proportional to its binding energy:

$$T_X \propto \frac{M}{r} \tag{7.19}$$

  This is essentially a restatement of the virial theorem. Since $M \propto r^3$, this gives

$$T_X \propto r^2 \propto M^{2/3} \tag{7.20}$$

  Then, we should also expect M_{gas} to be related to the total mass M by some fraction f_{gas}. If we assume f_{gas} is constant, then $M_{\text{gas}} = f_{\text{gas}} M$ and

$$Y_X \propto M^{5/3} \tag{7.21}$$

- **X-ray emission**: Refer to §7.Q101 for a detailed explanation, but in brief, the emissivity of X-rays from bremsstrahlung radiation in the ICM can be used to get estimates for the gas density and temperature, which can then be related to the mass of the cluster by assuming the ICM is in hydrostatic equilibrium. See equation (7.17).

103. Sketch the rotation curve of our Galaxy, with approximate scales on the axes. How can information about the rotation curve be determined from 21 cm observations?

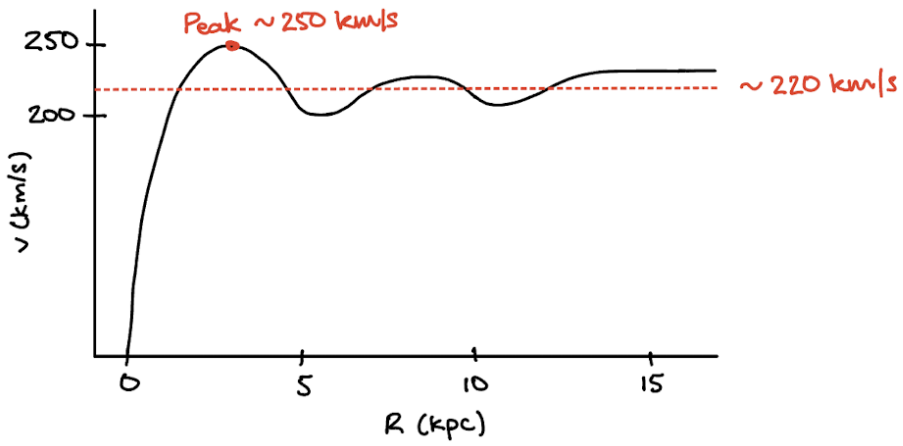


Figure 7.4: The rotation curve of the Milky Way drawn out to 17 kpc.

See Figure 7.4; also see Figure 7.2 for the actual data. H I 21 cm emission is a very useful tool for mapping the galactic rotation curve because virtually the entire galaxy is optically thin in this waveband, and the H I emission will be Doppler shifted based on the local motions of the gas[11,120].

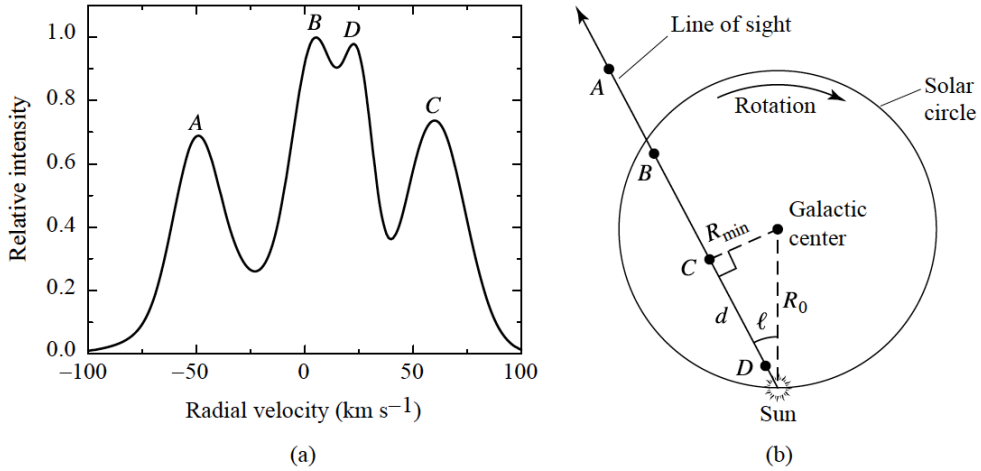


Figure 7.5: *Left*: A mock observed H I 21 cm emission line profile with distinct components coming from clouds A–D. *Right*: The geometry for the line of sight corresponding to the middle line profile with clouds A–D.[11]

However, to construct the rotation curve we also need distance measurements—the problem is that distance cannot be directly determined from H I 21 cm observations, so we have to be a bit clever about it. If we choose a line of sight that passes through the disk, we can use the so-called *tangent point method*.

- The observed H I profile along this line of sight will be a superposition of several gas clouds going at different velocities at different distances to us.
- If we assume the angular velocity Ω is a monotonically decreasing function with distance from the galactic center, then the radial velocity should attain a maximum where the line of sight is tangent to the local orbit (see point C in Figure 7.5b).
- Under this assumption, the distance from us to this cloud is directly related to the galactic longitude ℓ and solar distance from the galactic center R_0 (8 kpc): $d = R_0 \cos \ell$.
- And the distance from the cloud to the galactic center (which is what we're really interested in when looking at the rotation curve) is just $R_{\text{min}} = R_0 \sin \ell$.
- Then the radial velocity v_r is taken from the largest velocity component in the observed H I emission (see peak C in Figure 7.5a). Again, under the tangent point assumption, the radial velocity here is equal to the full velocity.
- (See §7.Q117 for a full description of the galactic rotation in the solar neighborhood relative to the Sun)

There are complications with using this method:

- Near $\ell = 90°$ and $\ell = 270°$ the method tends to break down because v_r becomes rather insensitive to changes in distance from the Sun.
- Within $\sim 20°$ of the galactic center, clouds can start having markedly noncircular motions, so the assumptions of the tangent point method break down and become no longer valid. This is potentially due to gravitational perturbations from the central bar.
- The method is not usable at all between $90° < \ell < 270°$ (i.e. pointing away from the galactic center) because there is no unique orbit for which a maximum radial velocity can be observed.

For data beyond R_0, we have to construct the rotation curve using a different approach. We can use velocities of globular clusters, but we have to use objects for which we can obtain a reliable distance estimate (i.e. parallaxes, Cepheids, etc.).

104. What is the density profile of a self-gravitating isothermal gas sphere? What is the corresponding rotation curve for this system?

The equation of state for an isothermal ($T = \text{const.}$) gas is the ideal gas law $P = \rho kT/m$. Starting with hydrostatic equilibrium[119,126]:

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{GM(r)\rho(r)}{r^2} \quad \longrightarrow \quad \frac{kT}{m}\frac{\mathrm{d}\ln\rho}{\mathrm{d}r} = -\frac{GM(r)}{r^2} \tag{7.22}$$

$$r^2\frac{\mathrm{d}\ln\rho}{\mathrm{d}r} = -\frac{GM(r)m}{kT} \tag{7.23}$$

Then differentiating with respect to r and using mass conservation ($\mathrm{d}M/\mathrm{d}r = 4\pi r^2\rho$):

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}\ln\rho}{\mathrm{d}r}\right) = -\frac{4\pi Gm}{kT}r^2\rho \tag{7.24}$$

Now if we consider the probability density in velocity space, we have a **Maxwell-Boltzmann distribution**. We consider only one dimension since we can only measure the velocity dispersion σ in one dimension:

$$f(E)\mathrm{d}E \propto \exp\left(-\frac{E}{kT}\right)\mathrm{d}E \quad \longrightarrow \quad f(v_i)\mathrm{d}v_i \propto \exp\left(-\frac{mv_i^2}{2kT}\right)\mathrm{d}v_i \tag{7.25}$$

Compare this to the standard form of a gaussian function with standard deviation $\sigma_i = \sigma$:

$$\exp\left(-\frac{v_i^2}{2\sigma_i^2}\right) \quad \longrightarrow \quad \sigma^2 = \frac{kT}{m} \tag{7.26}$$

Plugging this back in, we have

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}\ln\rho}{\mathrm{d}r}\right) = -\frac{4\pi G}{\sigma^2}r^2\rho \tag{7.27}$$

We can guess a power-law solution of the form $\rho = Cr^{-b}$, which results in

$$\frac{\mathrm{d}\ln\rho}{\mathrm{d}r} = -\frac{b}{r} \;\;,\;\; \frac{\mathrm{d}}{\mathrm{d}r}(-br) = -b \quad \longrightarrow \quad b = \frac{4\pi G}{\sigma^2}Cr^{2-b} \tag{7.28}$$

$$\therefore \;\; b = 2 \;\;,\;\; C = \frac{\sigma^2}{2\pi G} \tag{7.29}$$

Giving us a density profile for the isothermal sphere of

$$\boxed{\rho(r) = \frac{\sigma^2}{2\pi Gr^2}} \tag{7.30}$$

This is a simple r^{-2} power law (see Figure 7.6).

Using the Poisson equation to back out a potential:

$$\nabla^2\Phi = 4\pi G\rho \quad \longrightarrow \quad \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\Phi}{\partial r}\right) = \frac{2\sigma^2}{r^2} \tag{7.31}$$

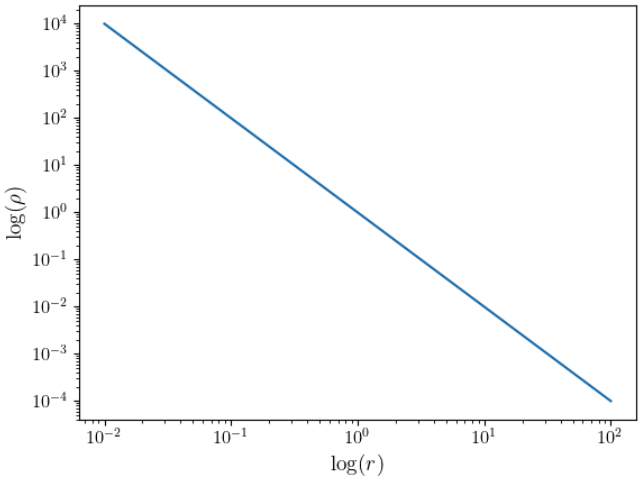


Figure 7.6: Plot of the density profile of an isothermal sphere.

Integrating twice, we have

$$r^2 \frac{\partial \Phi}{\partial r} = \int \mathrm{d}r\, 2\sigma^2 = 2\sigma^2 r + A \tag{7.32}$$

$$\Phi = \int \mathrm{d}r \left(\frac{2\sigma^2}{r} + \frac{A}{r^2} \right) = 2\sigma^2 \ln r - \frac{A}{r} + B \tag{7.33}$$

$$\tag{7.34}$$

Choosing $A = 0$ to prevent Φ from blowing up as $r \to 0$ and $B = -2\sigma^2 \ln r_0$

$$\boxed{\Phi(r) = 2\sigma^2 \ln\left(\frac{r}{r_0}\right)} \tag{7.35}$$

This results in a velocity profile that is **flat** and **independent of radius**:

$$\frac{v^2}{r} = -\frac{\mathrm{d}\Phi}{\mathrm{d}r} = -\frac{2\sigma^2}{r} \quad \longrightarrow \quad \boxed{v(r) = \sqrt{2}\sigma} \tag{7.36}$$

Another way of seeing this is calculating the mass, which we see becomes proportional to r:

$$M(r) = \int_0^r \mathrm{d}r\, 4\pi r^2 \rho(r) = \frac{2\sigma^2}{G} r \tag{7.37}$$

Thus, $v^2 = GM/r = 2\sigma^2$, as before.

105. What is two-body relaxation? For a self-gravitating cluster of N objects, each of mass m, with a velocity dispersion σ, what is the relaxation time? How long does it take for a massive object ($M \gg m$) to sink to the bottom of a cluster potential well?

Two-body relaxation is the gradual process by which a system becomes **virialized** over time by numerous 2-body interactions between particles in the system, driving the system to a stable equilibrium state where it no longer expands or contracts. Each 2-body interaction can be categorized as a **"strong"** or a **"weak" encounter** based on how close the 2 bodies get to each other and how much they affect each other's velocities. In a typical statistical mechanics example, like a collisional gas, the interactions between particles (gas molecules/atoms) are dominated by short-range (strong) interactions (i.e. collisions) governed by electromagnetic forces. Since the forces are short-ranged, the molecules travel at nearly constant velocity between interactions. However, stars in a galaxy are fundamentally different because the gravitational force is always attractive and thus acts on long ranges. Assuming a constant density of stars, the gravitational force per spherical volume goes like $r^{-2} \times r^3 = r$, so the most distant stars actually dominate the dynamics. Thus the stars will accelerate smoothly and gradually in response to density fluctuations at different points in the galaxy. **Note the similarities and differences between the following derivation and the derivation done for bremsstrahlung radiation in §5.Q74.**

For stars in a galaxy, strong encounters are extremely rare since the gravitational effects of the closest companion stars is sub-dominant. Weak encounters are much more common, with each one only slightly perturbing the velocity of the star. Because of this, the relaxation times for 2-body relaxation of a stellar population in a galaxy are unphysically large (orders of magnitude larger than the age of the universe). To see why, let's try to estimate it.

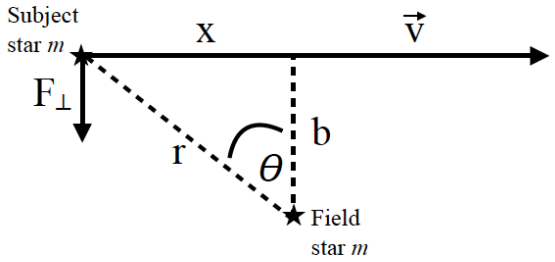


Figure 7.7: Diagram of a gravitational encounter between two stars[127].

What is the relaxation time?[126,127] For any encounter (strong or weak), the change in velocity $\delta v = |\delta \mathbf{v}|$ can be estimated by assuming a star of mass m passes by another star of the same mass (for simplicity, taking m as some average stellar mass). We also assume the second star is stationary such that the deflection $\delta \mathbf{v} \perp \mathbf{v}$, and $\delta v \ll v$ (see Figure 7.7).

Then, the component of force perpendicular to the direction of motion (where $x = vt$) is

$$F_\perp = \frac{Gm^2}{b^2 + x^2} \cos\theta = \frac{Gm^2 b}{(b^2 + x^2)^{3/2}} = \frac{Gm^2}{b^2[1 + (vt/b)^2]^{3/2}} \tag{7.38}$$

Using Newton's second law:

$$F_\perp \approx m\frac{\delta v}{\delta t} \quad \longrightarrow \quad \delta v = \frac{1}{m}\int_{-\infty}^{+\infty} \mathrm{d}t\, F_\perp = \frac{Gm}{b^2}\int_{-\infty}^{+\infty} \mathrm{d}t \left[1+\left(\frac{vt}{b}\right)^2\right]^{-3/2} \tag{7.39}$$

Using a substitution $s = vt/b$, $\mathrm{d}t = (b/v)\mathrm{d}s$:

$$\delta v = \frac{Gm}{bv}\int_{-\infty}^{+\infty} \frac{\mathrm{d}s}{(1+s^2)^{3/2}} \quad \longrightarrow \quad \delta v = \frac{2Gm}{bv} \tag{7.40}$$

(where the integral can be solved with the substitution $s = \tan u$, $\mathrm{d}s = \sec^2 u\ \mathrm{d}u$).

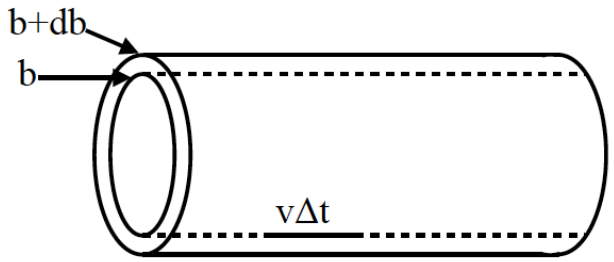


Figure 7.8: The shell sweeped out by a star as it travels along and interacts with other stars[127].

Now we consider a star that undergoes a number of these encounters as it travels along its orbit. It sweeps out the area of a cylindrical shell $2\pi b\mathrm{d}b$ over a length $v\Delta t$, having gravitational encounters with n stars per volume of the shell (see Figure 7.8). Thus, the total number of encounters is

$$N = 2\pi b\mathrm{d}b v\Delta t n \tag{7.41}$$

And each encounter changes the velocity by δv given by (7.40). The star makes a random walk through the galaxy, so after some number of encounters, while its *average* perpendicular velocity is 0, is *variance* is not, i.e.

$$\Delta v_{\mathrm{tot}}^2 = \sum_i (\delta v_i)^2 \approx \int_{b_{\min}}^{b_{\max}} \mathrm{d}b\, (2\pi b v\Delta t n)\left(\frac{2Gm}{bv}\right)^2 \tag{7.42}$$

$$= \frac{8\pi G^2 m^2 n\Delta t}{v}\int_{b_{\min}}^{b_{\max}} \frac{\mathrm{d}b}{b} \tag{7.43}$$

$$= \frac{8\pi G^2 m^2 n\Delta t}{v}\ln\left(\frac{b_{\max}}{b_{\min}}\right) \tag{7.44}$$

The minimum and maximum impact parameters can be estimated as $b_{\max} \sim R_{\mathrm{gal}}$ and $b_{\min} \sim 2Gm/v^2$ (i.e. setting $\delta v = v$ in (7.40) for strong interactions). This evaluates to $\ln\Lambda \sim 25$ (where $\Lambda \sim Rv^2/Gm \sim N$). Then we can estimate the relaxation time by setting $\Delta v_{\mathrm{tot}} = v$, i.e. the change in velocity imparted by the interactions is of the same order as the star's velocity ($v_\parallel \sim v_\perp$):

$$v^2 \approx \frac{8\pi G^2 m^2 n\ln\Lambda t_{\mathrm{relax}}}{v} \quad \longrightarrow \quad t_{\mathrm{relax}} \approx \frac{v^3}{8\pi G^2 m^2 n\ln\Lambda} \tag{7.45}$$

Substituting $N = \frac{4}{3}\pi R^3 n$ and $\langle v^2 \rangle = \sigma^2 = GNm/R$ from the virial theorem

$$n = \frac{N}{\frac{4}{3}\pi(GNm/\sigma^2)^3} = \frac{3\sigma^6}{4\pi G^3 N^2 m^3} \tag{7.46}$$

We have:

$$\boxed{t_{\text{relax}} \approx \frac{GN^2 m}{6\sigma^3 \ln N} \propto N^2} \tag{7.47}$$

Or, in terms of the crossing time $t_{\text{cross}} \approx R/\sigma = GNm/\sigma^3$:

$$t_{\text{relax}} \approx \frac{N}{6\ln N} t_{\text{cross}} \tag{7.48}$$

These relaxation timescales are unreasonably large for a typical galaxy where $N \approx 10^{11}$ and $t_{\text{cross}} \approx 10$ Gyr (so $t_{\text{relax}} \sim 10^9$ Gyr), whereas they can be reasonable for globular clusters with $N \approx 10^5$ and $t_{\text{cross}} \approx 1$ Myr (so $t_{\text{relax}} \sim 1$ Gyr). Thus, the stars in a galaxy must relax in some other way, i.e. **violent relaxation** (see §7.Q113). However, 2-body relaxation is feasible for the gas since, as explained, its collisional interactions are fundamentally different.

Sinking to the bottom of a potential well[128]. The free-fall timescale can be calculated by using conservation of energy. Say a particle falls from initial radius r_0 to some smaller radius r:

$$\frac{1}{2}mv(r)^2 = GMm\left(\frac{1}{r} - \frac{1}{r_0}\right) \quad \longrightarrow \quad v^2 = 2GM\left(\frac{1}{r} - \frac{1}{r_0}\right) = \frac{8\pi}{3} G r_0^2 \rho \left(\frac{r_0}{r} - 1\right) \tag{7.49}$$

We can integrate this from $r = r_0$ to $r = 0$ to find the free-fall time

$$\int_{r_0}^{0} \frac{\mathrm{d}r/r_0}{\sqrt{r_0/r - 1}} = \sqrt{\frac{8\pi}{3} G\rho} \int_0^{t_{\text{ff}}} \mathrm{d}t \tag{7.50}$$

$$\boxed{t_{\text{ff}} = \sqrt{\frac{3\pi}{32 G\rho}}} \tag{7.51}$$

The integral is a bit tricky and requires trig substitutions to solve. But what one can notice is that the integral is dimensionless, so we can say the integral is some constant of order unity and drop it, along with the other constants, to get

$$t_{\text{ff}} \sim \frac{1}{\sqrt{G\rho}} \tag{7.52}$$

(one can also use dimensional analysis to arrive at the same result). Then, rewriting in terms of N, m, and σ:

$$t_{\text{ff}} \sim \sqrt{\frac{R^3}{GNm}} \sim \sqrt{\frac{(GNm)^3}{GNm\sigma^6}} \quad \longrightarrow \quad \boxed{t_{\text{ff}} = \frac{GNm}{\sigma^3} \propto N} \quad , \quad t_{\text{ff}} \sim \frac{\ln N}{N} t_{\text{relax}} \tag{7.53}$$

Thus, t_{ff} is a factor of $\sim N$ smaller than t_{relax}, meaning a massive object will sink before having any strong encounters.

106. What is the "Local Group"?

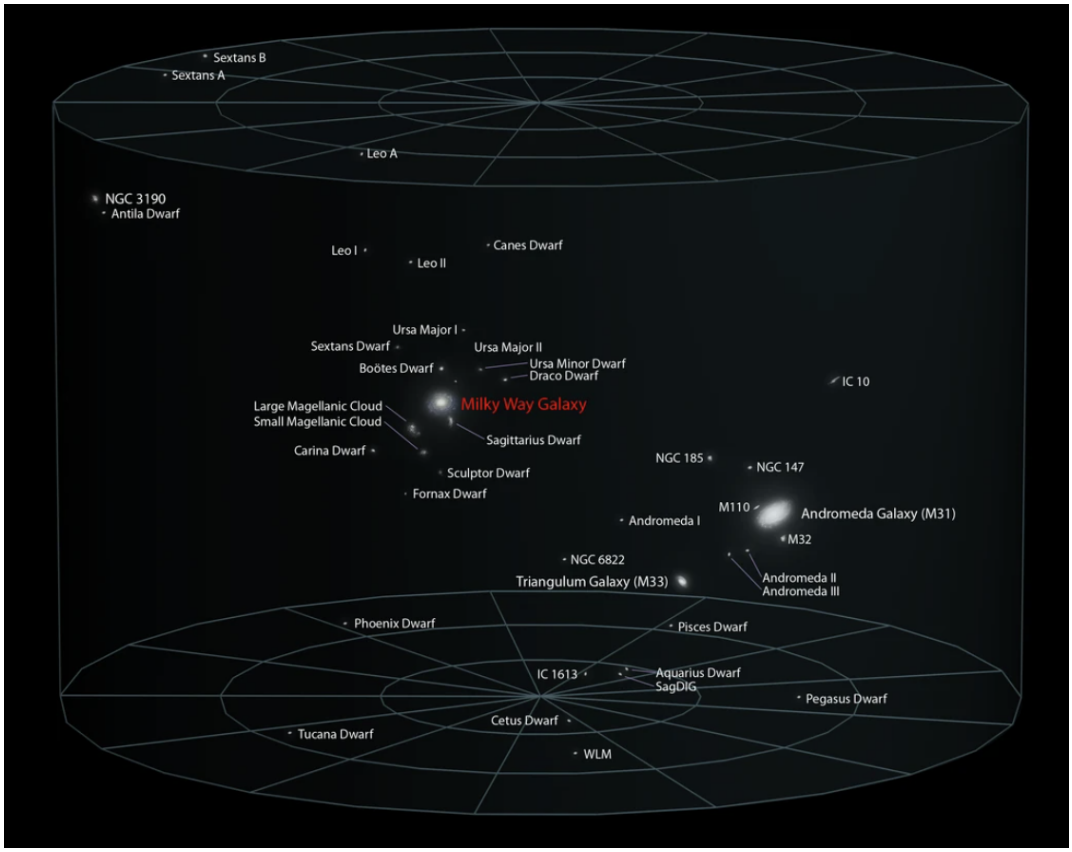


Figure 7.9: Map of the Local Group. Credit: SpaceWiki.

The **Local Group** is the galaxy group that the Milky Way belongs to[11] (see Figure 7.9). There are about 35 other galaxies within ∼1 Mpc of the Milky Way that are part of the Local Group, and its **zero velocity surface** (where galaxies would turn around if their velocities were suddenly directed outwards) is ∼1.2 Mpc from its barycenter. The most prominent member galaxies are:

- The Milky Way
- Andromeda (M31)
- The Triangulum Galaxy (M33)

The next most luminous are the Large and Small Magellanic Clouds, which are 2 of the 13 irregular galaxies in the Local Group. Finally, the rest are very small and faint dwarf ellipticals and dwarf spheroidals. These have all accumulated around the Milky Way and Andromeda.

The Milky Way and Andromeda are also approaching each other with a relative velocity of 119 km s^{-1}, and they are about 770 kpc apart, meaning they will collide in roughly 6.3 Gyr. The total mass of the Local Group is $\sim 4 \times 10^{12}\ M_\odot$ with a mass-to-light ratio $M/L = 57(M_\odot/L_\odot)$. Compared to $M/L \sim 3(M_\odot/L_\odot)$ for the luminous matter in the Milky Way. This suggests there is a large fraction of dark matter in the Local Group.

Why is $M/L \neq 1$ for the baryonic matter in the Milky Way? It would be if all of the stars were Sun-like, but the initial mass function favors low-mass stars, which have much lower luminosities.

107. What is the morphology-density relation for galaxies? Give several possible explanations for this trend.

The **morphology-density relation** is a correlation between galaxy morphology and the density of the environment they tend to be found in. Specifically, early-type galaxies (ellipticals) tend to be found more commonly in denser environments (like galaxy clusters) compared to late-type galaxies (spirals), which are more commonly found in less dense environments[107] (see Figure 7.10).

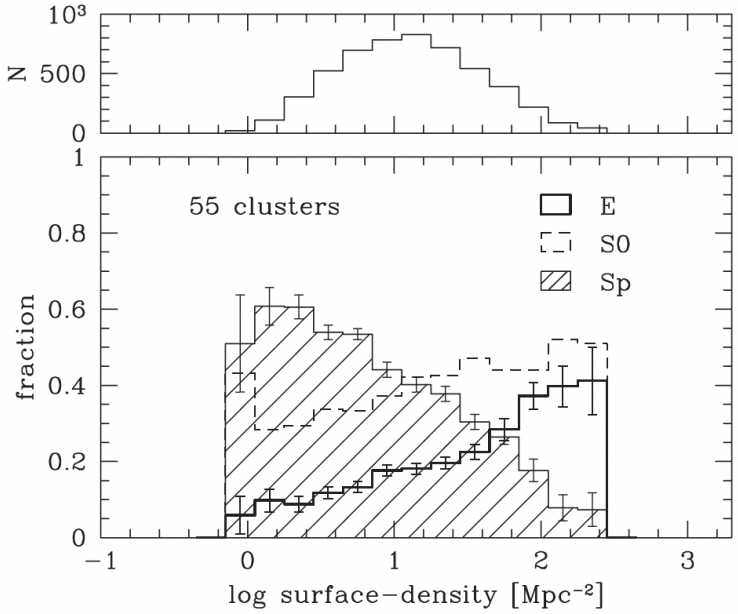


Figure 7.10: The morphology-density relationship[107].

This correlation was realized as early as the 1930s by Edwin Hubble and Milton Humason, but was quantified more accurately by Alan Dressler in the 1980s. He found that the fraction of spiral galaxies decreases from ∼60% in the lowest density regions to ∼10% in the highest density regions, while the elliptical fraction shows the opposite behavior.

Potential explanations[107]: In the standard paradigm, galaxies are believed to form as spirals, and major mergers can convert them into ellipticals. However, if the gas around an elliptical merger remnant is able to cool, it may form a new disk. Thus, there are two potential explanations for why galaxies in clusters are preferentially early-type:

1. More massive halos experience more major mergers
2. Gas is not able to cool as efficiently in more massive halos

At first glance, 1 seems reasonable (denser environments lead to more mergers), but this is actually incorrect. CDM cosmology predicts halos of all masses to have very similar average merger histories. To see why, consider this: the typical velocity dispersion of galaxies in a cluster is much larger than the typical internal velocity dispersion of each galaxy. So all encounters are "high-speed" encounters from the perspective of the galaxies, making them unlikely to become merger events. Instead, 2 is the most likely explanation. This may be the result of AGN feedback heating the gas, or it may be a result of the gas in these environments being evacuated due to ram pressure stripping.

108. What is the "Schechter luminosity function"? What is the luminosity of a typical bright galaxy? How does the luminosity function of galaxies compare to the mass function of halos from simulations?

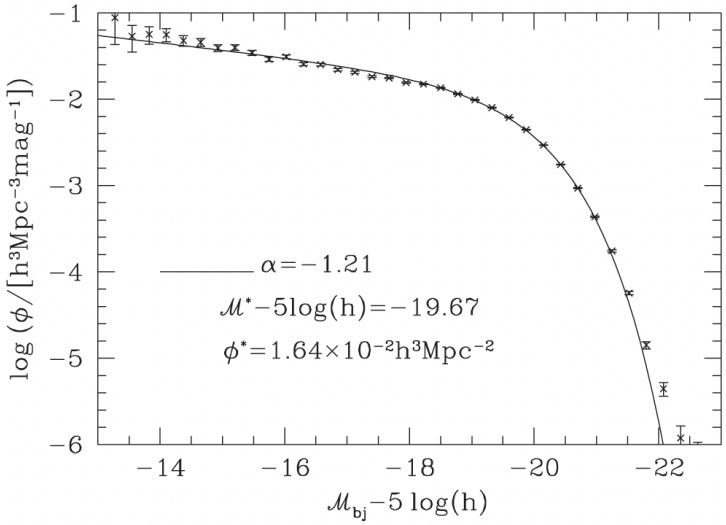


Figure 7.11: The Schechter luminosity function[107].

The **Schechter luminosity function** is a function that gives the number density of galaxies with luminosities between L and $L + \mathrm{d}L$. It is defined as[107,126]:

$$\boxed{\phi(L)\mathrm{d}L = \frac{\phi_*}{L_*}\left(\frac{L}{L_*}\right)^{\alpha} \exp\left(-\frac{L}{L_*}\right)\mathrm{d}L} \tag{7.54}$$

where $\alpha \approx -1.1$, $\phi_* \approx 5 \times 10^{-3} h_{70}^3$ Mpc^{-3}, and $L_* \approx 3 \times 10^{10}\ L_\odot$ (see Figure 7.11). Note that these parameters are just approximations for the general galaxy population of the unverse—in reality, the luminosity functions of galaxies in different clusters and fields can give different parameters. It's defined such that integrating over $\mathrm{d}L$ gives the mean number density n (galaxies Mpc^{-3}):

$$n = \int_0^\infty \mathrm{d}L\, \phi(L) = \phi_* \Gamma(\alpha + 1) \tag{7.55}$$

And integrating $\phi(L)L$ gives the mean luminosity density $\mathscr{L}$ ($L_\odot$ Mpc^{-3}):

$$\mathscr{L} = \int_0^\infty \mathrm{d}L\, \phi(L)L = \phi_* L_* \Gamma(\alpha + 2) \tag{7.56}$$

The luminosity of a typical bright galaxy is given by $L_* \sim 10^{10}\ L_\odot$ (which is also ∼luminosity of the Milky Way).

Halo mass functions from simulations (i.e. see question on **Press-Schechter theory**; §7.Q119) typi-

cally follow a more simple power-law trend where

$$\frac{\mathrm{d}n}{\mathrm{d}M} \propto \begin{cases} M^{-2} & M \ll M_* \\ 0 & M \gg M_* \end{cases} \tag{7.57}$$

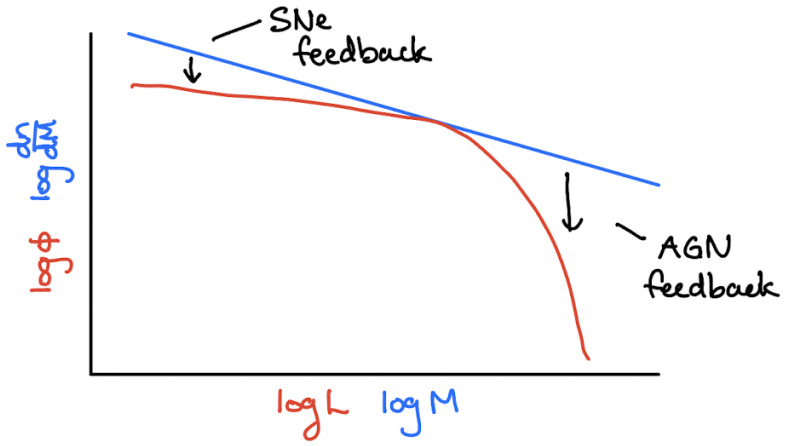


Figure 7.12: The Schechter function compared to the halo mass function from simulations.

This shows that there is an underprediction of the galaxy number density both at the low end and high end of the luminosity function. Both are thought to be due to suppressed star formation, but for different reasons. The suppression at the low end is thought to be caused by feedback from supernovae, which blows out gas that is needed to form new stars. The suppression at the high end is thought to be due to AGN feedback, which can similarly blow out gas and dust[119].

109. Describe Hubble's classification scheme for galaxies and explain why it is useful.

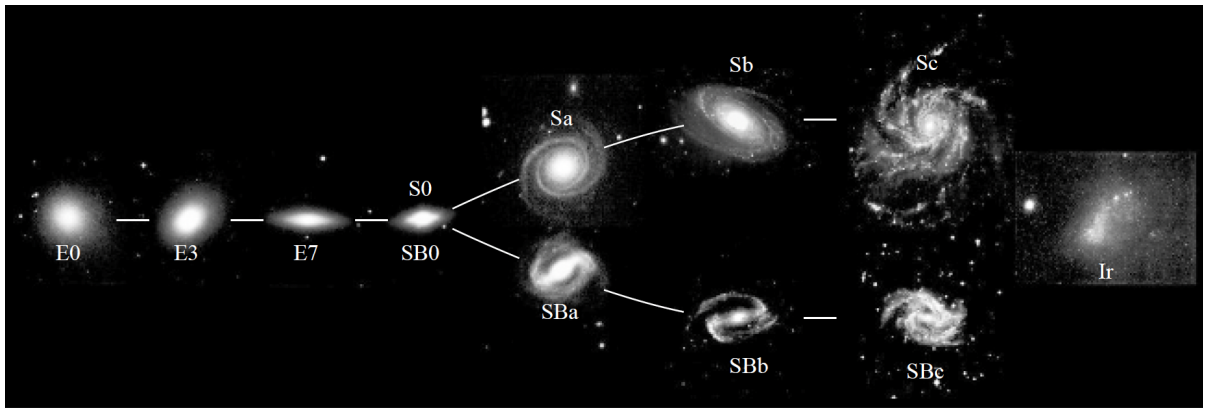


Figure 7.13: Hubble's tuning fork diagram[11]

Also known as the "tuning fork" due to its two-pronged shape (see Figure 7.13), the **Hubble sequence** subdivides galaxies into three main categories, with some further subdivisions, based on their observed morphologies[11]:

- **E**: Ellipticals
  - **S0/SB0**: Lenticulars (transitional between E and S; may also have a bar)
- **S**: Spirals
  - **S**: Normal Spirals
  - **SB**: Barred Spirals
- **Irr**: Irregulars

Ellipticals are spherically/ellipsoidally distributed with red colors indicative of older stellar populations. They are pressure supported with random motions, and they have relatively little gas and dust. Their brightness profiles follow a de Vaucouleurs distribution. They span a huge dynamic range in mass (see §7.Q102).

Spiral galaxies have a disklike structure with blue colors indicative of younger stellar populations. They are rotationally supported, and have much more gas and dust than ellipticals. Their brightness profiles follow an exponential distribution. They tend to span a smaller range in physical properties than ellipticals do (i.e. mass, see §7.Q102).

Thus, the Hubble sequence is useful in categorizing general trends in the galaxy population (support, color, age, gas content, brightness profile, mass, etc.) simplying by looking at morphologies.

E's are further subcategorized based on their *ellipticity* (how elongated they are):

- **E0**: Nearly spherical

  ↓
- **E7**: Highly elongated

S/SB's are further subcategorized based on their bulge-to-disk ratio ($L_{\text{bulge}}/L_{\text{disk}}$), how tightly wound their spiral arms are, and how smoothly distributed the stars in their arms are:

- **Sa/SBa**: Most prominent bulges ($L_{\text{bulge}}/L_{\text{disk}} \sim 0.3$), most tightly wound arms (pitch angles $\sim 6°$), smoothest distribution of stars

  ↓
- **Sc/SBc**: Smallest bulges ($L_{\text{bulge}}/L_{\text{disk}} \sim 0.05$), most loosely would arms ($\sim 18°$), most clumpy distribution of stars

Irr's are subcategorized based on if there are any hints of some organized structure:

- **Irr I**: Some hint of an organized structure (i.e. spiral arms)
- **Irr II**: Absolutely no hint of organization at all

The Irr subclassification has since been eliminated in favor of classifying Irr I's as a further extension of the S's up to **Sd/SBd**, **Sm/SBm**, and **Im**. These tend to be much less massive than the Sa–Sc types and are thus sometimes referred to as **dwarf spirals**. The truly irregular galaxies are then just designated as **Ir**.

Another modification has been the subdivision of S0's based on how much dust absorption is in their disks:

- **S0$_1$/SB0$_1$**: No discernable dust in the disk

  ↓
- **S0$_3$/SB0$_3$**: Significant amounts of dust in the disk

Thus, the full modern sequence is:

- E0, E1, ..., E6, E7, S0$_1$, S0$_2$, S0$_3$, Sa, Sab, Sb, Sbc, Sc, Scd, Sd, Sm, Im, Ir
- E0, E1, ..., E6, E7, SB0$_1$, SB0$_2$, SB0$_3$, SBa, SBab, SBb, SBbc, SBc, SBcd, SBd, SBm, Im, Ir

Alternative names for things

- Sometimes ellipticals are referred to as **"early-type"** galaxies while spirals are **"late-type"** galaxies. This terminology is primarily used by boomers and Katiya Fosdick. It originates from the fact that Hubble thought the tuning-fork diagram could be interpreted as an evolutionary sequence of galaxies, from ellipticals to spirals. We know today that this is not correct, and in fact the true evolutionary sequence likely goes in the opposite direction.
- Sometimes elliptical/spiral galaxies are referred to as **"spheroid"/"disk" galaxies** by hipsters.

Dwarf galaxies[11,107]: Dwarf galaxies have a diverse structure and do not tend to fall cleanly into the Hubble sequence, thus they have their own classifications:

- **dE**: Dwarf Ellipticals
- **dSph**: Dwarf Spheroidals
- **BCD**: Blue Compact Dwarfs

Dwarf ellipticals and dwarf spheroidals tend to be gas-poor systems with no young stars and low metallicities. BCDs, on the other hand, are gas-rich, and as the name suggests, are unusually blue, suggesting vigorous star formation.

There are also **cD galaxies**—see §7.Q111 for a description.

110. What is the "Tully-Fisher relation"? Derive this relation beginning with the virial theorem and using simplifying assumptions about galaxies. How can this relation be used to determine H_0?

The **Tully-Fisher relation** is a scaling relation between the luminosity L and maximum rotational velocity v_c of a spiral galaxy that says $L \propto v_c^4$. **This only works for spiral galaxies**. We can derive this by starting with the virial theorem[119]:

$$2T + U = 0 \tag{7.58}$$

We can write out the kinetic and potential energies in terms of a sum (i.e. over each star in a galaxy), which we can then rewrite in terms of some characteristic velocity v_c and radius R_c (see §7.Q101 for a detailed explanation of this part):

$$T = \frac{1}{2} M v_c^2 \ , \ U = -\frac{GM^2}{R_c} \tag{7.59}$$

Such that

$$v_c^2 = \frac{GM}{R_c} \tag{7.60}$$

Now we need to relate M and R_c to the luminosity L, which we can do if we make some simplifying assumptions. Let's assume that the galaxy is disk-shaped and is fully face-on such that its luminosity is given in terms of its average surface brightness I_0 (luminosity density) by

$$L \sim \pi R_c^2 I_0 \ \longrightarrow \ R_c \sim \sqrt{\frac{L}{\pi I_0}} \tag{7.61}$$

Now let's also assume the galaxy has a constant mass-to-light ratio $\Gamma = M/L$. Then, we can write (7.60) as

$$v_c^2 \sim \frac{G\Gamma L}{\sqrt{L/\pi I_0}} = G\Gamma\sqrt{\pi I_0 L} \tag{7.62}$$

Therefore,

$$L \sim \frac{1}{G^2\Gamma^2 \pi I_0} v_c^4 \tag{7.63}$$

Assuming Γ and I_0 are constant, then we get the Tully-Fisher relation

$$\boxed{L \propto v_c^4} \ \longrightarrow \ \boxed{v_c = 220 \text{ km s}^{-1} \left(\frac{L}{10^{10} L_\odot} \right)^{1/4}} \tag{7.64}$$

Where in the second equation values for the Milky Way are used as the normalization. The Tully-Fisher relation can be used as a **distance indicator** since the intrinsic luminosity L can be compared to the observed flux F via $L = 4\pi d_L^2 F$ (where d_L is the luminosity distance). The recessional velocity is easily obtained from the redshift. Thus, one can measure an object's redshift and v_c, back out L from the Tully-Fisher relation, and get the distance and recessional velocity. Then one can measure $H_0 = v_r/d$.

111. What are cD galaxies and where are they found? How might they be formed?

cD galaxies[11]: An extension of the elliptical class, these immensely bright galaxies are giant, sometimes up to 1 Mpc in diameter. The "D" here stands for "diffuse." They have a central region with high surface brightness surrounded by an extended, diffuse envelopse of much lower surface brightness. Hence, they typically have best-fitting Sérsic indices $\gg 4$. They are the most massive known galaxies with stellar masses in excess of $10^{12}\ M_\odot$, and may exceed mass-to-light ratios of $750h_{70}(M_\odot/L_\odot)$, implying vast amounts of dark matter.

They are typically found as the **Brightest Cluster Galaxy (BCG)** of large, dense galaxy clusters, located at the center of the cluster. This introduces some ambiguity in the outskirts of the galaxy as to where exactly the boundary should lie between the envelope that's part of the galaxy and the 'intracluster light' (ICL) composed of stars that belong to the cluster rather than one particular galaxy.

They are thought to have formed through the accretion of multiple galaxies in a cluster, in a process known as **"galactic cannibalism."**

- The velocity dispersion of galaxies in a cluster is typically much larger than the internal velocity dispersions of the galaxies themselves. Thus, galaxies in a cluster are unlikely to merge because they only experience high-speed encounters.
- There is one exception—dynamical friction causes the galaxies to lose energy in each interaction, causing them to gradually "sink" towards the center of the gravitational potential well of the cluster, which is where the central galaxy resides.
- Thus, over time, the central galaxy will tend to accrete these galaxies that have lost too much energy from dynamical friction.
- The dynamical friction timescale is shorter for more massive galaxies, so massive satellites will tend to be accreted first.

Nearby cDs frequently appear to have multiple nuclei, supporting this accretion scenario.

112. What is the "Faber-Jackson relation" for elliptical galaxies? How is this different than the "fundamental plane" for elliptical galaxies?

The **Faber-Jackson relation** is a scaling relation between the luminosity L and velocity dispersion σ of an elliptical galaxy that says

$$\boxed{L \propto \sigma^4} \tag{7.65}$$

This only works for elliptical galaxies (or the bulges of spiral galaxies). Note the similarities and differences with the Tully-Fisher relation for spiral galaxies ($L \propto v_c^4$). The power law index (4) is the same, but here we are using the velocity dispersion σ instead of the rotational velocity v_c. It can be derived the same way that we derived Tully-Fisher in §7.Q110. The Faber-Jackson relation notably has a high scatter since it covers a galaxy population with such a diverse range of properties (elliptical galaxies have a larger dynamic range in properties like mass compared to spiral galaxies).

To address the larger scatter, a more accurate scaling relation was developed called the **Fundamental Plane** which uses both the velocity dispersion σ and the effective radius of the galaxy R_e. One representation of this relationship looks like:

$$\boxed{L \propto \sigma^{2.65} R_e^{0.65}} \tag{7.66}$$

In this context, galaxies can be visualized as residing on some 2-dimensional "surface" within the 3-dimensional space of σ, R_e, and L.

113. What is a violent relaxation? How does the phase space distribution function it produces differ from that of an isothermal gas? How does the timescale for violent relation compare to that for weak 2-body interactions?

Violent relaxation is the process by which stars' energies are changed by being in a potential Φ that varies over both time and space. A star can gain or lose energy by passing through a potential well that is becoming shallower or deeper over time (see Figure 7.14). How a star's energy changes in general will depend in a complex way on the initial position and velocity of the star, even during a simple spherical collapse. The overall effect is to spread out the range of energies of the stars[126,119].

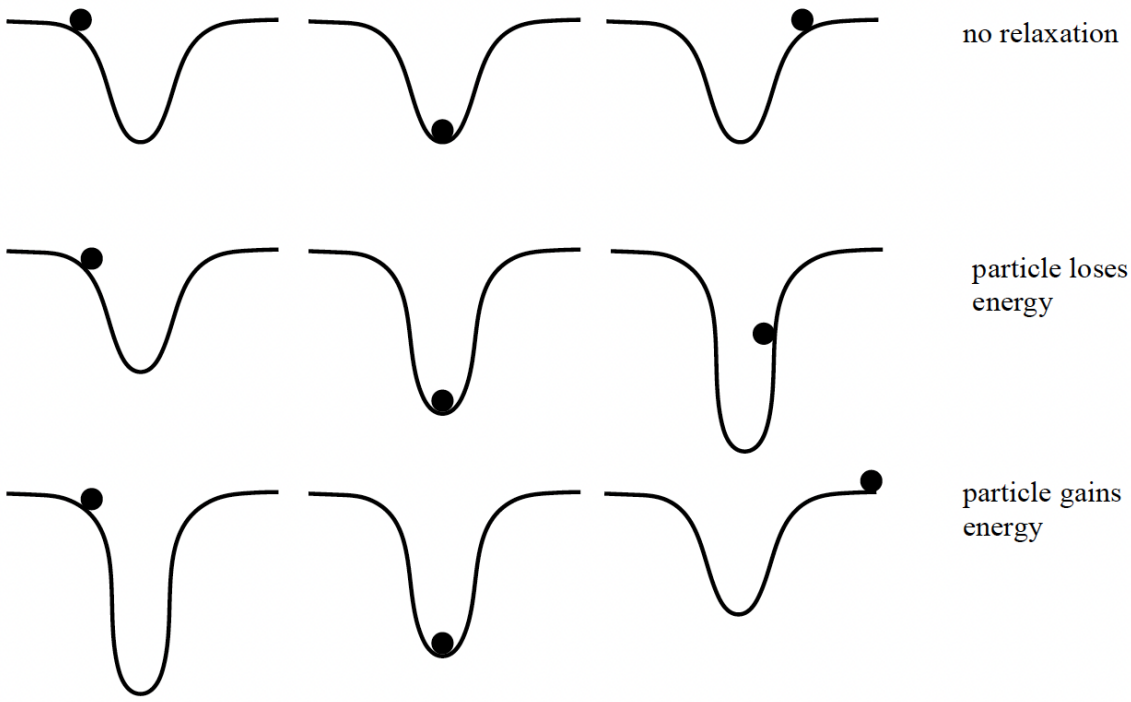


Figure 7.14: Examples of particles in a time-varying potential[127]

Phase space distribution function: Consider the energy per unit mass of a star $\mathscr{E} = E/m$:

$$\mathscr{E} = \frac{1}{2}v^2 + \Phi \tag{7.67}$$

Differentiating:

$$\frac{\mathrm{d}\mathscr{E}}{\mathrm{d}t} = \frac{\mathrm{d}\mathscr{E}}{\mathrm{d}\mathbf{v}}\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} + \frac{\mathrm{d}\mathscr{E}}{\mathrm{d}\Phi}\frac{\mathrm{d}\Phi}{\mathrm{d}t} \tag{7.68}$$

For conservative forces, $\mathrm{d}\mathbf{v}/\mathrm{d}t = -\nabla\Phi$:

$$\frac{\mathrm{d}\mathscr{E}}{\mathrm{d}t} = -\mathbf{v}\cdot\nabla\Phi + \frac{\mathrm{d}\Phi}{\mathrm{d}t} \tag{7.69}$$

And using the chain rule:

$$\frac{\mathrm{d}\Phi}{\mathrm{d}t} = \frac{\partial\Phi}{\partial\mathbf{x}}\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} + \frac{\partial\Phi}{\partial t} = \mathbf{v}\cdot\nabla\Phi + \frac{\partial\Phi}{\partial t} \tag{7.70}$$

Therefore,

$$\frac{\mathrm{d}\mathscr{E}}{\mathrm{d}t} = \frac{\partial\Phi}{\partial t} \tag{7.71}$$

This shows that the change in energy is independent of the mass of the star. This is in contrast to statistical mechanics where collisional relaxation tends to pump energy from the most massive particles to the least massive particles, establishing the equipartition of energy and causing more massive particle to tend to sink towards the center of the potential. Thus, an isothermal gas follows a **Maxwell-Boltzmann distribution** whereas the phase space distribution of violent relaxation is more uniform[107,126].

Timescale[119,127]: The timescale of violent relaxation can be estimated by averaging the change in energy over all particles:

$$t_{VR} = \left\langle \frac{\mathscr{E}^2}{(\mathrm{d}\mathscr{E}/\mathrm{d}t)^2} \right\rangle^{1/2} = \left\langle \frac{\mathscr{E}^2}{(\partial\Phi/\partial t)^2} \right\rangle^{1/2} \longrightarrow \boxed{t_{\mathrm{VR}} \sim \left\langle \frac{\Phi^2}{(\partial\Phi/\partial t)^2} \right\rangle^{1/2} \sim t_{\mathrm{ff}} \propto N} \tag{7.72}$$

Where in the penultimate step we estimated $\mathscr{E} \sim \Phi$ for any given star. This is similar to the free-fall timescale t_{ff} since it is the timescale over which the potential changes during the collapse. Therefore, the timescales for violent relaxation are much faster than those of 2-body relaxation and are feasible for the virialization of galaxies and galaxy clusters. The proportionality with N comes from (7.53).


Wii
Wii Are Resorting to Violence
12 Deaths Help Them
BONUS
Wii F■king Gun
INSIDE!
How is this allowed
made with mematic
EVERYONE 10+
ESRB
Holy Sh■t

114. Why do some galaxies have prominent spiral structure and others do not? Briefly describe the density wave theory of spiral structure.

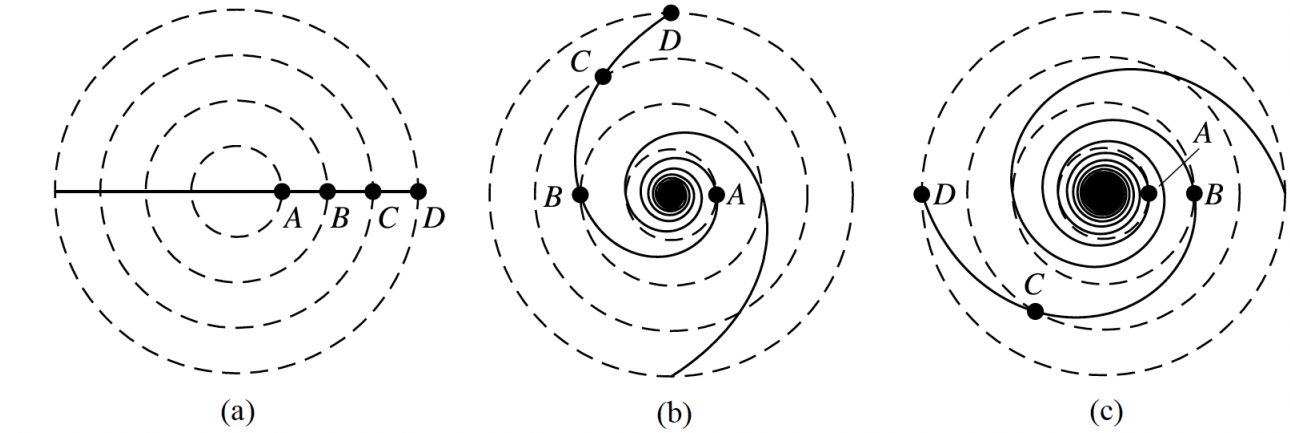


Figure 7.15: Illustration of the spiral winding problem[11]. Points A, B, C, and D start out in a line and are tracked over time as they rotate at different speeds, creating spiral arms that quickly wind up.

We can try to explain the spiral structure by differential rotation in the galactic disk. That is, because stars and gas closer to the galactic center rotate faster and those further out rotate slower, it presumably should lead to the natural formation of the spiral arms. This would mean that the spiral arms are *material arms* consisting of the same fixed set of stars, gas, and dust. Under this assumption, the timescale for one rotation is

$$t_{\rm rot} = \frac{2\pi r}{v} = 2\pi\sqrt{\frac{r^3}{GM}} \tag{7.73}$$

With typical numbers, $r \sim 10$ kpc and $M \sim 10^{11}\ M_\odot$, so $t_{\rm rot} \sim 10^8$ yr. Over the age of the universe, these spiral arms should have rotated ∼10s to 100s of times. However, after only a few rotations this would result in the spiral arms getting too wound up and the spiral structure would be lost. This is known as the **"winding problem"** (see Figure 7.15)[11].

The **Lin-Shu Density Wave Theory**[11] suggests that spiral arms are **quasistatic density waves**. In other words, the spiral arms are not *material arms* consisting of the same set of identifiable stars and gas clouds. Instead, they are regions in the galactic disk with densities greater than average, perhaps 10–20%. Stars and gas clouds pass through these density waves during their orbits like cars moving through a traffic jam.

Stars closest to the center of the galaxy will tend to orbit faster than the density waves, passing forward through them. The increase in density can then cause some gas clouds to collapse and form new stars. Further out, there is a point in the middle of the galaxy where the orbital periods of stars and gas are equal to that of the density wave. Further out still, the orbits of the stars and gas will become slower and thus they will pass through the density waves in the opposite direction (see Figure 7.16).

How do these density waves form? The density wave theory eliminates the winding problem, but then introduces the new question of how these density waves can be formed in the first place, which

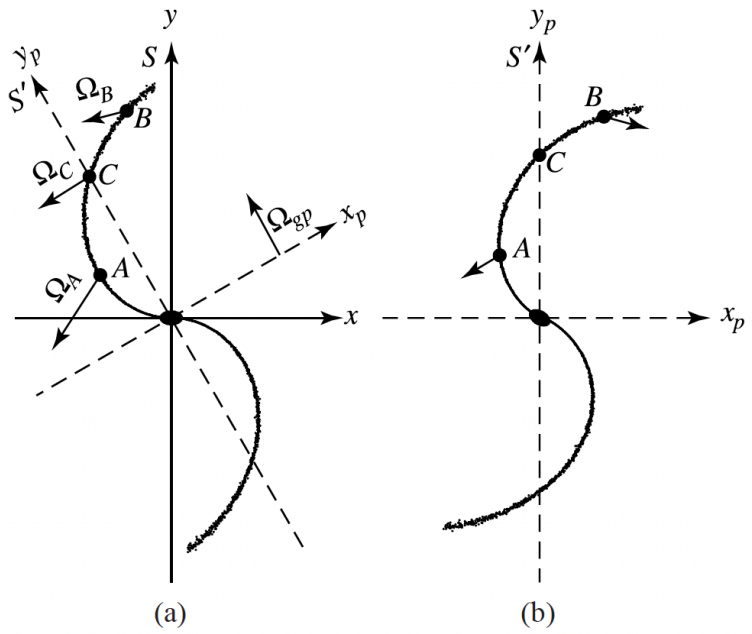


Figure 7.16: Illustration of the spiral arms as density waves in the observed frame S where the density wave has angular speed Ω_{gp}, and the corotating frame S' where the density wave is static[11].

has been the subject of extensive research since this theory was first proposed by Lin and Shu in the 1960s. We can consider a simplified picture of how this can arise by looking at the orbits of stars in a disk-like galaxy. The stars can be thought of as orbiting in an effective potential Φ_{eff}:

$$\Phi_{\text{eff}}(r,z) \equiv \Phi(r,z) + \frac{J_z^2}{2r^2} \tag{7.74}$$

using cylindrical coordinates, where Φ is the gravitational potential and J_z is the angular momentum in the z-direction per unit mass. Then the equations of motion are

$$\ddot{r} = -\frac{\partial \Phi_{\text{eff}}}{\partial r} \quad , \quad \ddot{z} = -\frac{\partial \Phi_{\text{eff}}}{\partial z} \tag{7.75}$$

The most simple solution is circular motion inside the midplane of the galaxy, where the Φ_{eff} derivatives are both 0 (we'll call this potential $\Phi_{\text{eff,m}}$). To find a more general solution, we can consider perturbations around this solution using $\rho = r - r_m$ where r_m is the radius of the minimum energy orbit. Then, Taylor expanding Φ_{eff}:

$$\Phi_{\text{eff}}(r,z) = \Phi_{\text{eff,m}} + \frac{\partial \Phi_{\text{eff}}}{\partial r}\bigg|_m \rho + \frac{\partial \Phi_{\text{eff}}}{\partial z}\bigg|_m z + \frac{1}{2}\frac{\partial^2 \Phi_{\text{eff}}}{\partial r \partial z}\bigg|_m \rho z + \frac{1}{2}\frac{\partial^2 \Phi_{\text{eff}}}{\partial r^2}\bigg|_m \rho^2 + \frac{1}{2}\frac{\partial^2 \Phi_{\text{eff}}}{\partial z^2}\bigg|_m z^2 + \ldots \tag{7.76}$$

Because of how we chose the minimum orbit, the first order derivatives are all 0. Then if we define

$$\kappa^2 \equiv \frac{\partial^2 \Phi_{\text{eff}}}{\partial r^2}\bigg|_m \quad , \quad \nu^2 \equiv \frac{\partial^2 \Phi_{\text{eff}}}{\partial z^2}\bigg|_m \tag{7.77}$$

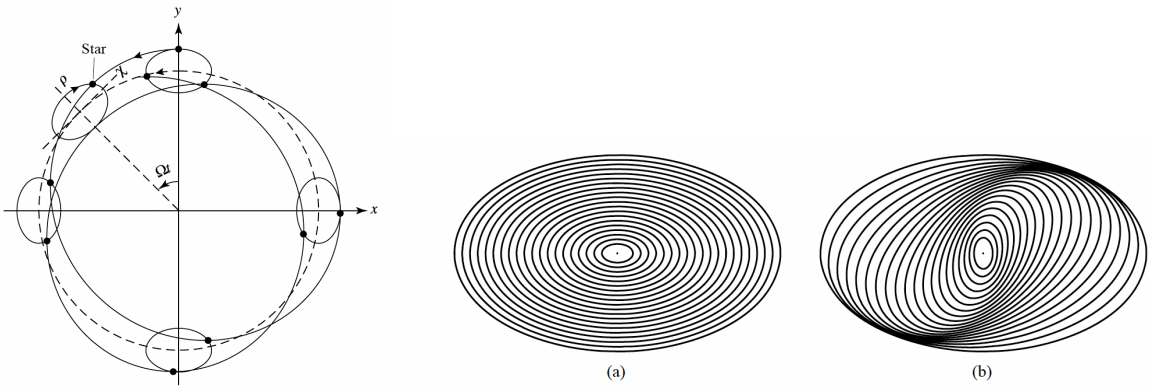


Figure 7.17: *Left*: A schematic of an orbit with epicycles. *Right*: The orbits of many stars viewed in a frame that rotates at an angular speed with $n = 1$ and $m = 2$ ($\Omega' = \Omega - \kappa/2$). The residual epicycle motions are shown in different configurations where (a) they have the same orientation, and (b) they are rotated relative to each other[11].

We have

$$\Phi_{\text{eff}} \approx \Phi_{\text{eff,m}} + \frac{1}{2}\kappa^2\rho^2 + \frac{1}{2}\nu^2 z^2 \tag{7.78}$$

which results in

$$\ddot{\rho} \approx -\kappa^2\rho \quad , \quad \ddot{z} \approx -\nu^2 z \tag{7.79}$$

In other words, the most general solution allows for simple harmonic motion in the r and z direction around the circular orbit, leading to **epicycles** (see Figure 7.17, left). κ is called the epicycle frequency and ν is the vertical oscillation frequency.

If we view these orbits in a frame that corotates with the orbital frequency Ω, then the stars will just be performing their epicycles around a stationary point, and the orbits will be closed. However, we can still achieve closed orbits in a frame where the star completes n orbits for every m epicycle oscillations (where n and m are integers). In this case, the angular frequency of the frame is given by

$$\Omega' = \Omega - \frac{n}{m}\kappa \tag{7.80}$$

and the orbits in this frame will look like ellipses or rose petal patterns depending on the values of n and m chosen. A particularly enlightening choice is $n = 1$ and $m = 2$, which shows residual epicycle motions that increase in density in such a way that they can create either bar-like structures or spiral arm-like structures depending on how the orbits are rotated relative to each other (see Figure 7.17, right). Resonances can also develop between Ω and Ω' that allow stars to tend into these patterns.

115. How many globular clusters does our Galaxy contain? How are they distributed in space? What fraction of the total mass of the Galaxy do they contain?

The Milky Way is known to contain at least **150 globular clusters** with distances ranging from **500 pc** to **120 kpc**. They are distributed nearly spherically around the galactic center. The vast majority of these have been found **within 42 kpc**, with only 6/150 having further distances between 69–123 kpc. Older metal poor clusters with $[\mathrm{Fe/H}] < -0.8$ exist in an extended spherical halo, while younger metal rich clusters with $[\mathrm{Fe/H}] > -0.8$ have a much flatter distribution and may even be associated with the thick disk. The total estimated mass of all the globular clusters is $\sim 10^7\ M_\odot$, which is a negligible amount of the total galaxy mass of $\sim 10^{12}\ M_\odot$ (a fraction of 10^{-5} or 0.001%)[11].

116. A globular cluster at a distance of 10 kpc, containing 10^6 stars, subtends an angle of 1/3 arcminute. Estimate the velocity dispersion among its stars.

Note: This question is approached from the perspective of not having a calculator, so all calculations should be able to be done quickly in your head.

We can use the virial theorem to estimate the velocity dispersion:

$$\sigma \sim \sqrt{\frac{GM}{R}} \tag{7.81}$$

The size R can be found using the distance and angular size:

$$R = \theta d = (1/3 \times 60)'' \times 10\,\text{kpc} = 200,000\,\text{AU} \tag{7.82}$$

$$200,000\,\text{AU} \approx 1\,\text{pc} \approx 3 \times 10^{18}\,\text{cm} \tag{7.83}$$

In the first line we used the definition of a parsec:

$$\theta\,[''] \times \text{d}\,[\text{pc}] = \text{D}\,[\text{AU}] \tag{7.84}$$

And in the second line we recalled the conversion:

$$206,265\,\text{AU} = 1\,\text{pc} \tag{7.85}$$

The mass M can be approximated using the number of stars and average mass of a star $\langle m \rangle \sim 0.5\,M_\odot$:

$$M \approx N\langle m \rangle = 10^6 \times 0.5\,M_\odot \approx 0.5 \times (2 \times 10^{33}\,\text{g}) \times 10^6 \approx 10^{39}\,\text{g} \tag{7.86}$$

Plugging in everything in CGS units (and approximating G):

$$\sigma \approx \sqrt{\frac{(6 \times 10^{-8}) \times 10^{39}}{3 \times 10^{18}}} \approx \sqrt{2} \times 10^{(31-18)/2}\,\text{cm}\,\text{s}^{-1} \approx \tag{7.87}$$

$$\approx 1.5 \times 10^{6.5}\,\text{cm}\,\text{s}^{-1} \approx 1.5 \times 3 \times 10^6\,\text{cm}\,\text{s}^{-1} \tag{7.88}$$

Thus

$$\boxed{\sigma \approx 45\,\text{km}\,\text{s}^{-1}} \tag{7.89}$$

117. What are the Oort A and B coefficients and what basic information about the Galaxy can be determined from them? What are the four types of observations that go into constraining these coefficients, and how precisely are each measured by current technology.

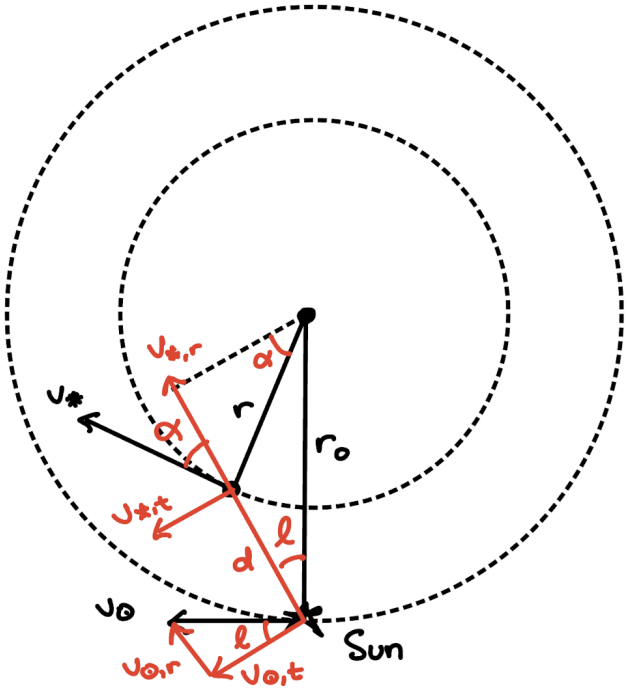


Figure 7.18: A map of the galaxy showing the relevant geometry for showing the relative velocities of a star of the Sun

The **Oort constants**[11] are a set of 2 parameters (A and B) that describe the differential galactic rotation in the solar neighborhood. They were defined in a specific way by Oort such that they could be constrained and studied by making observations of the velocities, distances, and galactic longitudes of nearby stars in the solar neighborhood. Thus, the Oort constants provide a direct relationship between these three quantities. We can derive this relationship, and the functional form of the Oort constants themselves, by using the geometry in Figure 7.18. The velocity components of the star relative to the Sun are

$$v_r = v_{*,r} - v_{\odot,r} = v_* \cos\alpha - v_\odot \sin\ell \tag{7.90}$$

$$v_t = v_{*,t} - v_{\odot,t} = v_* \sin\alpha - v_\odot \cos\ell \tag{7.91}$$

We can also note a few congruencies:

$$r\cos\alpha = r_0 \sin\ell \tag{7.92}$$

$$r\sin\alpha = r_0 \cos\ell - d \tag{7.93}$$

Then, rewriting v_r and v_t in terms of the angular velocities $\Omega = v_*/r$ and $\Omega_0 = v_\odot/r_0$:

$$v_r = v_* \frac{r_0}{r} \sin\ell - v_\odot \sin\ell = (\Omega - \Omega_0) r_0 \sin\ell \tag{7.94}$$

$$v_t = v_* \frac{r_0}{r} \cos\ell - v_\odot \cos\ell - v_* \frac{d}{r} = (\Omega - \Omega_0) r_0 \cos\ell - \Omega d \tag{7.95}$$

Now if we assume $\Omega(r)$ is a smoothly varying function over r, we can do a Taylor expansion about r_0:

$$\Omega(r) = \Omega_0(r_0) + \left.\frac{\mathrm{d}\Omega}{\mathrm{d}r}\right|_{r_0} (r - r_0) + \ldots \tag{7.96}$$

Thus we can get an approximate value for $\Omega - \Omega_0$ by taking the first order term:

$$\Omega - \Omega_0 \simeq \left.\frac{\mathrm{d}\Omega}{\mathrm{d}r}\right|_{r_0} (r - r_0) \tag{7.97}$$

We can also get an approximate relationship between d and $r_0 - r$ from the figure by assuming that $d \ll r_0$, i.e. we restrict ourselves to only the solar neighborhood:

$$r_0 - r \simeq d\cos\ell \tag{7.98}$$

We also recall some trig identities:

$$\begin{aligned} 2\sin\theta\cos\theta &= \sin 2\theta \\ 2\cos^2\theta - 1 &= \cos 2\theta \end{aligned} \tag{7.99}$$

Using (7.97), (7.98), and (7.99), we have:

$$v_r \simeq -\left.\frac{\mathrm{d}\Omega}{\mathrm{d}r}\right|_{r_0} r_0 d \sin\ell\cos\ell = -\frac{1}{2}\left.\frac{\mathrm{d}\Omega}{\mathrm{d}r}\right|_{r_0} r_0 d \sin 2\ell \tag{7.100}$$

$$v_t \simeq -\left.\frac{\mathrm{d}\Omega}{\mathrm{d}r}\right|_{r_0} r_0 d \cos^2\ell - \Omega d = -\frac{1}{2}\left.\frac{\mathrm{d}\Omega}{\mathrm{d}r}\right|_{r_0} r_0 d(\cos 2\ell + 1) - \Omega d \tag{7.101}$$

Finally, we can define the Oort constants A and B:

$$\boxed{A = -\frac{1}{2} r_0 \left.\frac{\mathrm{d}\Omega}{\mathrm{d}r}\right|_{r_0}} \tag{7.102}$$

$$\boxed{B = -\frac{1}{2} r_0 \left.\frac{\mathrm{d}\Omega}{\mathrm{d}r}\right|_{r_0} - \Omega} \tag{7.103}$$

We can also write these using the linear velocities instead of the angular velocities:

$$\left.\frac{\mathrm{d}\Omega}{\mathrm{d}r}\right|_{r_0} = \frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{v}{r}\right)_{r_0} = \frac{1}{r_0}\left.\frac{\mathrm{d}v}{\mathrm{d}r}\right|_{r_0} - \frac{v_0}{r_0^2} \tag{7.104}$$

Such that:

$$\boxed{A = -\frac{1}{2}\left(\left.\frac{\mathrm{d}v}{\mathrm{d}r}\right|_{r_0} - \frac{v_0}{r_0}\right)} \tag{7.105}$$

$$\boxed{B = -\frac{1}{2}\left(\left.\frac{\mathrm{d}v}{\mathrm{d}r}\right|_{r_0} + \frac{v_0}{r_0}\right)} \tag{7.106}$$

Then we can write the velocities in terms of A and B:

$$v_r \simeq Ad \sin 2\ell \tag{7.107}$$

$$v_t \simeq Ad \cos 2\ell + Bd \tag{7.108}$$

Therefore we see that A describes the velocity shear ($\mathrm{d}\Omega/\mathrm{d}r$) whereas B describes the overall galactic rotation (Ω), and we have a relationship between these parameters and the observable quantities v_r, v_t, d, and ℓ. We can use a combination of these parameters to get more physical quantities:

$$\Omega_0 = A - B \quad , \quad \left.\frac{\mathrm{d}v}{\mathrm{d}r}\right|_{r_0} = -(A+B) \tag{7.109}$$

And we can constrain these parameters using (7.107) and (7.108) with measurements of

1. Radial velocities (v_r) — from stellar spectra
2. Tangential velocities (v_t) — from proper motions (i.e. from *Gaia*)
3. Distances (d) — from parallaxes (also from *Gaia*)
4. Galactic longitudes (ℓ) — from where they are in the sky

Current best measurements of the Oort A and B constants are from the *Gaia* DR2 mission[129]:

$$A = 15.1 \pm 0.1\,\mathrm{km\,s^{-1}\,kpc^{-1}} \quad , \quad B = -13.4 \pm 0.1\,\mathrm{km\,s^{-1}\,kpc^{-1}} \tag{7.110}$$

118. Define the following simple "laws" and "profiles" and for what galaxy component or galaxy type each is most relevant: Sérsic, de Vaucouleurs, exponential, Einasto, NFW. Describe where observational data deviates from these idealized profiles.

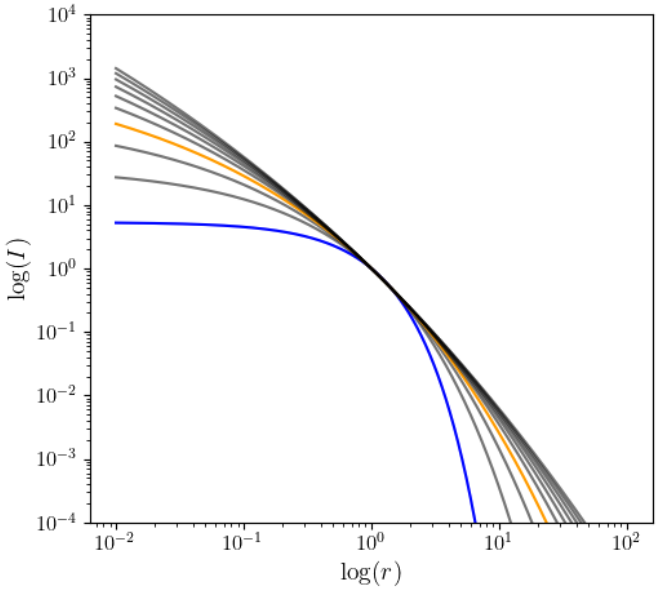


Figure 7.19: The Sérsic profile for different values of n ranging from 1 (bottom) to 10 (top). The blue line corresponds to an exponential profile ($n = 1$) and the orangle line corresponds to a de Vaucouleurs profile ($n = 4$).

Sérsic[126]:

$$\boxed{I(r) = I_e \exp\{-b_n[(r/r_e)^{1/n} - 1]\}} \tag{7.111}$$

where r_e is the effective radius, I_e is the surface brightness at the effective radius, and n is the **Sérsic index**. The exact value of b_n is determined such that half of the total luminosity is contained within r_e (assuming spherical symmetry), but an approximate formula is $b_n \simeq 2n - 0.324$. This profile is quite flexible and can fit the surface brightness profiles of different galaxy components depending on the value of the Sérsic index. For elliptical galaxies or bulges, $n \simeq 2$–4, while for disks, $n \simeq 1$. Since this is fitting the surface brightness, the profile only traces luminous matter (not dark matter). A plot of the Sérsic profile for different values of n is shown in Figure 7.19.

de Vaucouleurs[126]:

$$\boxed{I(r) = I_e \exp\{-7.669[(r/r_e)^{1/4} - 1]\}} \tag{7.112}$$

This is a special case of the Sérsic profile for $n = 4$. Thus, it is best at fitting elliptical galaxies and bulges.

Exponential[126]:

$$\boxed{I(r) = I_d \exp(-r/r_d)} \tag{7.113}$$

This is another special case of the Sérsic profile for $n = 1$, but has been redefined in terms of an exponential scale length r_d and intensity at the scale length I_d. Thus, it is best at fitting disks. Spiral arms may cause bumps in the brightness profile at certain radii that are not fit well by this profile.

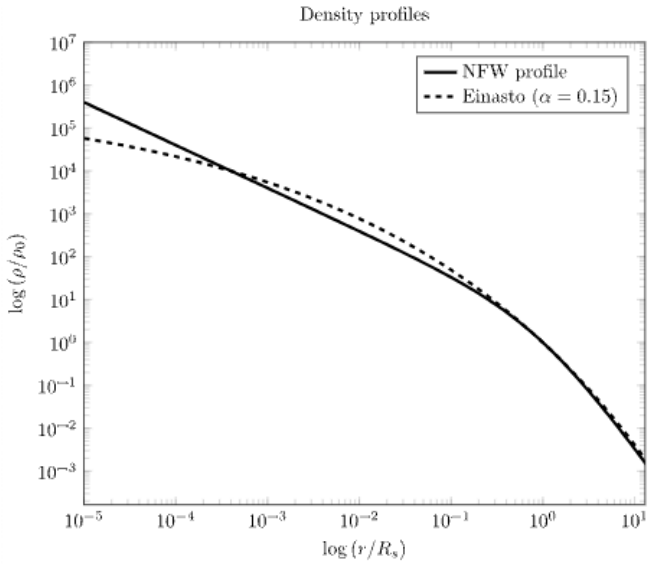


Figure 7.20: A comparison between the Einasto (dashed) and NFW (solid) density profiles. Notice how the NFW profile is more cuspy for low r.

Einasto[107]:

$$\boxed{\rho(r) = \rho_{-2} \exp\left\{\frac{-2}{\alpha}\left[\left(\frac{r}{r_{-2}}\right)^{\alpha} - 1\right]\right\}} \tag{7.114}$$

where r_{-2} is the radius where the logarithmic slope is -2, and $\rho_{-2} = \rho(r_{-2})$. The best-fit power law index is typically $0.12 \lesssim \alpha \lesssim 0.25$. This profile fits the density profiles of dark matter halos. It does not have a central r^{-1} cusp like the NFW profile does—it becomes shallower as $r \to 0$. In general it seems to fit dark matter halos better than the NFW profile. See a comparison of these two in Figure 7.20.

Navarro, Frenk, & White (NFW)[107]:

$$\boxed{\rho(r) = \rho_{\text{crit}} \frac{\delta_{\text{char}}}{(r/r_s)(1 + r/r_s)^2}} \tag{7.115}$$

where r_s is a scale radius, and δ_{char} is a characteristic overdensity. This profile also fits the dark matter halo density. It has a logarithmic slope of -1 as $r \to 0$ and gradually changes to -3 as $r \to \infty$. It resembles an isothermal sphere with a slope of -2 at $r \sim r_s$. Has a divergent central density (core-cusp problem; see §7.Q125).

119. What is "Press-Schechter theory" and why is it wrong?

First, let's introduce some basic concepts. In the early universe (before structure formation), the density at any given point can be described relative to the average density $\bar{\rho}$ using an overdensity field:

$$\delta(\mathbf{x}) = \frac{\rho(\mathbf{x}) - \bar{\rho}}{\bar{\rho}} \tag{7.116}$$

We can smooth this overdensity field over some length scale R, i.e. by using a window function W such that

$$\delta_s(\mathbf{x}; R) \equiv \int \mathrm{d}^3\mathbf{x}'\, \delta(\mathbf{x}') W(\mathbf{x} + \mathbf{x}'; R) \tag{7.117}$$

The length scale R then corresponds to a mass scale $M \propto \bar{\rho} R^3$, with a prefactor depending on the shape of the window function.

Press-Schechter theory[107] states that the probability of a dark matter halo having a mass $> M$ is equal to the probability of an overdensity in the early universe (smoothed over a length scale R) being above some critical density, $\delta_s > \delta_{\rm crit}$ (where $\delta_{\rm crit} \approx 1.68$). In other words,

$$P(> M) = P(\delta_s > \delta_{\rm crit} | M) \tag{7.118}$$

If we assume $\delta_s(\mathbf{x})$ can be described by a gaussian random field, then the probability of having a smoothed overdensity δ_s is given by a normal distribution

$$P(\delta_s) = \frac{1}{\sqrt{2\pi}\sigma(M)} \exp\left(- \frac{\delta_s^2}{2\sigma^2(M)} \right) \tag{7.119}$$

And the probability of having any overdensity above $\delta_{\rm crit}$ is found by integrating up to infinity:

$$P(\delta_s > \delta_{\rm crit} | M) = \frac{1}{\sqrt{2\pi}\sigma(M)} \int_{\delta_{\rm crit}}^{\infty} \mathrm{d}\delta_s \exp\left(- \frac{\delta_s^2}{2\sigma^2(M)} \right) = \frac{1}{2} \mathrm{erfc}\left(\frac{\delta_{\rm crit}}{\sqrt{2}\sigma(M)} \right) \tag{7.120}$$

Then, according to the Press-Schechter formalism, this is equal to $P(> M)$. However, there is a problem. As $M \to 0$, $\sigma(M) \to \infty$ (see (7.127)), meaning $P(> M) \to 1/2$. **Evidently, only half the mass of the universe is contained within collapsed halos.** This is obviously ludicrous, with the reason being that we only allowed regions which are initially overdense to collapse and form halos, but there may be underdense subregions within larger overdense regions that are smaller than our smoothing scale R that would inevitably collapse as a part of the larger overdense region (see Figure 7.21). Thus, Press & Schechter[130] included a "fudge factor" of 2 to account for these initially underdense regions that become part of collapsed halos.

We can now construct the halo mass function $n(M)\,\mathrm{d}M$, i.e. the number density of halos with masses between M and $M + \mathrm{d}M$. We can start with $\partial P(> M)/\partial M \times \mathrm{d}M$ which gives the fraction of the universe in halos with masses between M and $M + \mathrm{d}M$. Then, multiplying by $\bar{\rho}$ makes it the total mass per volume contained within $(M, M + \mathrm{d}M)$, and dividing by M gives the number density per volume within $(M, M + \mathrm{d}M)$:

$$n(M)\mathrm{d}M = \frac{\bar{\rho}}{M} \frac{\partial P(> M)}{\partial M} \mathrm{d}M \tag{7.121}$$

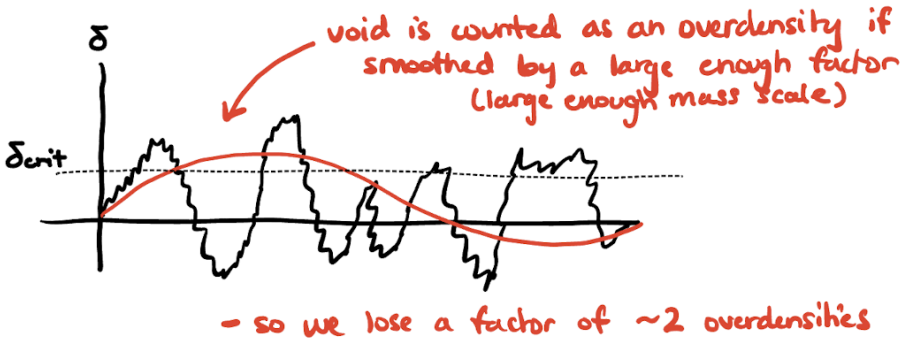


Figure 7.21: An illustration of why Press-Schechter theory undercounts the number of halos by a factor of 2.

Included the fudge factor of 2 and evaluating the derivative:

$$n(M)\mathrm{d}M = 2\frac{\bar{\rho}}{M}\frac{\partial P(\delta_s > \delta_{\rm crit}|M)}{\partial\sigma}\left|\frac{\mathrm{d}\sigma}{\mathrm{d}M}\right|\mathrm{d}M \tag{7.122}$$

$$= 2\frac{\bar{\rho}}{M}\frac{\delta_{\rm crit}}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\delta_{\rm crit}^2}{2\sigma^2}\right)\left|\frac{\mathrm{d}\sigma}{\mathrm{d}M}\right|\mathrm{d}M \tag{7.123}$$

$$= \sqrt{\frac{2}{\pi}}\frac{\bar{\rho}}{M^2}\frac{\delta_{\rm crit}}{\sigma}\exp\left(-\frac{\delta_{\rm crit}^2}{2\sigma^2}\right)\left|\frac{\mathrm{d}\ln\sigma}{\mathrm{d}\ln M}\right|\mathrm{d}M \tag{7.124}$$

We can also write this out explicitly in terms of the mass[119] by realizing that the mass variance is given by integrating the power spectrum $P(k)$ using the Fourier-transformed window function $\tilde{W}(\mathbf{k};R)$

$$\sigma^2(M) = \frac{1}{2\pi^2}\int_0^\infty \mathrm{d}k\, P(k)\tilde{W}(\mathbf{k};R)k^2 \tag{7.125}$$

If we assume a power-law power spectrum $P(k) \propto k^n$ and a gaussian kernel (though the result is the same for a spherical top-hat kernel), then we have

$$\sigma^2(M) \propto \int \mathrm{d}k\, k^{n+2} \propto k^{n+3} \tag{7.126}$$

Recalling that $M \propto R^3 \propto k^{-3}$,

$$\boxed{\sigma(M) \propto k^{(n+3)/2} \propto M^{-(n+3)/6}} \tag{7.127}$$

Therefore we can define some mass scale M_* for which $\sigma^2(M_*) = \delta_{\rm crit}^2/2$, which allows us to normalize $\sigma^2(M)$ into:

$$\sigma^2(M) = \frac{1}{2}\delta_{\rm crit}^2\left(\frac{M}{M_*}\right)^{-(n+3)/3} \tag{7.128}$$

Then,

$$\frac{\mathrm{d}\sigma}{\mathrm{d}M} = -\frac{\delta_{\mathrm{crit}}}{\sqrt{2}}\frac{n+3}{6M}\left(\frac{M}{M_*}\right)^{-(n+3)/6} \tag{7.129}$$

And, plugging these back into (7.124), we have

$$\boxed{n(M)\mathrm{d}M = \frac{n+3}{3\sqrt{\pi}}\frac{\bar{\rho}}{M^2}\left(\frac{M}{M_*}\right)^{(n+3)/6}\exp\left[-\left(\frac{M}{M_*}\right)^{(n+3)/3}\right]\mathrm{d}M} \tag{7.130}$$

For $M \ll M_*$ this acts like $\mathrm{d}n/\mathrm{d}M \propto M^{-2}$, and for $M \gg M_*$ $\mathrm{d}n/\mathrm{d}M$ quickly goes to 0.

120. What is "biased galaxy formation"?

This refers to the fact that galaxies are not perfect tracers of mass in the universe, i.e. $\rho_{\text{gal}}(\mathbf{x}) \neq \rho(\mathbf{x})$. This can be probed using the **two-point correlation function** for galaxies $\xi_{\text{gal}}(r)$ (see definition in §7.Q122) and comparing it to the two-point correlation for all matter ξ_{m} (which can be derived the same way as ξ_{gal}). If galaxies were a perfect tracer of matter, these two should be identical. So we can quantify the bias b_{gal} by saying

$$\xi_{\text{gal}}(r) = b_{\text{gal}}^2 \xi_{\text{m}}(r) \tag{7.131}$$

Since ξ is the Fourier transform of the matter power spectrum $P(k)$ (see §7.Q122), we can also write b_{gal} in terms of the power spectrum normalization on $8h^{-1}$ Mpc scales (chosen to follow conventions). Using (7.128), we can define

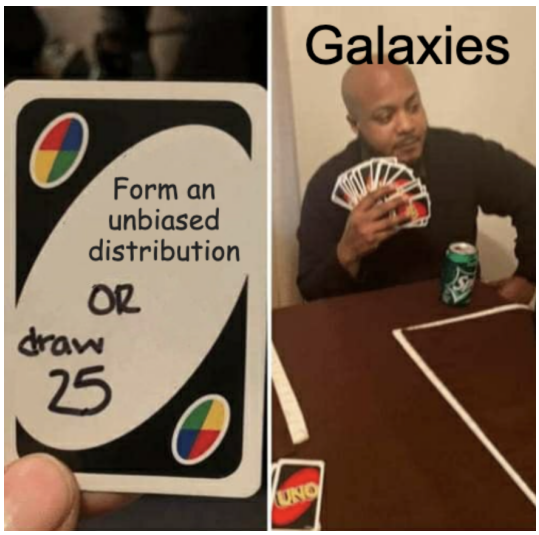
$$\sigma(M) = \sigma_8 \left(\frac{M}{M_8} \right)^{-(n+3)/6} \quad , \quad M_8 = \frac{4\pi}{3} (8h^{-1} \text{ Mpc})^3 \bar{\rho} \tag{7.132}$$

Then, the bias becomes

$$b_{\text{gal}} = \sqrt{\frac{\xi_{\text{gal}}}{\xi_{\text{m}}}} = \frac{\sigma_{8,\text{gal}}}{\sigma_{8,\text{m}}} \tag{7.133}$$

If $b_{\text{gal}} > 1$, galaxies are more tightly clustered than all matter, whereas if $b_{\text{gal}} < 1$, they are less tightly clustered. Measuring observationally from *WMAP*, *Planck*[6], and SDSS, studies have found $\sigma_{8,\text{gal}} \approx 1$ and $\sigma_{8,\text{m}} \approx 0.8$, i.e. $b_{\text{gal}} \approx 1.25$, thus it appears galaxies are more tightly clustered[127,119].

Why are galaxies biased? This is because baryonic matter is subject to different physics than dark matter, and dark matter is the dominant matter component in the universe. It may also be partially because of how galaxy formation works, but this is generally poorly understood.



[6] $\sigma_{8,\text{m}}$ is measured from the gravitational lensing anisotropies in the CMB—the modeled lensing potential gives a matter power spectrum, which can be compared with Ω_m to give $\sigma_{8,\text{m}}$[131]. Then, $\sigma_{8,\text{gal}}$ can just be measured by the observed galaxy 2-point correlation function.

121. What fractions of each of the following is composed of "dark matter": (i) The solar neighborhood, (ii) a typical galaxy like the Milky Way, (iii) a typical galaxy cluster like the Coma cluster, (iv) The universe. How are each of these estimates arrived at?

Note fractions are all quoted relative to the total matter (dark + luminous) in the system.

(i) The solar neighborhood: $\lesssim 15\%$

First, we get an estimate for the local surface density of baryonic matter (stars and gas) in the solar neighborhood Σ_b by measuring a sample of objects within some distance. Then, we can get an estimate for the local total matter density (dark + baryonic) by measuring the vertical velocities of local stars (i.e. motion perpendicular to the disk), since the vertical component of the gravitational force $K_z \approx 2\pi G\Sigma(z)$. This relation is derived from Poisson's equation in cylindrical coordinates if we assume the gravitational field $\mathbf{K}$ is axisymmetric (i.e. $\partial\mathbf{K}/\partial\phi = 0$):

$$\nabla \cdot \mathbf{K} = -4\pi G\rho \tag{7.134}$$

$$\frac{\partial K_z}{\partial z} - \frac{1}{r}\frac{\partial}{\partial r}(rK_r) = \frac{\partial K_z}{\partial z} - \frac{1}{r}\frac{\partial}{\partial r}(v_c^2) = -4\pi G\rho \tag{7.135}$$

$$\frac{\partial K_z}{\partial z} \approx -4\pi G\rho \tag{7.136}$$

$$\int_{-h}^{h} \mathrm{d}z\, \frac{\partial K_z}{\partial z} \approx -4\pi G \int_{-h}^{h} \mathrm{d}z\, \rho \tag{7.137}$$

$$|K_z| \approx 2\pi G\Sigma \tag{7.138}$$

In the third line we assumed v_c does not change with radius (flat rotation curve). It turns out this is not correct, but it's a good first-order approximation. Thus if we can determine the vertical component of the force $K_z(r,z) \sim GM(<r)z/r^3$, we can determine $\Sigma = |K_z|/2\pi G$. Once we have Σ_b and Σ, we can convert to ρ_b and ρ, and from there it's a simple matter to get the dark matter fraction:

$$f_{\mathrm{DM}} = \frac{\rho - \rho_b}{\rho} \tag{7.139}$$

Recent studies have found values in the solar neighborhood of $\rho_b \approx 0.084\,M_\odot\,\mathrm{pc}^{-3}$ and $\rho \approx 0.097\,M_\odot\,\mathrm{pc}^{-3}$, giving $f_{\mathrm{DM}} \simeq 13\%$[132].

(ii) A typical galaxy like the Milky Way: $\sim$ 90–95%

Firstly, we can use the rotation curve of a spiral galaxy to get an estimate of the total (dark + luminous) mass of the galaxy. Or, for an elliptical galaxy, we can use the virial theorem. See §7.Q101 for details. Then, we can separately get an estimate of the stellar mass in a couple of ways:

- Measure the luminosity of the galaxy L and adopt a mass-to-light ratio $\Gamma = M/L$ to convert this to a stellar mass. If you don't have a good estimate for the mass-to-light ratio, see the other 2 methods.
- Get an optical spectrum of the galaxy and fit it with **stellar population synthesis** models to obtain a total stellar mass. You get a collection of spectra for single stars of all different types (age, metallicity, mass) and convolve these with an assumed initial mass function (i.e. Salpeter) assuming some fixed age & metallicity—this is called a "simple stellar population" (SSP). Then

convolve the SSPs with an assumed star formation history (single/double exponential, discrete bursts, periodic, etc.) to get the model for the total starlight in the galaxy. This can also be done with broadband spectroscopic data if you fit the entire SED and impose some constraints (i.e. the infrared luminosity emitted by dust is reprocessed starlight, so it must be balanced with the optical luminosity directly from the stars).

- Use scaling relations for the stellar mass based on other properties, like colors in specific filters, that have been calibrated by doing the above procedure on lots of galaxies.

  Side note: About 10–15% of the *visible* mass of the Milky Way is from the ISM, not the stars. Obviously this is not accounted for when fitting stellar population synthesis models, but the gas and dust masses can be estimated from the strengths of emission lines (which depend on the column density).

In any case, once we have both the total and stellar masses, we can get the fraction of dark matter just like in (i). For the Milky Way, $M_{\text{tot}} \sim 3 \times 10^{12}\, M_\odot$ and $M_* \sim 10^{11}$, so $f_{\text{DM}} \simeq 95\%$.

(iii) A typical galaxy cluster like the Coma cluster: $\sim$ 80–95%[120,133]

For clusters, the baryonic mass is dominated by the hot ICM (which accounts for up to 15% of the *total* mass), so we need a different approach to measure the baryonic matter. One method that gets us measures of both ρ_b and M_{tot} in one fell swoop is using the X-ray emission from thermal bremsstrahlung. See §7.Q101 for a detailed explanation of how this works. Essentially, we are able to relate the X-ray emission to the thermodynamic properties of the gas, and then we can relate the gas properties to the total mass if we assume hydrostatic equilibrium, see (7.17).

The total mass could also be obtained from weak lensing or scaling relations, but either way we will need an estimate of the baryonic gas density. For the Coma cluster, $M_b \sim 1 \times 10^{14}\, M_\odot$ and $M_{\text{tot}} \sim 1.5 \times 10^{15}\, M_\odot$, corresponding to $f_{\text{DM}} \simeq 93\%$[134].

(iv) The universe: $\sim$ 85% (or $\sim$ 23% of the total energy density)

The dark matter content of the universe depends on the cosmological parameters Ω_m and Ω_b, which are constrained by the CMB power spectrum, BAO, and BBN. See questions §8.Q129, §8.Q144, and §8.Q146 for details. Current values are $\Omega_b \simeq 0.04$ and $\Omega_m \simeq 0.27$, which gives $f_{\text{DM}} = (\Omega_m - \Omega_b)/\Omega_m \simeq 85\%$.

122. What is the galaxy correlation function? How do the correlation functions of galaxies and clusters of galaxies differ? How is the galaxy correlation function used to constrain H_0?

We can ask the question: "What is the probability of finding 2 galaxies within some volumes ΔV_1, ΔV_2, separated by a distance $\mathbf{r}$?"[119,107] The probability of finding any galaxy within a volume element ΔV at position $\mathbf{x}$ is

$$P_1(\mathbf{x}) = n(\mathbf{x})\Delta V = \bar{n}\frac{\rho(\mathbf{x})}{\bar{\rho}}\Delta V \tag{7.140}$$

Then, for the two galaxies, we multiply the probabilities

$$P_{12}(\mathbf{x},\mathbf{r}) = \frac{\bar{n}^2}{\bar{\rho}^2}\rho(\mathbf{x})\rho(\mathbf{x}+\mathbf{r})\Delta V_1\Delta V_2 \tag{7.141}$$

Then, if we average over all $\mathbf{x}$ to get the average probability in all space, we have

$$P_{12}(\mathbf{r}) = \frac{\bar{n}^2}{\bar{\rho}^2}\langle\rho(\mathbf{x})\rho(\mathbf{x}+\mathbf{r})\rangle\Delta V_1\Delta V_2 \tag{7.142}$$

The probability therefore depends on the quantity

$$\langle\rho(\mathbf{x})\rho(\mathbf{x}+\mathbf{r})\rangle = \bar{\rho}^2\langle(\delta(\mathbf{x})+1)(\delta(\mathbf{x}+\mathbf{r})+1)\rangle \tag{7.143}$$

$$= \bar{\rho}^2\langle\delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{r}) + \delta(\mathbf{x}) + \delta(\mathbf{r}) + 1\rangle \tag{7.144}$$

$$= \bar{\rho}^2(1 + \langle\delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{r})\rangle) \tag{7.145}$$

Whereas, if galaxies were completely randomly distributed, we would expect the quantity in angled brackets to go to 0 such that $\mathrm{d}P_{12}(\mathbf{r}) = \bar{n}^2\mathrm{d}V_1\mathrm{d}V_2$. Therefore, we quantify the non-uniformity in the distribution by defining the **"two-point correlation function"** for galaxies:

$$\boxed{\xi_{\mathrm{gal}}(r) \equiv \langle\delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{r})\rangle} \tag{7.146}$$

If $\xi_{\mathrm{gal}} = 0$, galaxies are completely uncorrelated and their distribution is random. If $\xi_{\mathrm{gal}} > 0$, galaxies prefer to cluster together, and if $\xi_{\mathrm{gal}} < 0$, they prefer to stay apart from each other. Notice that we have also claimed ξ_{gal} is a function of the magnitude r, independent of the direction of $\mathbf{r}$, which is because the universe is homogeneous and isotropic.

Also note that ξ_{gal} is the Fourier transform of the power spectrum $P(k)$ of the perturbation field:

$$\xi_{\mathrm{gal}}(r) = \frac{1}{(2\pi)^3}\int \mathrm{d}^3\mathbf{k}\,P(k)e^{i\mathbf{k}\cdot\mathbf{r}} \tag{7.147}$$

How do we measure this observationally? Say we make a galaxy survey and measure $S(r)$ galaxy pairs with separations between $r \pm \Delta r/2$ out of a total sample size of N_{sam} galaxy pairs (note: for a sample with N_{gal} galaxies, there are $N_{\mathrm{gal}}(N_{\mathrm{gal}}-1)/2$ unique pairs). Then our measured probability is

$$P_{12}(r) = \frac{S(r)}{N_{\mathrm{sam}}} = \bar{n}^2(1+\xi_{\mathrm{gal}})\Delta V_1\Delta V_2 \tag{7.148}$$

Then, we can also created a simulated galaxy sample where we assume the galaxy distribution is completely random and extract the number of pairs $R(r)$ with separations between $r \pm \Delta r/2$, just like before, with a total number of simulated pairs N_{sim}. This gives us a probability

$$P_{12,\text{sim}}(r) = \frac{R(r)}{N_{\text{sim}}} = \bar{n}^2 \Delta V_1 \Delta V_2 \tag{7.149}$$

Then we just divide (7.148) by (7.149) to isolate ξ_{gal}:

$$\boxed{\xi_{\text{gal}}(r) = \frac{N_{\text{sim}}}{N_{\text{sam}}} \frac{S(r)}{R(r)} - 1} \tag{7.150}$$

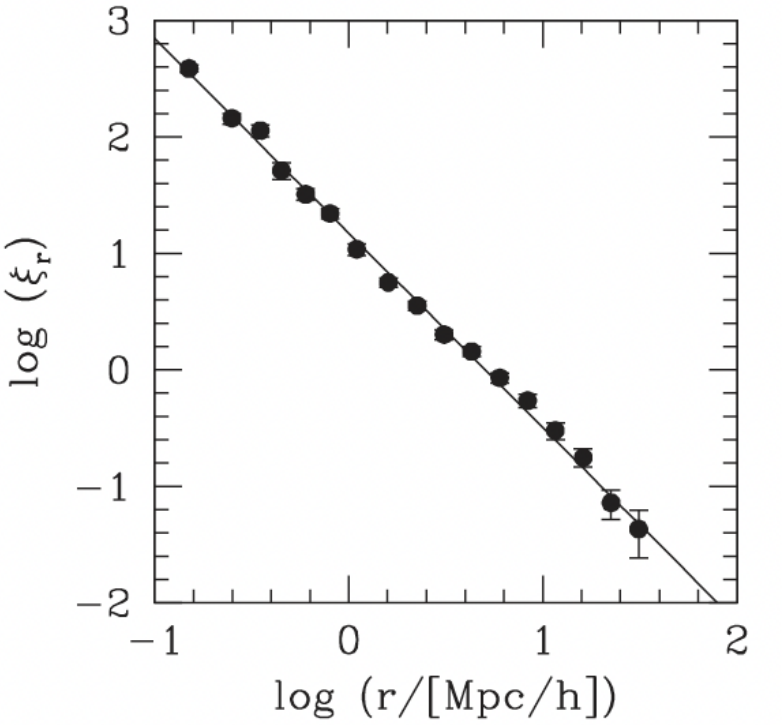


Figure 7.22: A measurement of the 2-point correlation function for galaxies[107].

Statistical studies measuring the correlation function of galaxies have found that a power law approximates the two-point correlation function well:

$$\boxed{\xi_{\text{gal}}(r) \simeq \left(\frac{r}{r_0}\right)^{\gamma}} \tag{7.151}$$

For galaxies, the best-fit values are $\gamma \simeq -1.67$ and $r_0 \simeq 7h_{70}^{-1}$ Mpc (see Figure 7.22). This tells us that galaxies that are closer together are more likely to cluster together (ξ_{gal} is big for small r), while galaxies that are further apart are nearly completely randomly distributed ($\xi_{\text{gal}} \to 0$ as $r \to \infty$).

For galaxy clusters, the best fit values are $\gamma \simeq -2.11$ and $r_0 \simeq 23h_{70}^{-1}$ Mpc (see reference 135). Thus, in general **galaxy clusters are much more correlated than galaxies themselves**, yielding $\xi_{\text{cluster}} \simeq 30\xi_{\text{gal}}$.

Measuring H_0: The Baryon Acoustic Oscillations (BAO) create a visible excess in the correlation function at a scale of ∼150 Mpc. Why? See questions §8.Q129 and §8.Q131 for an explanation of BAO. The important point here is that the oscillations in the baryons cause a slight but noticeable increase in the 2-point correlation function at the scale that corresponds to the first BAO harmonic because the density perturbations in the early universe reach this length scale before collapsing, preferentially forming more galaxies at these separations.

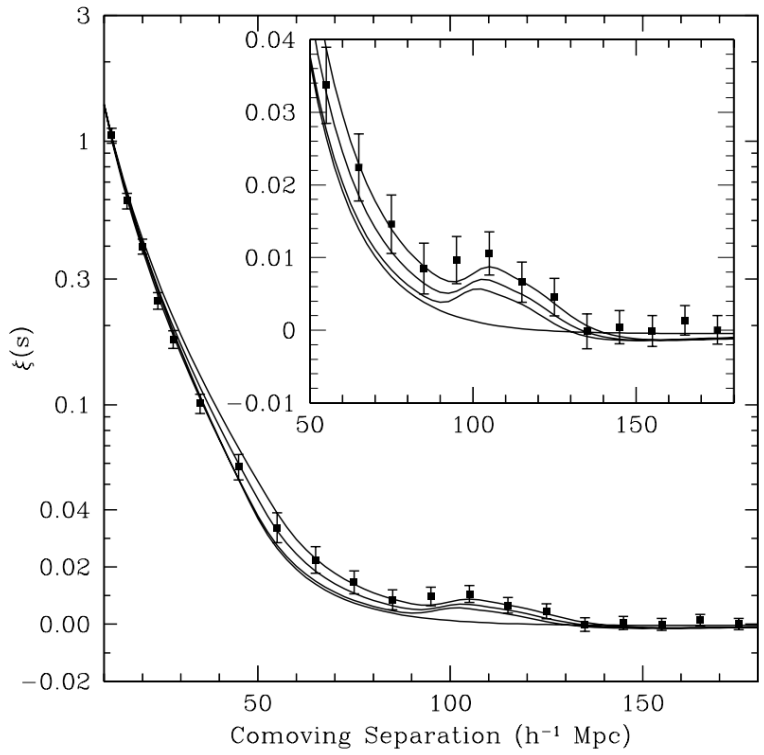


Figure 7.23: Another measurement of the 2-point correlation function of galaxies, this time using SDSS data and showing the small peak at $\sim 110h^{-1}$ Mpc due to the BAO[136].

We can use the BAO as a **"standard ruler"** in cosmology since we know what its intrinsic length scale should be at the time of recombination (analogous to the "standard candles" for measuring distance from objects whose intrinsic luminosities we know). It should be ∼the distance that sound could travel from the Big Bang to the epoch of recombination (i.e. the sound horizon), or $r_s \sim 0.14$ Mpc (see §8.129). We can then extrapolate the intrinsic length scale into a comoving length scale at $z = 0$ of $r_0 = r_s(1 + z_{\text{recomb}}) \sim 150$ Mpc.

Then, we can use the galaxy correlation function in the angular direction to measure the angular size $\Delta\theta$ of the BAO at some local redshift by looking for the small peak in the function. With these, we have a measurement of $d_A(z)$ at a local redshift, which allows us to put a constraint on H_0 assuming that we know the other cosmological parameters (Ω_m, Ω_Λ, etc.) well enough[136]:

$$d_A(z) = \frac{r(z)}{\Delta\theta(z)} \quad \longrightarrow \quad \frac{r_0}{(1+z)\Delta\theta(z)} = \frac{c}{H_0}\frac{1}{1+z}\int_0^z \frac{\mathrm{d}z'}{E(z')} \tag{7.152}$$

We can also apply this in the radial direction by measuring the separation in redshift Δz that the

BAO peak occurs at. Then we can relate this to the comoving distance

$$r_0 = \frac{c}{H_0}\int_z^{z+\Delta z}\frac{\mathrm{d}z'}{E(z')} \quad\longrightarrow\quad r_0 \simeq \frac{c}{H_0}\frac{\Delta z}{E(z)} \tag{7.153}$$

In both cases, we get an estimate of H_0 since all other parameters (r_0, $\Delta\theta$, Δz, z) are known. Note that often our resolution isn't high enough to separate the angular and radial separation components, so in this case people often perform an analysis to obtain a volume-averaged distance

$$d_V(z) = \left[(1+z)d_A(z)\right]^{2/3}\left[\frac{cz}{H(z)}\right]^{1/3} \quad\longrightarrow\quad d_V(z) = \left[\frac{r_0}{\Delta\theta(z)}\right]^{2/3}\left[\frac{r_0 z}{\Delta z}\right]^{1/3} = \frac{r_0}{\Delta\Theta} \tag{7.154}$$

where $\Delta\Theta \equiv \Delta\theta^{2/3}(\Delta z/z)^{1/3}$ is some kind of average dimensionless 3D separation[137,136].

123. What are "active galactic nuclei"? What is the "unified model of AGN", and how does it explain the variety of AGN that are observed? How and when are AGN important for galaxy evolution?

An object is classified as an **Active Galactic Nucleus (AGN)** if, roughly speaking, it fulfills at least one of the following[107]:

1. A compact nuclear region that is much brighter than a region of the same size in a normal galaxy
2. Non-stellar/non-thermal continuum emission
3. Strong emission lines
4. Variability in the conintuum/emission lines on relatively short time scales

There are a variety of different subcategories of AGN that are observed with different properties, often put in classifications based on observations from a particular wavelength regime[11].

- **Seyfert Galaxies (Sy)**: Galaxies that have a nucleus that is optically bright and pointlike, exhibiting a featureless power-law continuum and bright emission lines. These are radio-quiet and may exhibit UV/X-ray excess.
  - **Seyfert I**: Shows broad emission lines with speeds up to 1000–5000 $\mathrm{km\,s^{-1}}$ and narrow emission lines with speeds ~ 500 $\mathrm{km\,s^{-1}}$.
  - **Seyfert II**: Shows only narrow emission lines with speeds ~ 500 $\mathrm{km\,s^{-1}}$ and generally have a weaker continuum (mostly stellar).
  - There are also intermediate types (i.e. **Seyfert 1.5**) that display broad lines only for permitted transitions (like Hα) and not forbidden transitions (like [O III]).
- **Radio Galaxies**: Galaxies that are extremely bright at radio wavelengths (radio-loud). May display radio lobes, with one connected to the central nucleus by a jet. May also exhibit UV/X-ray excess.
  - **Broad Line Radio Galaxies (BLRGs)**: Correspond to Seyfert Is. Have bright pointlike nuclei and faint, hazy envelopes.
  - **Narrow Line Ratio Galaxies (NLRGs)**: Correspond to Seyfert IIs. Found in giant or supergiant elliptical galaxies (i.e. cD galaxies).
  - Also sometimes subdivided into **"steep-spectrum"** and **"flat-spectrum"** radio quasars (SSRQs and FSRQs) based on their radio spectral index α.
- **Quasi-Stellar Radio Sources (Quasars)**: Radio galaxies that have pointlike (or "star-like") optical counterparts with highly redshifted spectra ($z \gtrsim 0.1$) containing broad emission lines. These are some of the most distant objects known in the universe and thus they have some of the highest intrinsic luminosities $\sim 10^{45}$–10^{48} $\mathrm{erg\,s^{-1}}$ ($\sim 10^5\times$ higher than a normal Milky-Way-like galaxy). They emit an excess of UV light compared to stars and are quite blue in appearance, exhibiting a "big blue bump" feature in their SEDs. They also tend to exhibit lots of narrow absorption lines at smaller redshifts than the quasar itself, due to intermediate gas clouds along our line of sight.
- **Quasi-Stellar Objects (QSOs)**: A superset of quasars that do not have to have a corresponding radio counterpart. It has been discovered that $\sim 90\%$ of QSOs are radio quiet. QSOs can also

be broken up into **Type I** and **Type II** based on whether or not they exhibit broad emission lines.

- **Blazars**: Display rapid variability and highly linearly polizarized light in optical wavelengths. All blazars are radio-loud and display a UV/X-ray excess.
  - **BL Lacs**: Characterized by rapid timescales of variability. Luminosities may change up to 30% in 24 hours and may change by a factor of 100 over longer time periods. Have strongly polarized power-law continua (30–40% linear polarization) and very faint emission lines at high redshifts. About 90% reside in elliptical galaxies. Named after **BL Lacertae**, the most well known object in this category.
  - **Optically Violently Variable Quasars (OVVs)**: Similar to BL Lacs, but much more luminous, and may display broad emission lines.

In addition to these categories, radio-loud AGN (radio galaxies, quasars, and blazars) can be further subdivided using the **Faranroff-Riley (FR) Luminosity Classes**[11]:

- **FR-I Objects**: Defined from having the extent of the radio jets being $< 50\%$ of the full extent of the radio source. Generally have diminishing radio luminosities with increasing distance from the center of the jets. Often have 2 recognizable jets, and the jets tend to be curved.
- **FR-II Objects**: Defined from having the extent of the radio jets being $> 50\%$ of the full extent of the radio source. Tend to be the most bright at the ends of the radio lobes. Often only have 1 recognizable jet, and the jet tends to be straight.

A note on terminology: While the above definitions are the "technically correct" usages of the above terms, the way a lot of these terms are actually used today has somewhat morphed. Generally, **"quasar"** refers to both quasars with radio counterparts and QSOs without radio counterparts. Thus, additional modifiers for "radio-loud quasars" and "radio-quiet quasars" are utilized to make this distinction. Additionally, **"QSO"** is sometimes used as an abbreviation for "quasar"[11].

The Unified AGN Model: This model suggests that, despite the observational differences between all of the types of AGN, they can all be explained as the result of the same underlying phenomenon: a **Supermassive Black Hole (SMBH)** at the center of a galaxy that has a surrounding disk of gas and dust called an **accretion disk** from which it is actively accreting material. Around the accretion disk is also a larger dusty torus. The differences in each class of AGN, then, can be explained by differences in viewing angle and accretion rate of the AGN[11].

If we are viewing the AGN from an angle above the disk/torus on the side that the radio jet is, we see a radio-loud BLRG or Type I QSO (which one we see depends on the accretion rate—more powerful accretion fuels QSOs). If our viewing angle is more oblique such that our line of sight passes through the dusty torus, our view of the broad line region (BLR) immediately around the accretion disk is obscured, so we only see the narrow line region (NLR) at larger spatial extents and hence view a NLRG or Type II QSO. And if our viewing angle is directly along the line of sight of the radio jet, we see a blazar—either a BL Lac or an OVV depending on the accretion rate. Finally, if we are viewing from the side of the disk/torus that *doesn't* have the radio jet, then for low accretion rates we'll see either a Seyfert I or Seyfert II (depending on the viewing angle), and for high accretion rates we'll see a radio-quiet QSO of Type I or II (again depending on the viewing angle). See Figure 7.24 for a visual summary of this model.

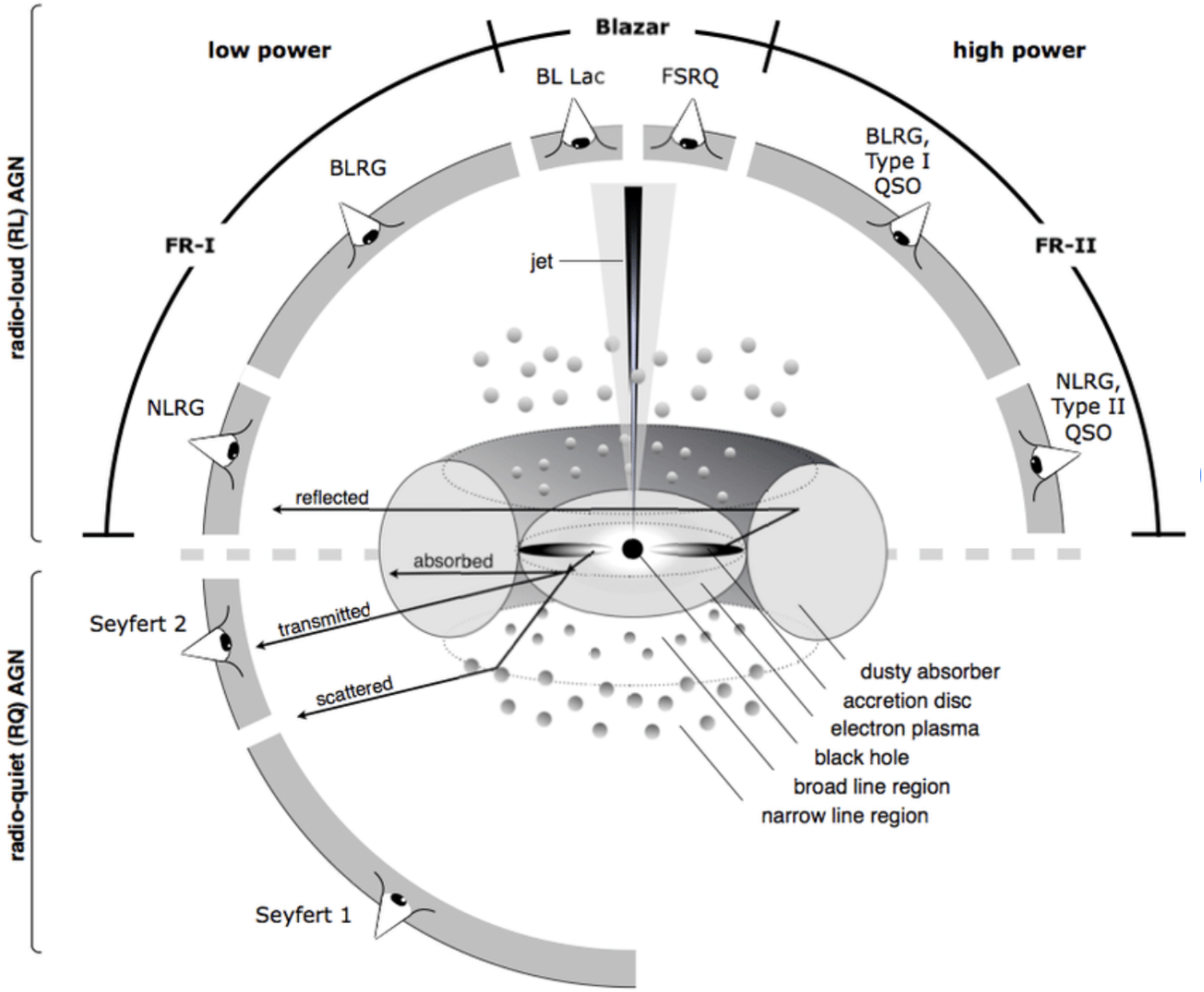


Figure 7.24: A diagram of the AGN unified model, showing what type of AGN we will see depending on the viewing angle and the accretion rate[138]

Therefore, this model simultaneously explains the differences in the X-ray and optical properties of Type I and Type II AGN, the radio properties of radio-loud and radio-quiet AGN, the radio properties of FR-I and FR-II AGN, and the extreme variability and polarization of blazars (since we are looking directly down the jet axis).

The SEDs of AGN are extraordinarily flat over ∼15 orders of magnitude in frequency, such that they are often described by a simple $F_\nu \propto \nu^\alpha$ power law. An example is shown in Figure 7.25. There is structure, however, particularly in the form of bumps in the IR and UV. There are a variety of emission mechanisms involved in creating the full SED, including radio synchrotron emission, thermal emission from the accretion disk (peaking in the UV ⟶ the "big blue bump") and dusty torus (peaking in the IR), and inverse Compton scattering up to the X-rays[107,120].

Importance in galaxy evolution: Recall the difference between the observed luminosity function of galaxies and the model halo mass function from simulations (§7.Q108). AGN feedback is thought

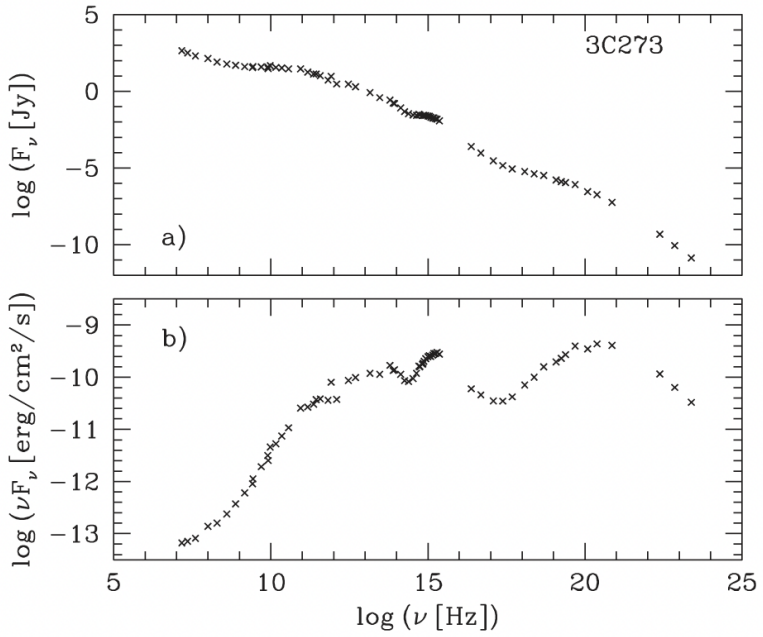


Figure 7.25: An example SED of the quasar 3C 273[107].

to suppress star formation, causing a dip in the observed luminosity function at high luminosities/masses. AGN feedback comes in 2 main forms[107]:

- **Radiative processes**: UV and X-ray photons from the AGN can ionize atoms, and the excess energy of the photon goes into kinetic energy, heating the gas through photoionization heating. This is thought to be a significant contributor to the reionization of the universe. It can heat the surrounding IGM up to $\sim 10^4$ K, suppressing gas cooling and therefore suppressing star formation.

  Radiation pressure on the dusty torus can transfer momentum from the radiation field to the gas (since the gas is coupled to the dust) and, if the force is strong enough, it can cause the gas to flow out (in an "outflow") from the AGN in a momentum-driven wind.

  Photons from the AGN can also heat the gas through Compton scattering (scattering between a photon and an electron), which transfers energy to the electron if the temperature of the gas is lower than the effective Compton temperature of the radiation field.

  Although it's highly uncertain, estimates of the efficiency of these radiative processes gives $\sim 5\%$—that is, 5% of the energy in the radiation field is transferred into the gas. Usage of this efficiency in simulations shows that such an energy injection has a significant impact on the host galaxy, producing gas outflows that quench the SMBH growth and suppress star formation.

- **Mechanical processes**: Radio jets from the AGN can inflate "bubbles" or cavities of relativistic gas. These bubbles then rise buoyantly and expand adiabatically, depositing energy into the surrounding gas through shocks until the expansion velocity becomes lower than the sound speed, after which energy is deposited through $P\mathrm{d}V$ work. These processes are thought to be responsible for quenching cooling flows in galaxy clusters.

  As with the radiative processes, it is thought that mechanical feedback can efficiently quench cooling and suppress star formation.

124. Describe the following relations and what they are useful for: (i) Kennicutt-Schmidt; (ii) M-sigma relation; (iii) the Madau plot; (iv) the star-forming main sequence.

(i) Kennicutt-Schmidt: A scaling relation between the local star formation surface density in a given region, $\Sigma_{\rm SFR}$, and the local gas surface density $\Sigma_{\rm gas}$:

$$\Sigma_{\rm SFR} = 2.5\times10^{-4}\,M_\odot\,{\rm yr}^{-1}\,{\rm kpc}^{-2}\left(\frac{\Sigma_{\rm gas}}{1\,M_\odot\,{\rm pc}^{-2}}\right)^{1.4} \quad \text{or} \quad \boxed{\Sigma_{\rm SFR}\propto\Sigma_{\rm gas}^{1.4}} \tag{7.155}$$

where the power law index $n\approx1.4\pm0.15$ (from Refs. 139,140). Allows us to predict star formation rates given a more easily observable quantity in the gas density. SFRs are hard to measure directly—must either use scaling relations or model spectra with stellar populations using an assumed IMF, age, metallicity, and dust extinction, all of which come with uncertainties.

(ii) $M-\sigma$ relation: A scaling relation between the mass of a black hole and the velocity dispersion in a bulge/elliptical galaxy:

$$M_{\rm BH} = 3\times10^{8}\,M_\odot\left(\frac{\sigma}{200\,{\rm km\,s}^{-1}}\right)^{4.38} \quad \text{or} \quad \boxed{M_{\rm BH}\propto\sigma^{4.38}} \tag{7.156}$$

where the power law index ranges from 3.75–5.1 depending on the sample used and the exact definition of σ. The above value of 4.38 is taken from Ref. 141. This is useful because it is hard to measure black hole masses directly, whereas σ is very easy to measure. It also gives us a hint that galaxies are intrinsically linked to their SMBHs and coevolve, since measurements of σ often come from well outside the sphere of influence of the BH.

(iii) Madau plot: A plot of the cosmic star formation history, i.e. average star formation rate over time/redshift. It has a distinct peak near a redshift of $z\sim2$, called "cosmic noon." See Figure 7.26.

This is around the same time when the number density of quasars also peaks, inferring some connection between AGN and host galaxy activity, like $M-\sigma$. Suggests that many metals were formed early from the deaths of massive stars (this is corroborated by ICM metallicity measurements).

(iv) The star-forming main sequence: A scaling relation between stellar mass M_* and star formation rate SFR that also evolves over time[143]:

$$\log{\rm SFR}(t) = (0.84-0.026t)\log M_* - (6.51-0.11t) \quad \text{or} \quad \boxed{{\rm SFR}\propto M_*^{0.5}}\ (\text{at } t=13) \tag{7.157}$$

(t is the age of the Universe in Gyr). This describes most galaxies quite well, but there are outliers. Particularly "red and dead" ellipticals, which fall below the main sequence, and starbursts, which lie above it. Figure 7.27 shows this relation at a few selected redshifts.

This relation suggests that galaxies spend most of their lifetimes forming stars in a regulated way, and it is useful for understanding galaxy evolution and stellar populations.

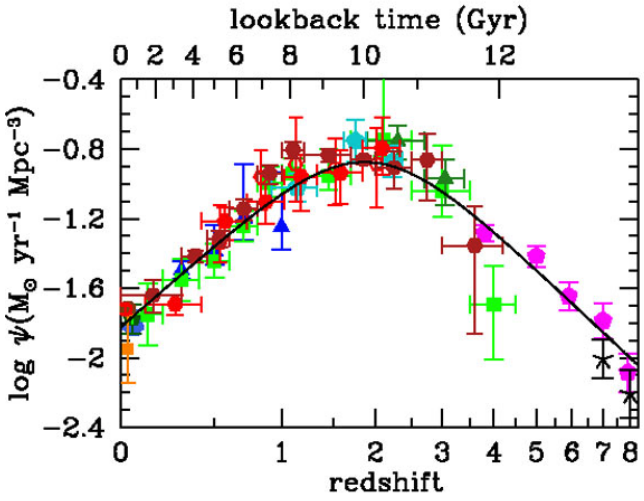


Figure 7.26: The Madau plot of SFR vs. redshift[142].

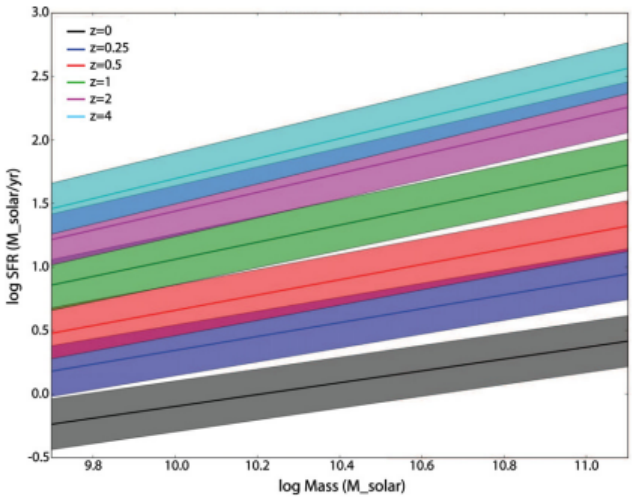


Figure 7.27: The SFR main sequence at a few different redshifts[143].

125. Describe the following galaxy-formation problems: (i) the missing satellite problem; (ii) the core-cusp problem; (iii) the cooling flow problem. Which of these are solved? What are their solutions?

(i) The missing satellite problem: This refers to the discrepancy in the number of subhalos predicted by CDM simulations and the number of satellite galaxies observed around larger galaxies (see Figure 7.28).

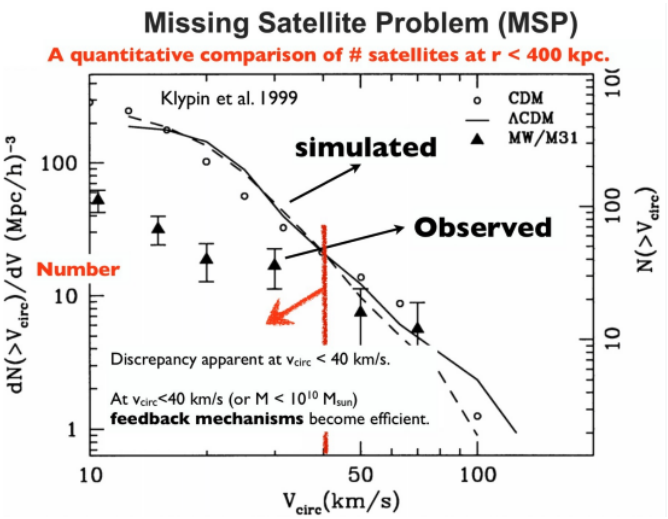


Figure 7.28: Plot showing the difference in observed vs predicted number of satellite galaxies[144].

Potential solutions:

- The thought is that star formation is less efficient in the least massive DM halos, but it is not known how star formation is being suppressed in these halos.
- Recent surveys from SDSS (and in the future LSST) have found lots of extremely faint dwarf galaxies that have somewhat resolved this tension. Nora Shipp (formerly MIT) has done work demonstrating that if one carefully considers the detectability limits of faint satellite dwarf galaxies, this problem is resolved, and in fact there may actually be a "too many satellites" problem.
- Accurate modeling of baryonic physics in these simulations (like FIRE) can also resolve this tension.

(ii) The core-cusp problem: CDM simulations of dark matter halos predict "cuspy" density profiles described by an NFW profile (see §7.Q118) that sharply increases like r^{-1} at small radii. However, observed rotation curves (and thus density profiles) of dwarf galaxies are much more "cored", meaning they have a constant density at small radii (see Figure 7.29)[145].

Potential solutions:

- Baryonic feedback (i.e. from SNe, AGN, or a rotating spiral arm bar) can "puff up" the dark matter distribution and form a core
- Warm or self-interacting DM
- MOND

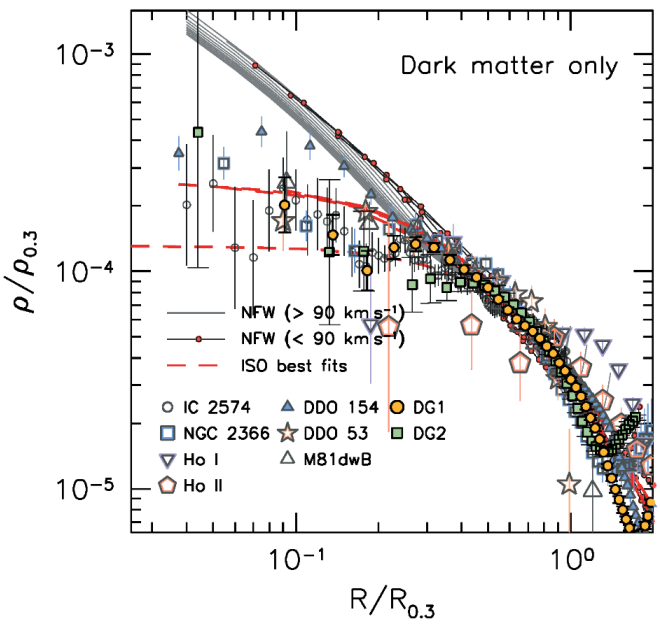


Figure 7.29: Demonstration of the different density profiles of observed galaxies and simulated halos[146]

(iii) The cooling flow problem[7]: The hot intracluster medium (ICM) in galaxy clusters typically has temperatures on the order of 10^7 K and emits X-rays via thermal bremsstrahlung. This radiation causes it to lose energy and cool down over time. If the gas is allowed to cool down enough, it can start radiating energy through other mechanisms, primarily through line emission. How long will it take the gas to cool? We can examine this problem by calculating the *cooling time*, which is just the time it will take to radiate away all of its thermal energy. We learned in §5.Q84 that the cooling function $\Lambda(T)$ describes the cooling rate of a system as a function of its temperature—specifically, $n_H n_e \Lambda(T)$ is the cooling rate per unit volume. $\Lambda(T)$ takes into account cooling from all mechanisms, including thermal bremsstrahlung and line emission. We can compare this to the thermal energy per unit volume, which is

$$\mathscr{E} = \frac{3}{2} nkT \tag{7.158}$$

Thus,

$$\boxed{t_{\text{cool}} = \frac{\mathscr{E}}{\mathscr{C}} = \frac{3nkT}{2n_H n_e \Lambda(T)} \approx 3.3 \times 10^9 \frac{T_6}{n_{-3}\Lambda_{-23}(T)} \text{ yr}} \tag{7.159}$$

So, since $\Lambda \propto T^{1/2}$, we see that $t_{\text{cool}} \propto T^{1/2}\rho_{\text{gas}}^{-1}$. In a typical cluster, the cooling rates of most of the ICM's volume are much longer than the Hubble time $t_H \propto H_0^{-1}$. But, since the hot gas is in hydrostatic equilibrium, this requires it to get denser towards the cluster center. This means the cooling time must get shorter. Within the central ~100 kpc of some clusters, the cooling times can reach below 10 Gyr, meaning we should be able to see evidence of cooling. As the gas cools, its temperature drops. In order to maintain the same pressure to support the weight of the surrounding gas, it must flow inwards to increase in density. This process is known as a **cooling flow**. Clusters

[7] This is what I do research on, so this section may be more detailed than necessary...

with cooling times that drop below the age of the universe in their cores are known as **cool core clusters**. See Figure 7.30 for an example of cooling time as a function of radius.

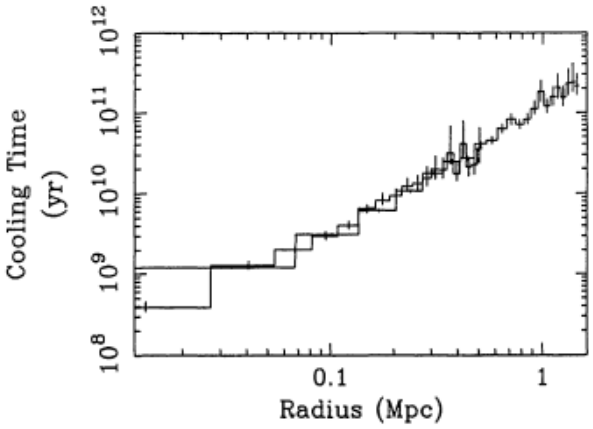


Figure 7.30: A plot of cooling time as a function of radius from the center of Abell 478[147]

We can calculate the mass deposition rate by noticing that the radiated luminosity should be equal to the change in the thermal energy and $P\mathrm{d}V$ work on a blob of gas as it falls inwards, assuming the cooling process is isobaric. Notice, since the enthalpy is defined as the internal energy plus PV, the radiated energy is simply equal to the change in enthalpy $\Delta H \equiv \Delta E + P\Delta V$ (this is a standard result from thermodynamics for isobaric processes, see e.g. https://en.wikipedia.org/wiki/Isobaric_process). Therefore, the radiated luminosity is the time derivative of ΔH:

$$L = \frac{\mathrm{d}}{\mathrm{d}t}(\Delta H) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{3}{2}Nk\Delta T + P\Delta V\right) = \frac{3}{2}\dot{N}k\Delta T + \frac{\mathrm{d}}{\mathrm{d}t}(P\Delta V) \tag{7.160}$$

Then we use $N = \rho V/\mu m_p = M/\mu m_p \longrightarrow \dot{N} = \dot{M}/\mu m_p$, and with the ideal gas law, $P\Delta V = Nk\Delta T = Mk\Delta T/\mu m_p$, so both terms combine with a factor of 5/2:

$$\boxed{L = \frac{5}{2}\frac{\dot{M}}{\mu m_p}k\Delta T} \tag{7.161}$$

If we want to consider *all* of the cooling from some maximum temperature T down to 0 (or roughly 0), then we can just replace ΔT with T. Solving for $\dot{M}$ gives:

$$\boxed{\dot{M}(<r) = \frac{2\mu m_p}{5kT}L(<r)} \tag{7.162}$$

This is known in the literature as the **classical cooling rate**, since we've made a lot of assumptions—mainly that the cooling flow is steady and isobaric and that it remains constant over several gigayears. In a typical cluster, classical cooling rates measured from the hot X-ray emitting ICM range from $\dot{M} \sim 10 - 100\,M_\odot\,\mathrm{yr}^{-1}$, while some have up to $\dot{M} \gtrsim 1000\,M_\odot\,\mathrm{yr}^{-1}$. This implies that large quantities of cold gas are being deposited within the BCG, which should then be able to form stars. However, the measured star formation rates in BCGs are typically only a few $M_\odot\,\mathrm{yr}^{-1}$, 2 orders of magnitude less than the deposition rate. This is the first puzzle that makes up what is known as the **cooling flow problem**, but there's more.

We can further break this down to obtain the spectrum of a cooling flow by considering the differential form of equation (7.161):

$$\mathrm{d}L = \frac{5}{2}\frac{\dot{M}}{\mu m_p} k \mathrm{d}T = n_e n_H \Lambda(T) \mathrm{d}V \tag{7.163}$$

Now, to get the specific luminosity, we need a per-unit-frequency version of the cooling function, $\Lambda_\nu(T)$, that specifies the cooling rate as a function of temperature for an individual frequency, as opposed to $\Lambda(T)$ which is the cooling rate at temperature T integrated over all frequencies. With this, we have

$$\mathrm{d}L_\nu = n_e n_H \Lambda_\nu(T) \mathrm{d}V \tag{7.164}$$

Comparing this to the previous equation, we find the form of the cooling flow spectrum looks like[147]

$$\boxed{L_\nu = \frac{5}{2}\frac{\dot{M}}{\mu m_p} k \int_0^{T_{\max}} \frac{\Lambda_\nu(T)}{\Lambda(T)} \mathrm{d}T} \tag{7.165}$$

where $\dot{M}$ becomes a sort of normalization factor. Note that $\Lambda_\nu(T)$ can be written as $\varepsilon_\nu / n_e n_H$ where ε_ν is the emissivity (energy per unit time per unit volume per unit frequency). Thus, the cooling flow spectrum can be calculated if we know the emissivity and the cooling function. This is theoretically simple for bremsstrahlung radiation (see §5.Q74), but for line emission is becomes exceedingly complicated, and oftentimes simulations (such as CLOUDY) need to be used to compute ε_ν as a function of ν and T. For this reason, cooling rates from specific emission lines are typically quoted simply with Γ factors:

$$\boxed{L_{\text{line}} = \Gamma_{\text{line}} \dot{M}} \quad \text{where} \quad \Gamma_{\text{line}} \equiv \frac{\alpha}{2}\frac{k}{\mu m_p} \int_0^{T_{\max}} \frac{\Lambda_{\text{line},\nu}(T)}{\Lambda(T)} \mathrm{d}T \tag{7.166}$$

The factor α is 5 for isobaric cooling and 3 for isochoric cooling. Notice that by recasting the integral over temperature to an integral over time, we can recover a simpler expression for Γ[148]:

$$\mathrm{d}t = \frac{\alpha}{2}\frac{nk}{n_H n_e \Lambda(T)} \mathrm{d}T \quad \longrightarrow \quad \Gamma_{\text{line}} = \int \frac{\varepsilon_\nu}{\rho} \mathrm{d}t \tag{7.167}$$

X-ray spectroscopy of cool core clusters has revealed an absence of spectral lines with energies between 1–2 keV, leading to the cooling rates measured from these lines to be far lower than the classical cooling rates. Evidently, the gas cools down to about 1/3 of the mean temperature and then vanishes. This is the second and final puzzle that makes up the **cooling flow problem**.[107]

Potential solutions: The classical mass deposition rate we calculated in (7.162) assumed that the gas is not being reheated at all. The leading hypothesis to explain the cooling flow dilemma proposes that the gas is being reheated via feedback from an AGN in the BCG. Specifically, **mechanical feedback** from the radio jets inflates bubbles of hot gas in the ICM that buoyantly rise, dispersing energy throughout the cluster. See question §7.Q123 for a more detailed description of AGN feedback processes. The details of this process are still widely unclear. Open questions:

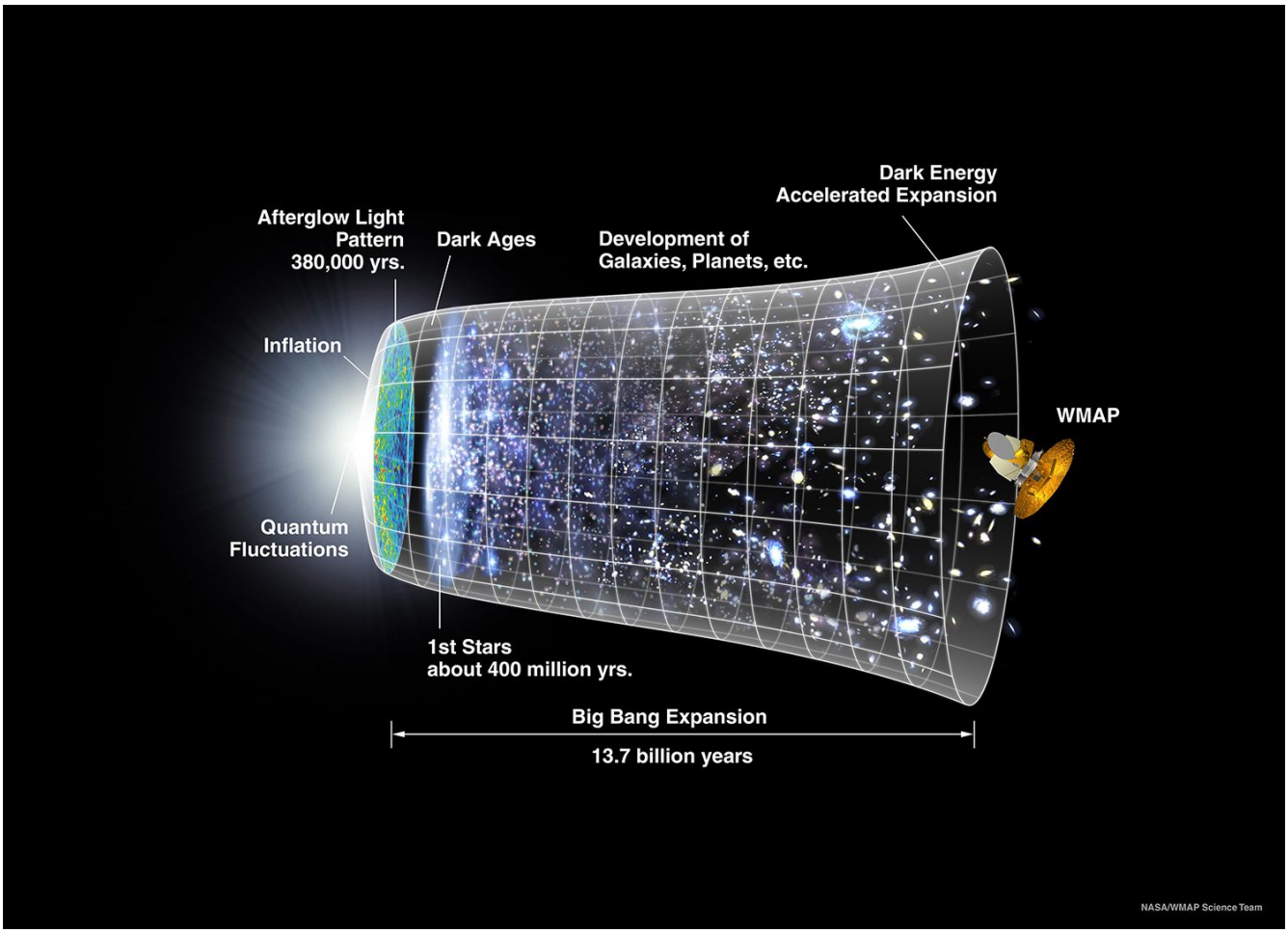
- How do the AGN jets efficiently couple to the ICM and inflate the bubbles?
- What are the duty cycles of the AGN and how does this affect cooling?

- How does AGN feedback evolve over time?

An additional source of heating that has been proposed is **thermal conduction**—i.e. the temperature gradient in the ICM may induce heat transport from the outside inwards via conduction. However, the conductivity of the ICM is not well constrained[107].

An alternative explanation is that there is a significant population of cold clouds that absorb the softer X-ray emission, making it appear as though the cooling abruptly stops when in reality it does not. As such, these are called **"hidden cooling flows."** This explains the lack of 1–2 keV X-ray emission lines, but AGN feedback is still required to explain the lack of star formation and prevent massive cooling rates in hotter gas[149].

8 Cosmology


Dark Energy
Accelerated Expansion
Afterglow Light
Pattern
380,000 yrs.
Dark Ages
Development of
Galaxies, Planets, etc.
Inflation
WMAP
Quantum
Fluctuations
1st Stars
about 400 million yrs.
Big Bang Expansion
13.7 billion years
NASA/WMAP Science Team

126. Summarize the state of the experimental field of 21 cm cosmology and what it hopes to accomplish.

See §5.Q80 for a description of the 21 cm spin-flip transition of neutral Hydrogen.

Experimental 21 cm cosmology attempts to utilize this transition to directly probe the amount of neutral Hydrogen in the universe as a function of redshift. The hope is that by doing so we can better constrain when the **epoch of reionization** occurred, which will allow us to better understand how and when the first stars and quasars formed and evolved. Reionization is the period when our universe transitioned from mostly neutral to mostly ionized (due to radiation from the first stars and QSOs). Currently it is only constrained to be somewhere between $z \sim 5-15$ (see Question §8.Q141; Figure 8.18)[150,151].

How is this done?

The CMB radiation is essentially used as a backlight—CMB photons at 21 cm can be absorbed by neutral hydrogen and excite this transition, so we should expect to see a dip in the CMB wherever there is neutral Hydrogen. We can use the radiative transfer equation (see §12.11) to predict what kind of radiation we will see:

$$\frac{\mathrm{d}I_\nu}{\mathrm{d}s} = j_\nu - \alpha_\nu I_\nu \quad \longrightarrow \quad \frac{\mathrm{d}I_\nu}{\mathrm{d}\tau_\nu} = S_\nu - I_\nu \tag{8.1}$$

where $S_\nu = j_\nu/\alpha_\nu$ is the source function. If we assume S_ν is constant over s (the line of sight), we can integrate and obtain

$$I_\nu(s) = I_\nu(0)e^{-\tau_\nu(s)} + S_\nu(1-e^{-\tau_\nu(s)}) \tag{8.2}$$

If we then assume the optically thin regime, so $\tau_\nu \ll 1$ and $e^{-\tau_\nu(s)} \simeq 1-\tau_\nu(s)$, we have

$$I_\nu(s) - I_\nu(0) \simeq (S_\nu - I_\nu(0))\tau_\nu(s) \tag{8.3}$$

Or, written in terms of the brightness temperature (see §5.Q85) and spin temperature (see §5.Q80):

$$T_b(s) - T_b(0) = (T_s - T_b(0))\tau_\nu(s) \tag{8.4}$$

Here, T_s is the spin temperature of the H I gas in the IGM, $T_b(0)$ is the brightness temperature of the CMB at 21 cm, and $T_b(s)$ is the observed brightness temperature. Thus, 21 cm cosmology focuses on measuring the temperature difference $\Delta T_b \equiv T_b(s) - T_b(0)$.

The difference in temperatures depends on the optical depth, and the optical depth depends on the amount of neutral H I at a given redshift that absorbs CMB radiation. Thus, we can relate ΔT_b to the neutral Hydrogen fraction and the baryon density Ω_b by

$$\tau_\nu(s) \propto n_{\mathrm{HI}}(z)\mathrm{d}s \propto X_{\mathrm{HI}}(z)\Omega_b(z)\frac{c}{H(z)}\frac{1}{1+z}\mathrm{d}z \tag{8.5}$$

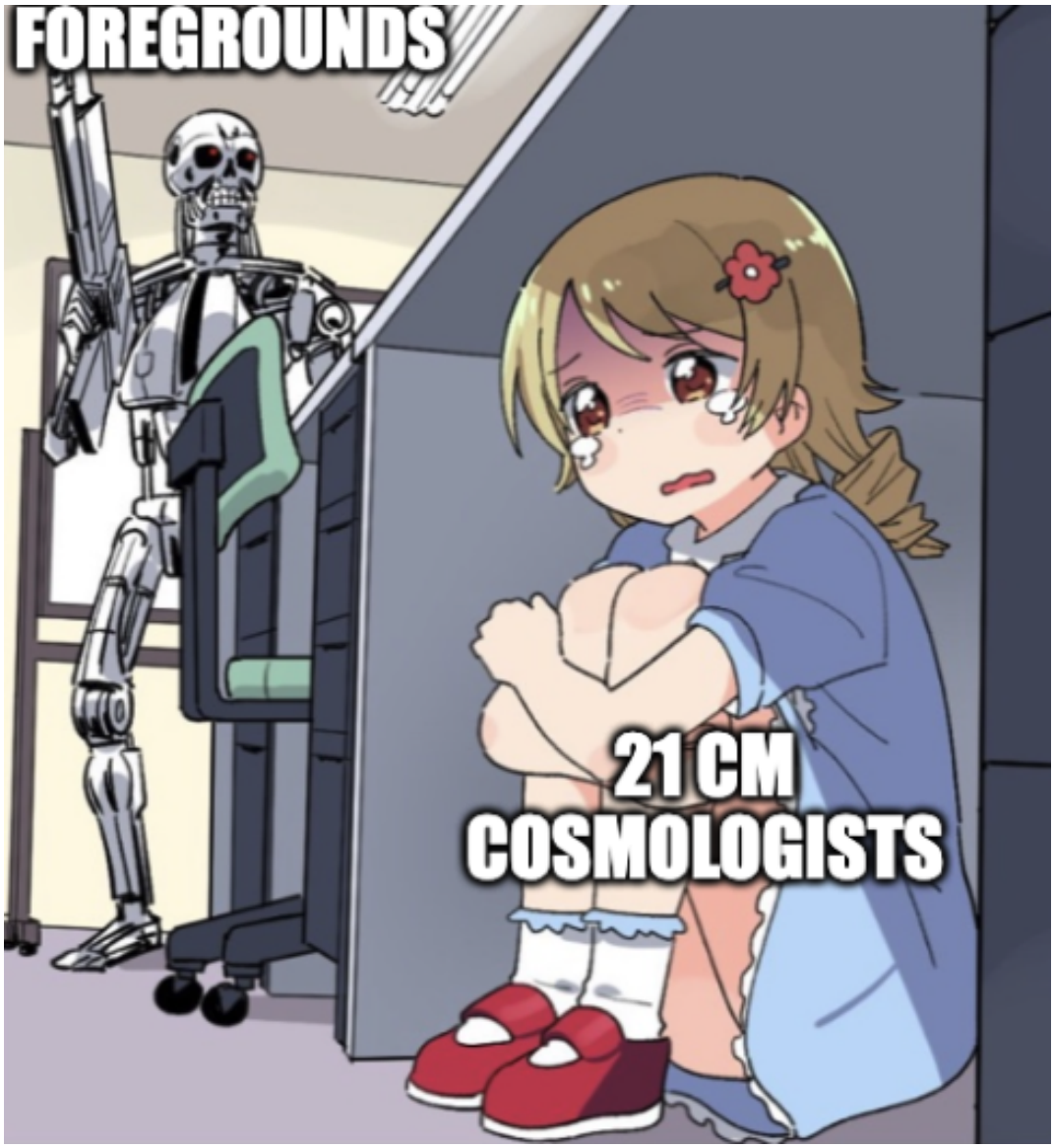
Current state of the field:

- CHIME: Measures between 400–800 MHz, which is sensitive to an H I 21 cm signal at a redshift of $\sim 0.8-2.5$. Wide FOV radio telescope made up of cylinders rathers than dishes. Can also measure the BAO. Images must be stacked to increase SNR.

- HERA: Measures between 120–200 MHz, which is sensitive to an H I 21 cm signal at a redshift of $\sim 6-13$. Array of dishes with a large collecting area (10^5 m^2). Will allow us to constrain the 21 cm power spectrum (which measures the correlation between ionization bubbles).
- LOFAR: 115–230 MHz and 30–80 MHz
- MWA: 80–300 MHz
- PAPER: 100–200 MHz
- SKA: 50–350 MHz, 0.95–1.76 GHz, 4.6–14 GHz, 0.13–1.1 GHz

Challenges:

The main challenge is accounting for **foreground emission**, which is orders of magnitude brighter than the cosmological signal that we are trying to measure. This makes the post-processing very difficult—Haochen Wang at MIT/MKI does this. The foreground comes from galactic synchrotron emission (i.e. from AGN radio jets, XRBs, SNe remnants, the ISM, etc.; see §5.Q75) and unresolved extragalactic sources.



127. Summarize the state of the experimental CMB field and what it hopes to accomplish.

At a redshift of $z \sim 1100$, the universe's ambient temperature ($T \sim 3000$ K) finally becomes cool enough to allow protons and electrons to form neutral atoms in a process called **recombination** (confusingly named since this is the first time they are combining...it should just be called "combination", but I digress). This causes the number of free electrons to plummet, which in turn greatly reduces the optical depth of the universe due to electron scattering, making the universe "transparent" to photons for the first time. **The Cosmic Microwave Background (CMB)** is radiation from this time that we observe today, which has cooled down due to redshift. Today we observe the CMB as a blackbody spectrum that peaks at 2.726 K and is uniform across the whole sky up to a $\Delta T/T$ of 1 part in 10^5 (see Figure 8.1).

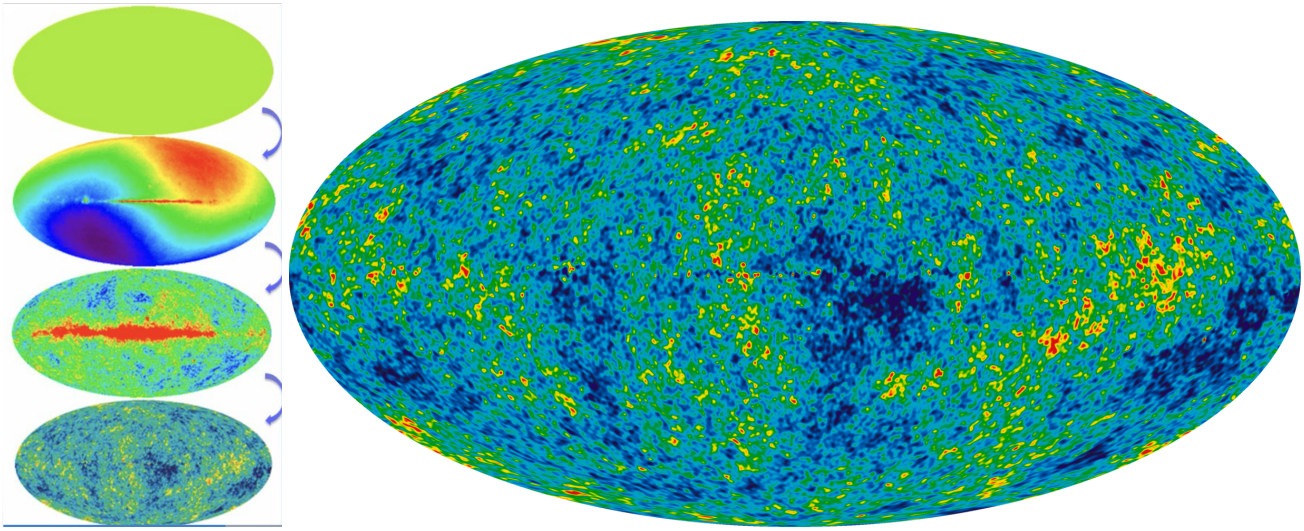


Figure 8.1: A map of the CMB as seen by WMAP. The left panel shows the different stages of subtracting systematic effects. First, the main signal of $T = 2.726$ K is subtracted off. Then, there are also signals due to the motion of the Earth and the Milky Way (causing a dipole moment), and emission from the Milky Way (i.e. synchrotron emission from XRBs and SNe remnants, free-free emission, and dust emission from the ISM), which leaves a strip through the center. The final map leaves only the residual temperature differences.

Measuring the **anisotropies** in the CMB provides us with direct probes of matter over- and under-densities in the early universe. By measuring the **power spectrum** of the CMB, we can probe these temperature differences on different angular scales and compare them directly to what should be expected based on cosmological models, which allows us to constrain parameters like the matter density, baryon density, cosmological constant, and Hubble constant (see §8.Q128 and §8.Q129 for more information).

Measuring the **polarization** of the CMB can also provide us with evidence of primordial gravitational waves, but a true detection of these has yet to be done.

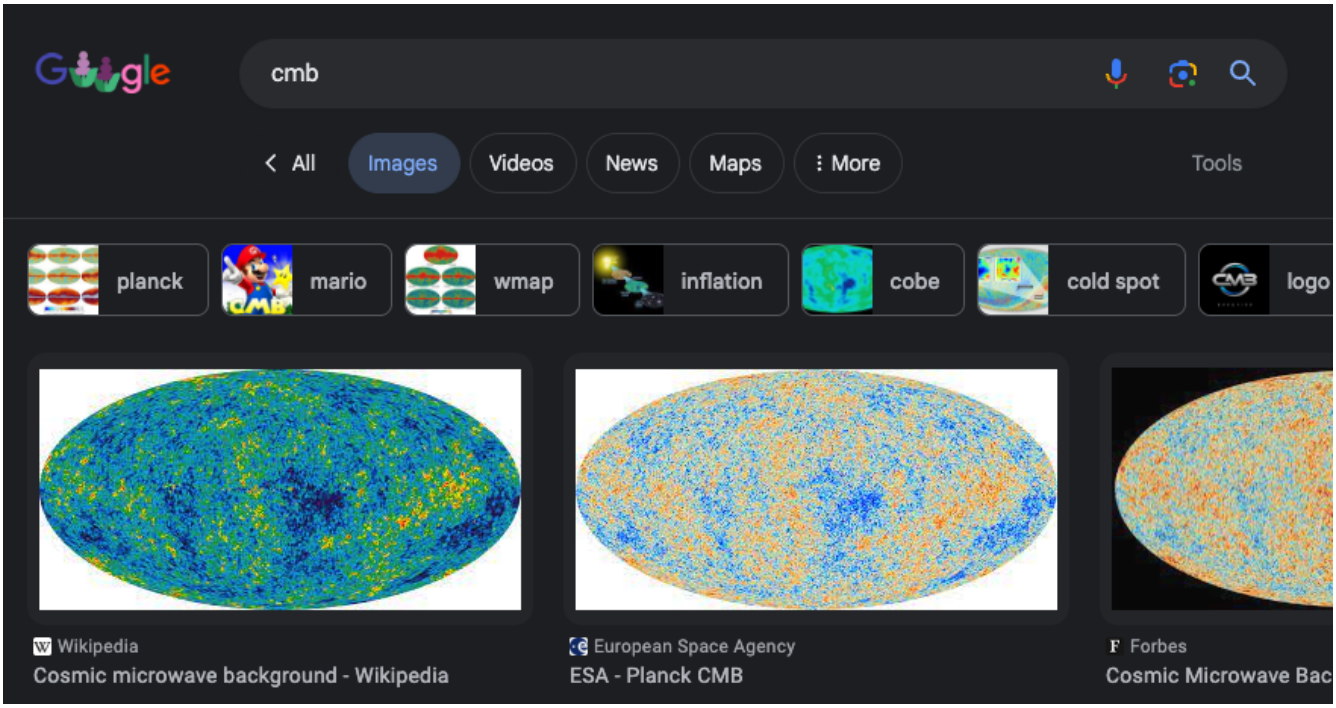
A brief history:

- 1965: Arno Penzias and Robert Wilson, working at Bell Labs, serendipitously discover a 4 K excess temperature in their antenna, and are able to attribute it to the CMB after consulting

with Robert H. Dicke, Jim Peebles, and David Wilkinson (all professors at Princeton). This discovery would later win the 1978 Nobel Prize.

- 1980: Some theoretical work on inhomogeneities is done, and the Soviet satellite RELKIT-1 is able to place limits on these.
- 1992: NASA launches the COBE satellite, which measures anisotropies of $\Delta T/T \sim 10^{-5}$. Probes the low-ℓ (large angular separations) part of the power spectrum. This wins the 2006 Nobel Prize.
- 2001: NASA launches the WMAP satellite, which extends the precision to measure up to the third peak in the power spectrum.
- 2007: SPT comes online at the South Pole and starts taking measurements of the CMB at fine angular resolutions.
- 2013: ESA launches the Planck satellite, which gets the best measurements of the CMB power spectrum to date.
- 2014: BICEP2 claims the detection of B-mode polarization, which would be evidence of primordial gravitational waves. This is later discovered to be a false positive due to dust polarization. See §5.Q79.

Also, we can't forget about the important contributions of Mario to the field of experimental CMB cosmology. After all, he's one of the first google results for "CMB"!



128. What are the anisotropies in the CMB? How do their amplitudes depend on angular scale?

The anisotropies in the CMB are temperature differences from the expected 2.726 K on the order of $\delta T \equiv \Delta T/T \sim 10^{-5}$. We can measure these anisotropies by analogy to the 2-point correlation function (i.e. (7.146)):

$$C(\Delta\theta) = \langle \delta T(\theta,\phi)\delta T(\theta+\Delta\theta,\phi)\rangle \tag{8.6}$$

We can decompose the temperature fluctuations into spherical harmonics, where we have

$$\delta T(\theta,\phi) = \sum_{\ell=0}^{\ell_{\max}} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\theta,\phi) \tag{8.7}$$

We can get the coefficients $a_{\ell m}$ by recalling that $Y_{\ell m}$ are orthonormal, i.e.

$$\int \mathrm{d}\Omega\, Y_{\ell_1 m_1}(\theta,\phi) Y^*_{\ell_2 m_2}(\theta,\phi) = \delta_{\ell_1\ell_2}\delta_{m_1 m_2} \tag{8.8}$$

Thus, we have

$$a_{\ell m} = \int \mathrm{d}\Omega\, \delta T(\theta,\phi) Y^*_{\ell m}(\theta,\phi) \tag{8.9}$$

We can then define the power spectrum in direct relation to (8.6) by

$$C_\ell = \langle |a_{\ell m}|^2\rangle = \frac{1}{2\ell+1}\sum_{m=-\ell}^{\ell} |a_{\ell m}|^2 \tag{8.10}$$

where the average is over m. This is related to $C(\Delta\theta)$ by

$$C(\Delta\theta) = \frac{1}{4\pi}\sum_{\ell}(2\ell+1) C_\ell P_\ell(\cos\Delta\theta) \tag{8.11}$$

Often when plotting the CMB power spectrum we plot the quantity $\ell(\ell+1)C_\ell$. On large angular scales, there has not been enough time yet for sound crossing, thus no structures have formed/collapsed. Then, if the CMB directly reflects the matter power spectrum $P(k) \propto k$, the quantity $\ell(\ell+1)C_\ell$ should be constant[107,119].

Primary anisotropies

1. **The Sachs-Wolfe (SW) Effect** ($\ell \lesssim 100$)
    - Photons last scattered in a potential well will experience gravitational redshift as they climb out of it, causing them to become cooler and redder.
    - From Tegmark (1995)[152]: matter overdensities also cause time dilation effects, meaning the photons we see from these regions are actually from slightly earlier (hotter) times. They are also just intrinsically hotter, so they produce hotter photons. These two effects are proportional to the gravitational potential at the time of last scattering, $\sim \Phi/3$.
    - Overall, matter overdensities will cause the CMB to look slightly cooler, while underdensities will look slightly hotter.

2. **The Large-Scale Doppler Effect** ($\ell \lesssim 100$)
    - Photons last scattered in matter moving away from us will experience a Doppler redshift, causing them to become cooler and redder.

3. **Acoustic Oscillations** ($100 \lesssim \ell \lesssim 1000$)
    - Once baryons enter the sound horizon, pressure is able to balance gravity. This causes overdensities in the baryon-photon fluid that are in the process of collapsing to rebound and start oscillating around an equilibrium point.
    - The sound speed in the baryon-photon fluid is $\sim c/\sqrt{3}$, so the sound horizon occurs at angular scales of $\theta_s \sim 1°$, corresponding to $\ell = 180°/\theta = 180$.
    - Then there are harmonics at 2ℓ, 3ℓ, etc.

4. **Silk Damping** ($\ell \gtrsim 1000$)
    - Photons have a finite mean free path between each scattering event, so they are not perfectly coupled to the baryons. Thus, the photons execute a random walk sequence. If they get far enough, they can diffuse out of a perturbation and take some baryons with them. This tends to even out any differences in temperature between hot and cold regions within some length scale.
    - The mean free path of photons due to Thomson scattering is $\lambda = 1/n_e\sigma_T$, and each photon is expected to experience N scatterings, where N is given by

$$N = \frac{1}{3}\frac{ct_{1100}}{\lambda} \tag{8.12}$$

      where t_{1100} is the age of the universe at $z = 1100$, and the factor of $1/3$ accounts for the 3D nature of the random walk.
    - We can estimate t_{1100} in the matter-dominated regime as (see §12.12)

$$t_{1100} \simeq \frac{2}{3}\frac{1}{H(1100)} \simeq 10^{13}\,\mathrm{s} \tag{8.13}$$

    - Similarly, we can estimate n_e from the baryon density Ω_b:

$$n_e = \frac{\rho_e}{m_p}\Omega_b(1+z)^3 \simeq 313\,\mathrm{cm}^{-3} \tag{8.14}$$

    - Then $r_{\mathrm{silk}} = \sqrt{N}\lambda$ is the average distance expected to be traveled after N random walk steps.

$$r_{\mathrm{silk}} = \sqrt{\frac{1}{3}\lambda c t_{1100}} = \sqrt{\frac{1}{3}\frac{1}{n_e\sigma_T}ct_{1100}} \simeq 7.5\,\mathrm{kpc} \tag{8.15}$$

    - Which corresponds to a mass scale of

$$M_{\mathrm{silk}} = \frac{4}{3}\pi r_{\mathrm{silk}}^3\bar{\rho} \simeq 10^{13}\,M_\odot \tag{8.16}$$

    - Thus, anisotropies above this length scale will tend to be smoothed out[153,119].

Secondary anisotropies

1. **The Integrated Sachs-Wolfe (ISW) Effect**
    - This is similar to the SW effect, except applied to time-changing potentials. Thus, it is only important on large angular scales (small ℓ). In a flat universe, this can happen for two separate reasons.
    - The **early** ISW effect occurs before the universe becomes matter dominated (i.e. when it is radiation dominated)—in this regime, the density perturbations grow more slowly than the scale factor, causing the potential to decay until the universe becomes matter dominated.
    - The **late** ISW effect occurs after the universe becomes dominated by the cosmological constant, which once again causes the gravitational potential to decay over time.
2. **The Sunyaev-Zel'dovich (SZ) Effect**
    - CMB photons can gain energy if they are inverse Compton scattered by high energy electrons (i.e. in the ICM of galaxy clusters). Important only for the highest ℓ's (smallest scales).
    - Comes in 2 forms: thermal SZ and kinematic SZ. For more details, see §8.Q147.
3. **Gravitational Lensing**
    - The angular difference that we see between two photons can be changed by gravitational lensing, causing small perturbations in the observed anisotropies on the order of a few percent.
4. **Thomson Scattering**
    - A photon can be scattered by an electron on its way from $z = 1100$ to us, especially after reionization. This means the location we measure a photon from may be different from where it actually came from. This has the effect of "smearing out" the anisotropies because we measure a weighted average of the temperature from the $z = 1100$ CMB.

The CMB power spectrum, with each of these anisotropies labeled, is shown in Figure 8.2.

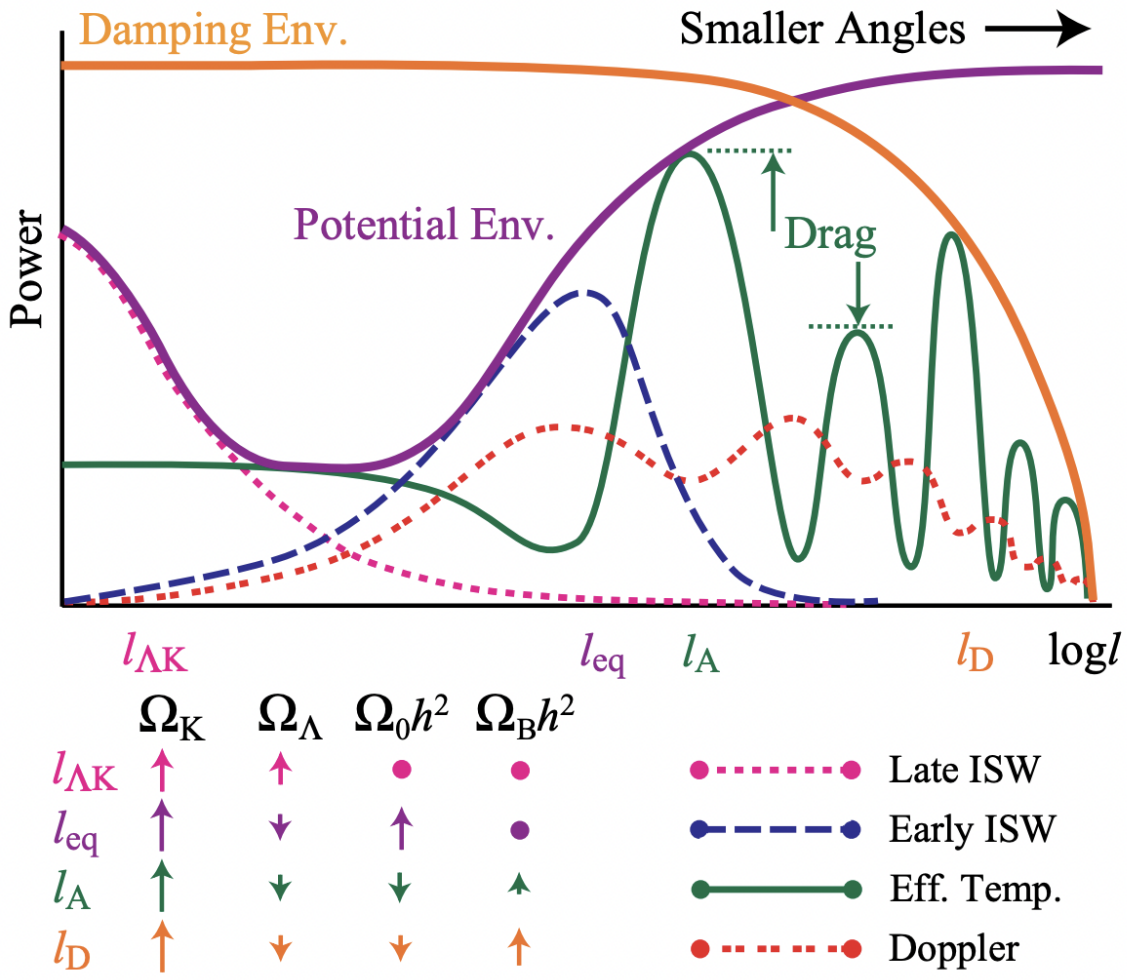


Figure 8.2: The CMB power spectrum, showing the anisotropies from late ISW (magenta), early ISW (blue), Doppler shifts (red), acoustic oscillations (green), and Silk damping (orange). Note that the regular SW effect is included in the green curve along with the acoustic oscillations[154]. The legend also shows how different cosmological parameters affect each of these anisotropies, which we will expand on more in the next question (§8.Q129).

129. What are acoustic peaks in the CMB power spectrum? What do their locations and amplitudes tell us?

The peaks in the CMB power spectrum are a result of the evolution of overdensities in the early universe that causes them to undergo acoustic oscillations for a window of time. For a detailed explanation of this process, see §8.Q131.

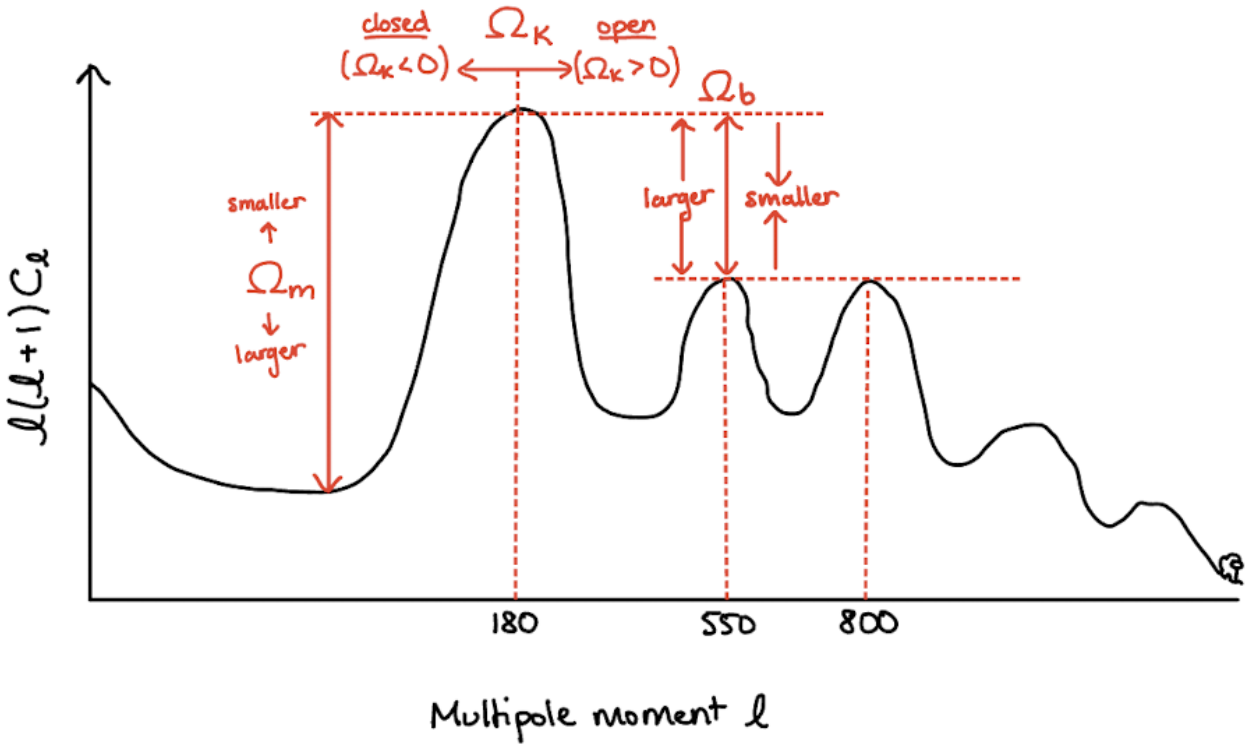


Figure 8.3: The CMB power spectrum, annotated with the information we can obtain from the locations and sizes of the acoustic peaks.

A sketch of the CMB power spectrum is shown in Figure 8.3. We notice that the first acoustic peak occurs at a multiple moment of $\ell \simeq 180$, the second at $\ell \simeq 550$, and the third at $\ell \simeq 800$. The locations and amplitudes of these peaks depend on the cosmological parameters of the universe—namely, the matter density Ω_m, baryon density Ω_b, and curvature density Ω_K.

The location of the first peak[119]:

We can derive the location of the first peak by asking the question: "At what angular scales should we see the lowest harmonic oscillation period?"

Different oscillation periods corresponds to having more oscillations in between the time when the an overdensity enters the sound horizon and when it hits recombination. Thus, the first, lowest harmonic occurs when there is exactly 1/2 of an oscillation between these two points—that is, the oscillation is at its first maximum compression exactly at recombination. This corresponds to sound being able to cross the overdensity exactly 1 time before recombination. The first few harmonics are depicted visually in Figure 8.4 (for an explanation of this figure, see §8.Q131).

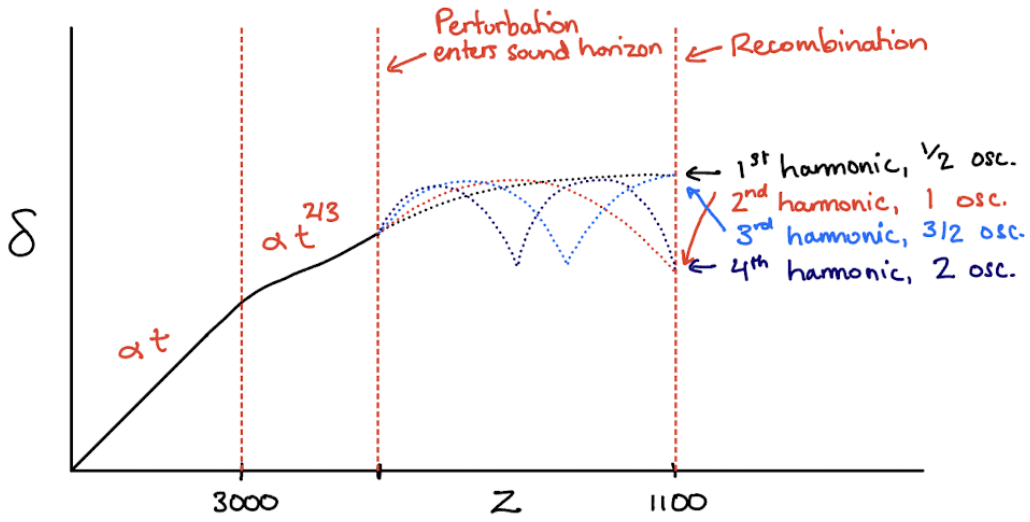


Figure 8.4: The evolution of an overdensity perturbation over cosmic time, highlighting the period of baryon acoustic oscillations with different harmonic periods.

We can find the horizon scale for the matter-dominated universe at recombination ($z = 1100$) with the horizon distance derived in Appendix section §12.12. We just have to replace the speed of light c with the speed of sound $c_s = c/\sqrt{3}$:

$$r_s = 3c_s t \quad , \quad t \simeq \frac{2}{3}\frac{1}{H(z)} \simeq \frac{2}{3}\frac{1}{H_0\sqrt{\Omega_m(1+z)^3}} \tag{8.17}$$

Which corresponds to angular scales of

$$\theta_s = \frac{r_s}{d_A} \quad , \quad d_A = \frac{1}{1+z}\frac{c}{H_0}\int_0^z \frac{\mathrm{d}z'}{E(z')} \tag{8.18}$$

Here, d_A is the angular diameter distance. So we have

$$d_A \simeq \frac{1}{1+z}\frac{c}{H_0\sqrt{\Omega_m}}\int_0^z \frac{\mathrm{d}z'}{(1+z')^{3/2}} = \frac{1}{1+z}\frac{c}{H_0\sqrt{\Omega_m}}\left[\frac{-2}{\sqrt{1+z'}}\right]_0^z \tag{8.19}$$

$$\simeq \frac{1}{1+z}\frac{c}{H_0\sqrt{\Omega_m}}\left(2 - \frac{2}{\sqrt{1+z}}\right) \simeq \frac{2c}{zH_0\sqrt{\Omega_m}} \tag{8.20}$$

Where we approximated $1+z \simeq z$ and $1/\sqrt{1+z} \simeq 0$. Putting it all together, we get about 1 degree

$$\theta_s = \frac{r_s}{d_A} \simeq \frac{1}{\sqrt{3z}} \simeq 1^\circ \tag{8.21}$$

Which corresponds to a multipole moment of

$$\ell = \frac{\pi}{\theta} = \frac{180}{1} = 180 \tag{8.22}$$

Note that in this derivation we assumed $c_s = c/\sqrt{3}$ which gives a sound horizon of $r_s \simeq 0.26$ Mpc, but in actuality, considering the relative densities of baryons and photons in the early universe, it's closer to $c/\sqrt{6}$. Modern measurements from the Planck collaboration lead to a smaller sound horizon of ~ 0.14 Mpc and an $\ell \sim 220$[155].

Also notice that in calculating the angular diameter distance d_A, we assumed a flat universe ($\Omega_K = 0$). In fact, if we had an open or a closed universe, the effect on d_A at a redshift of 1100 would be drastic! See Figure 8.5 for a visual representation of this effect. In essence, an open universe with hyperbolic geometry would cause θ_s to be smaller for the same size object, while a closed universe with spherical geometry would cause θ_s to be larger. We capture this effect by absorbing it into the angular diameter distance. Note that this effect does not just apply to the first peak, but all of the peaks. Thus, **the locations of the peaks depend on the curvature parameter Ω_K.** Results have found that this is consistent with $\Omega_K \simeq 0$, i.e. our universe is spatially flat.

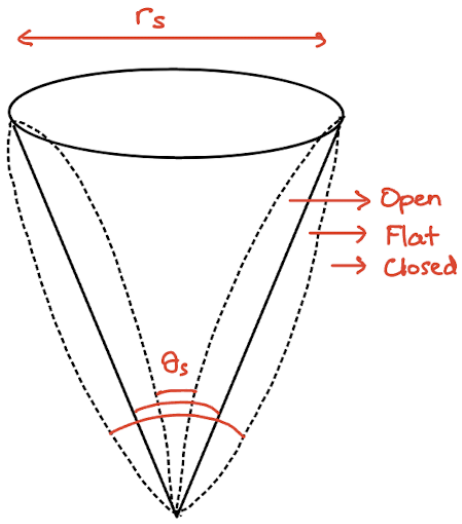


Figure 8.5: A diagram showing how having an open or closed geometry would change the relationship between r_s and θ_s, causing the angular diameter distance to grow or shrink.

The absolute amplitude of the peaks[119]:

Consider the approximation for the age of the universe in the matter-dominated regime (valid at recombination):

$$t \simeq \frac{2}{3}\frac{1}{H(z)}\bigg|_{z=1100} \simeq \frac{2}{3}\frac{1}{H_0\sqrt{\Omega_m(1+z)^3}} \tag{8.23}$$

Notice that $t \propto \Omega_m^{-1/2}$. In other words, if Ω_m is larger, the universe is younger at a redshift of 1100. This gives less time for structure to grow in the early universe, and thus there is less overall power in the CMB power spectrum. Thus, **the amplitude of the whole spectrum depends on the matter parameter Ω_m.** Results show $\Omega_m \simeq 0.27$.

The relative amplitude of the peaks[119,156]:

Recall from our discussion about the peak locations that the first harmonic has 1/2 of an oscillation, the second harmonic has a full oscillation, etc. In general, the odd harmonics correspond to points of

maximum compression while the even harmonics correspond to points of **maximum rarefaction**. So, we have to ask: how strong will the compressions and rarefactions be relative to each other?

In a primordial overdensity, the total mass M will be dominated by dark matter. If we increase the **baryonic mass** m_b, the total mass will barely change, but the small mass of the matter that is actually oscillating (the baryons) will increase. Conceptually, this is just like mass loading a spring! If the spring has more mass, it will fall further (maximum compression) since it starts with more gravitational potential energy (GMm_b/r^2), but during the rebound it will return to the same spot it started at (maximum rarefaction) since we have barely changed the total mass $M \gg m_b$. See this depicted visually in Figure 8.6.

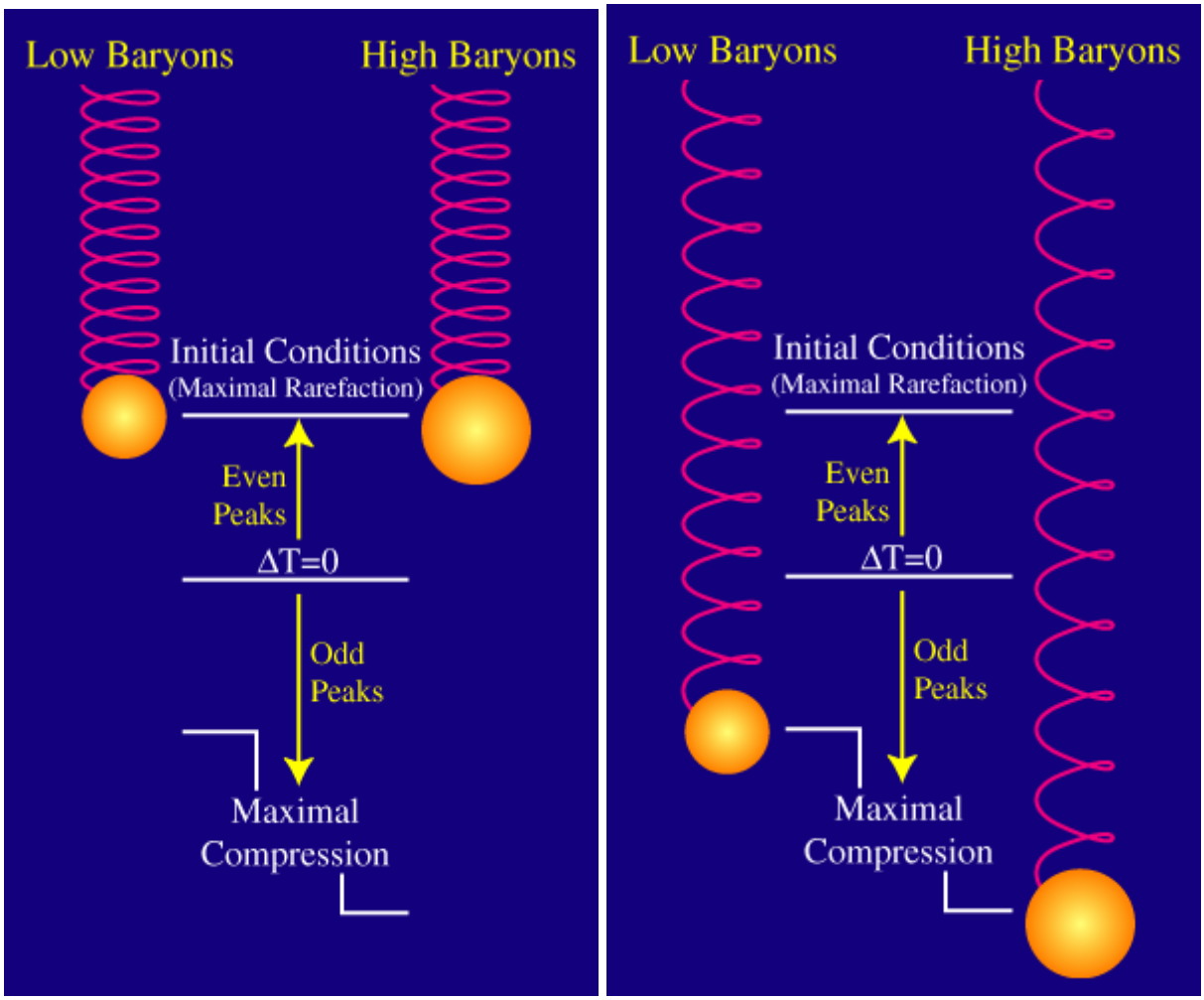


Figure 8.6: The effect of mass loading a spring on the relative strength of compression and rarefaction[156].

This has the effect of enhancing the compressions (odd peaks) relative to the rarefactions (even peaks). In other words, **the relative amplitudes of the even and odd peaks depend on the baryon density Ω_b.** Results give a baryonic contribution of $\Omega_b \simeq 0.05$.

The Temperature of the CMB[11]

The CMB temperature can directly give us the radiation density Ω_r. This is because the CMB is a near-perfect blackbody, so we can simply relate the temperature of the blackbody to the photon's

energy density using

$$\rho_\gamma c^2 = u_\gamma = aT^4 \tag{8.24}$$

Where $T = 2.7$ K. In fact, we can generalize this to include neutrinos as well. If we call the effective number of degrees of freedom shared by the neutrinos and the photons $g_* \approx 3.363$ (this is a weighted sum since photons contribute 2 d.o.f. and neutrinos contribute 6 d.o.f.), then we have

$$\rho_r = \frac{u_r}{c^2} = \frac{g_*}{2}\frac{aT^4}{c^2} \tag{8.25}$$

Then we just use the definition of Ω_r

$$\Omega_r = \frac{\rho_r}{\rho_{\text{crit}}} = \frac{8\pi G\rho_r}{3H_0^2} = \frac{4\pi G g_* aT^4}{3H_0^2 c^2} \tag{8.26}$$

Which gives $\Omega_r \simeq 8.24 \times 10^{-5}$.

Miscellaneous[119]:

From the previous sections, we have estimates of Ω_K, Ω_m, Ω_b, and Ω_r. From these, we can infer the remaining parameters:

$$\Omega_\Lambda = 1 - \Omega_K - \Omega_m - \Omega_r \tag{8.27}$$

$$\Omega_c = \Omega_m - \Omega_b \tag{8.28}$$

where Ω_Λ is the density from the cosmological constant, and Ω_c is the density of cold dark matter. This gives $\Omega_\Lambda \simeq 0.73$ and $\Omega_c \simeq 0.22$. We can also estimate Ω_Λ from the late-time integrated Sachs-Wolfe effect (which causes the slight increase in power at the lowest ℓ's).

Note: Some of the above effects are also degenerate with h ($\equiv H_0/100$ km s^{-1} Mpc^{-1}). For example, notice in equation (8.23) that t is actually proportional to $t \propto \Omega_m^{-1/2} h^{-1}$. Thus, we can get a general idea of h, but it will be degenerate with what we have for Ω_m. The baryon density also has the same degeneracy. This is why in the CMB results papers these values are often reported as $\Omega_m h^2$ and $\Omega_b h^2$. This degeneracy can partly be broken by analyzing the gravitational lensing anisotropies in the CMB. These require a model of the lensing potential, which gives a matter power spectrum and a measurement of the parameter $\sigma_8 \Omega_m^{1/4}$.

130. Write down the Jeans equation for a disturbance propagating in a self-gravitating medium. From this show how to find the critical wavenumber for propagating modes. What is the Jeans mass?

This question is a bit involved. The full derivation of the Jeans equation is likely not necessary to know since they only ask for us to write it down, so one can probably skip the first section here.

Derivation of the Jeans Equation[119,127]:

Let's say we have a total of N particles, each with their own positions $\mathbf{x}$ and velocities $\mathbf{v}$. Then, we can define a **phase space distribution function** $f(\mathbf{x},\mathbf{v},t)\mathrm{d}^3\mathbf{x}\mathrm{d}^3\mathbf{v}$ which gives the probability of finding a particle within $(\mathbf{x},\mathbf{x}+\mathrm{d}^3\mathbf{x})$ and $(\mathbf{v},\mathbf{v}+\mathrm{d}^3\mathbf{v})$ at time t.

$$f(\mathbf{x},\mathbf{v},t) \equiv \frac{1}{N}\frac{\mathrm{d}N}{\mathrm{d}^3\mathbf{x}\mathrm{d}^3\mathbf{v}} \quad \text{such that} \quad \int \mathrm{d}^3\mathbf{x}\mathrm{d}^3\mathbf{v}\, f(\mathbf{x},\mathbf{v},t) = 1 \tag{8.29}$$

If we make the following assumptions: 1) there are no sources or sinks (particles are not created or destroyed), and 2) there are no hard scatterings; then we can apply the continuity equation with our distribution function in 6D phase space, with a 6D phase vector $\mathbf{w} \equiv (\mathbf{x},\mathbf{v})$:

$$\frac{\partial f}{\partial t} + \nabla\cdot(f\dot{\mathbf{w}}) = 0 \quad\longrightarrow\quad \frac{\partial f}{\partial t} + \frac{\partial}{\partial \mathbf{x}}(f\dot{\mathbf{x}}) + \frac{\partial}{\partial \mathbf{v}}(f\dot{\mathbf{v}}) = 0 \tag{8.30}$$

If we have a gravitational potential Φ, then $\dot{\mathbf{v}} = -\nabla\Phi$, so we have

$$\boxed{\frac{\partial f}{\partial t} + \mathbf{v}\cdot\frac{\partial f}{\partial \mathbf{x}} - \nabla\Phi\cdot\frac{\partial f}{\partial \mathbf{v}} = 0} \tag{8.31}$$

This is the **Collisionless Boltzmann Equation**. If we integrate over velocity, we obtain the zeroth moment

$$n \equiv \int \mathrm{d}^3\mathbf{v}\, f(\mathbf{x},\mathbf{v},t) \quad\longrightarrow\quad \frac{\partial n}{\partial t} + \frac{\partial}{\partial \mathbf{x}}(n\mathbf{v}) - \nabla\Phi\times f\Big|_{\mathbf{v}=-\infty}^{\mathbf{v}=+\infty} = 0 \tag{8.32}$$

The last term goes to 0 since common sense dictates our probability of having an infinite velocity should be 0. Thus, we recover the 3D spatial continuity equation. Note: n is **not** a number density here, it is a probability density of finding a particle within a volume $(\mathbf{x},\mathbf{x}+\mathrm{d}^3\mathbf{x})$.

$$\frac{\partial n}{\partial t} + \frac{\partial}{\partial \mathbf{x}}(n\mathbf{v}) = 0 \tag{8.33}$$

Now, if we multiply by a velocity component v_j and then integrate over velocity, we obtain the first moment. From now on, I'll stick with index notation for all the vectors to avoid any confusion

$$\frac{\partial}{\partial t}\int \mathrm{d}^3\mathbf{v}\, v_j f + \sum_{i=1}^{3}\frac{\partial}{\partial x_i}\int \mathrm{d}^3\mathbf{v}\, v_i v_j f - \sum_{i=1}^{3}\frac{\partial\Phi}{\partial x_i}\int \mathrm{d}^3\mathbf{v}\, v_j\frac{\partial f}{\partial v_i} = 0 \tag{8.34}$$

There are a few useful quantities we can use to simplify the above expression—the average velocity

and the velocity dispersion:

$$\langle v_i \rangle = \frac{1}{n} \int \mathrm{d}^3\mathbf{v} v_i f(\mathbf{x}, \mathbf{v}, t) \ , \ \langle v_i v_j \rangle = \frac{1}{n} \int \mathrm{d}^3\mathbf{v}\, v_i v_j f(\mathbf{x}, \mathbf{v}, t) \ , \ \sigma_{ij}^2 \equiv \langle v_i v_j \rangle - \langle v_i \rangle \langle v_j \rangle \tag{8.35}$$

These terms, along with the continuity equation (8.33) allows us to write the first term in (8.34) as

$$\frac{\partial}{\partial t}(n\langle v_j \rangle) = \frac{\partial n}{\partial t}\langle v_j \rangle + n\frac{\partial \langle v_j \rangle}{\partial t} = -\langle v_j \rangle \sum_{i=1}^{3} \frac{\partial}{\partial x_i}(n\langle v_i \rangle) + n\frac{\partial \langle v_j \rangle}{\partial t} \tag{8.36}$$

We can rewrite the second term as

$$\sum_{i=1}^{3} \frac{\partial}{\partial x_i}[n(\sigma_{ij}^2 + \langle v_i \rangle \langle v_j \rangle)] \tag{8.37}$$

And for the third term, we integrate by parts to move the derivative

$$\int \mathrm{d}^3\mathbf{v}\, v_j \frac{\partial f}{\partial v_j} = v_j f \Big|_{-\infty}^{\infty} - \int \mathrm{d}^3\mathbf{v} \frac{\partial v_j}{\partial v_i} f = -\int \mathrm{d}^3\mathbf{v} \delta_{ij} f \tag{8.38}$$

$$\longrightarrow \sum_{i=1}^{3} \delta_{ij} \frac{\partial \Phi}{\partial x_i} \int \mathrm{d}^3\mathbf{v} f = n\frac{\partial \Phi}{\partial x_j} \tag{8.39}$$

Putting it all together from (8.36), (8.37), and (8.39), we have

$$n\frac{\partial \langle v_j \rangle}{\partial t} - \langle v_j \rangle \sum_{i=1}^{3} \frac{\partial}{\partial x_i}(n\langle v_i \rangle) + \sum_{i=1}^{3} \frac{\partial}{\partial x_i}[n(\sigma_{ij}^2 + \langle v_i \rangle \langle v_j \rangle)] + n\frac{\partial \Phi}{\partial x_j} = 0 \tag{8.40}$$

Further simplifying, we obtain the **Jeans Equation**:

$$\boxed{\frac{\partial \langle v_j \rangle}{\partial t} + \sum_{i=1}^{3} \langle v_i \rangle \frac{\partial \langle v_j \rangle}{\partial x_i} = -\frac{1}{n} \sum_{i=1}^{3} \frac{\partial}{\partial x_i}(n\sigma_{ij}^2) - \frac{\partial \Phi}{\partial x_j}} \tag{8.41}$$

With the interpretation of each term as follows, from left to right:

- Acceleration of the particles
- Kinematic viscosity / shear
- Thermal pressure gradient
- Gravity

Now we can simplify the notation a bit if we introduce the thermal pressure $P = \rho\sigma_{ij}^2$ and redefine $\mathbf{v}$ as the average velocity vector, i.e. $\mathbf{v} \equiv (\langle v_x \rangle, \langle v_y \rangle, \langle v_z \rangle)$. Then, we have the **Jeans Equation** (once again) as:

$$\boxed{\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\nabla\Phi - \frac{1}{\rho}\nabla P} \tag{8.42}$$

Applying the Jeans Equation to a propagating disturbance[119]:

Let's say our disturbance is characterized by small parameter ε such that we can write the density, velocity, pressure, and potential in our medium as some constant values plus a small perturbation term:

$$\rho(\mathbf{x},t) = \rho_0 + \varepsilon\rho_1(\mathbf{x},t) \tag{8.43}$$
$$\mathbf{v}(\mathbf{x},t) = \mathbf{v}_0 + \varepsilon\mathbf{v}_1(\mathbf{x},t) \tag{8.44}$$
$$P(\mathbf{x},t) = P_0 + \varepsilon P_1(\mathbf{x},t) \tag{8.45}$$
$$\Phi(\mathbf{x},t) = \Phi_0 + \varepsilon\Phi_1(\mathbf{x},t) \tag{8.46}$$

Then, we assume the background is uniform and static, i.e. we assume $v_0 = P_0 = \Phi_0 = 0$, and $\rho_0 = \text{const}$. These assumptions are actually inconsistent, since according to Poisson's equation we would have

$$\nabla^2\Phi_0 = 4\pi G\rho_0 \quad\longrightarrow\quad 0 = 4\pi G\rho_0 \tag{8.47}$$

This is called the **Jeans Swindle**. Legend has it that this technique was once used by a ne'er-do-well to convince a merchant that some denim garments had absolutely no potential on the market, despite the fact that they had the density of an average pair of trousers. Believing them worthless, the merchant relinquished the clothing and went about his day, completely unaware of the true potential of his wares. The ne'er-do-well later sold off the denim attire for a profit of 4 homemade pies and ate well for the next 2 days.

For our analysis, where we only consider a small perturbation, the Jeans Swindle actually does not affect our results at all, so we'll just ignore it and continue on.

First we use the **Jeans equation**, ignoring anything of order $\mathcal{O}(\varepsilon^2)$ and throwing away all terms with $v_0 = P_0 = \Phi_0 = 0$:

$$\frac{\partial(\varepsilon\mathbf{v}_1)}{\partial t} + [(\mathbf{v}_0 + \varepsilon\mathbf{v}_1)\cdot\nabla](\varepsilon\mathbf{v}_1) = -\nabla(\varepsilon\Phi_1) - \frac{1}{\rho_0 + \varepsilon\rho_1}\nabla(\varepsilon P_1) \tag{8.48}$$
$$\frac{\partial\mathbf{v}_1}{\partial t} = -\nabla\Phi_1 - \frac{1}{\rho_0}\nabla P_1 \tag{8.49}$$

With the definition of the sound speed $c_s^2 = \partial P/\partial\rho$, we have

$$\frac{\partial\mathbf{v}_1}{\partial t} = -\nabla\Phi_1 - \frac{c_s^2}{\rho_0}\nabla\rho_1 \tag{8.50}$$

Then, we can use the **continuity equation**, doing the same thing

$$\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{v}) = 0 \tag{8.51}$$

$$\varepsilon\frac{\partial\rho_1}{\partial t} + \nabla\cdot(\varepsilon\rho_0\mathbf{v}_1) = 0 \quad\longrightarrow\quad \frac{\partial\rho_1}{\partial t} + \rho_0\nabla\cdot\mathbf{v}_1 = 0 \tag{8.52}$$

Now we differentiate this with respect to time

$$\frac{\partial^2 \rho_1}{\partial t^2} + \rho_0 \frac{\partial}{\partial t}(\nabla \cdot \mathbf{v}_1) = \frac{\partial^2 \rho_1}{\partial t^2} + \rho_0 \nabla \cdot \frac{\partial \mathbf{v}_1}{\partial t} = 0 \tag{8.53}$$

And we plug in our result from the Jeans equation for the velocity derivative

$$\frac{\partial^2 \rho_1}{\partial t^2} + \rho_0 \left(-\nabla^2 \Phi_1 - \frac{c_s^2}{\rho_0} \nabla^2 \rho_1 \right) = 0 \tag{8.54}$$

And, using Poisson's equation, we have

$$\boxed{\frac{\partial^2 \rho_1}{\partial t^2} - c_s^2 \nabla^2 \rho_1 = 4\pi G \rho_1 \rho_0} \tag{8.55}$$

This is a wave equation for the density perturbation traveling at a speed c_s with a source term for the gravity. If we assume a generic wave solution of the form

$$\rho_1 = C e^{i(\mathbf{k}\cdot\mathbf{x} - \omega t)} \tag{8.56}$$

Then we recover the relationship

$$-\omega^2 \rho_1 + c_s^2 \mathbf{k}^2 \rho_1 = 4\pi G \rho_1 \rho_0 \quad \longrightarrow \quad \omega^2 = c_s^2 \mathbf{k}^2 - 4\pi G \rho_0 \tag{8.57}$$

We have two possible cases here:

- $\omega^2 > 0 \quad \longrightarrow \quad$ exponent is imaginary $\quad \longrightarrow \quad \rho_1$ is stable and propagates as a wave
- $\omega^2 < 0 \quad \longrightarrow \quad$ exponent is real $\quad \longrightarrow \quad \rho_1$ is unstable and grows exponentially (cloud collapses)

We can see that the bounding case is at $\omega = 0$. We can solve for k is this case, where we obtain

$$k^2 = \frac{4\pi G \rho_0}{c_s^2} \tag{8.58}$$

This corresponds to a length scale of

$$\lambda_J^2 = \left(\frac{2\pi}{k} \right)^2 = \frac{\pi c_s^2}{G \rho_0} \quad \longrightarrow \quad \boxed{\lambda_J = c_s \sqrt{\frac{\pi}{G \rho_0}}} \tag{8.59}$$

This is the **Jeans length**! This is an upper limit on the size of the density perturbation before it will collapse, since it corresponds to the above cases like so:

$$\omega^2 = 4\pi \left(\frac{\pi c_s^2}{\lambda^2} - G \rho_0 \right) \tag{8.60}$$

- $\lambda < \lambda_J \quad \longrightarrow \quad \omega^2 > 0 \quad \longrightarrow \quad$ stable
- $\lambda > \lambda_J \quad \longrightarrow \quad \omega^2 < 0 \quad \longrightarrow \quad$ unstable (collapse)

The Jeans Mass[119]:

We can recast the Jeans length in terms of a mass scale called the **Jeans mass**

$$\boxed{M_J = \frac{4}{3}\pi\lambda_J^3\rho_0 = \frac{4}{3}\pi\rho_0 c_s^3\left(\frac{\pi}{G\rho_0}\right)^{3/2}} \tag{8.61}$$

Note: This long and complicated derivation for the Jeans length can be done roughly in a much quicker method. Intuitively, in a cloud of gas we have gravity balanced by pressure. If gravity compresses faster than a sound wave can cross the cloud, then there is not enough time for pressure to push back against gravity. Then, all we have to do is compare the sound-crossing time t_s to the free-fall time t_{ff} (7.52):

$$t_{ff} \sim \frac{1}{\sqrt{G\rho}} \tag{8.62}$$

$$t_s \sim \frac{r}{c_s} \tag{8.63}$$

$$t_{ff} \sim t_s \longrightarrow \text{collapse!} \tag{8.64}$$

This gives us an estimate for the Jeans length of

$$\frac{\lambda_J}{c_s} \sim \frac{1}{\sqrt{G\rho}} \longrightarrow \boxed{\lambda_J \sim \frac{c_s}{\sqrt{G\rho}}} \tag{8.65}$$

Which is only off from the true value by a factor of $\sqrt{\pi}$.

131. Qualitatively, how does a density fluctuation grow over cosmic history and how does the answer differ for dark matter and baryons?

As we saw in §8.Q130, a perturbation in density will collapse on itself if its length scale λ is larger than its Jeans length λ_J. In the early universe, photons and electrons are strongly coupled, so we have a ram pressure from photons of $P = \frac{1}{3}\rho c^2$. The sound speed in this case is $c_s = \sqrt{\partial P/\partial \rho} = c/\sqrt{3}$. Using equation (8.59) for the Jeans length, we have

$$\lambda_J = \frac{c}{\sqrt{G\rho}}\sqrt{\frac{\pi}{3}} \tag{8.66}$$

Now let's compare this to the horizon scale r_H in a radiation-dominated universe. In this case, we have a scale factor $a \propto t^{1/2}$ and $\dot{a} \propto \frac{1}{2}t^{-1/2}$. So, from the Friedman equations, we have

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{1}{4t^2} = \frac{8\pi G\rho}{3} \quad \longrightarrow \quad t = \left(\frac{32\pi G\rho}{3}\right)^{-1/2} \tag{8.67}$$

And our horizon scale is

$$r_H = 2ct = \frac{c}{\sqrt{G\rho}}\sqrt{\frac{3}{8\pi}} < \lambda_J \tag{8.68}$$

We can do the same thing in a matter-dominated universe. This time, $a \propto t^{2/3}$ and $\dot{a} \propto \frac{2}{3}t^{-1/3}$, so

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4}{9t^2} = \frac{8\pi G\rho}{3} \quad \longrightarrow \quad t = (6\pi G\rho)^{-1/2} \tag{8.69}$$

And this gives a horizon scale

$$r_H = 3ct = \frac{c}{\sqrt{G\rho}}\sqrt{\frac{3}{2\pi}} < \lambda_J \tag{8.70}$$

We can see in **both cases** $r_H < \lambda_J$, so theoretically pressure should be able to keep the overdensity from collapsing for the radiation-dominated and matter-dominated universe. However, if we look at the very earliest stages in the history of the universe, the overdensities actually **are collapsing**. This happens because the universe literally has not existed long enough for sound waves to be able to cross the whole overdensity, so the effects of pressure (which propagate through sound waves) cannot counteract the effects of the overdensity's own gravity.

Eventually, at some redshift between $3000 > z > 1100$ (i.e. after the universe becomes matter-dominated, but before recombination), the first sound waves will have propagated across the overdensity. We call this event **entering the sound horizon** because the overdensity suddenly feels the effects of the first pressure wave, which causes it to rebound and start expanding for a bit (since, as we proved, if the overdensity could actually feel the pressure, it should be stable against collapse). However, since this is only one pressure wave, the effect quickly dissipates and gravity takes over again, causing the overdensity to collapse until the second sound wave crosses the horizon some time later. This process repeats many times, creating what are known as **Baryon Acoustic Oscillations (BAO)** (though this term most often specifically refers to the *first* oscillation peak). See this represented visually in the overdensity over time in Figure 8.7. These oscillations are what cause the acoustic peaks in the CMB power spectrum (see §8.Q129).

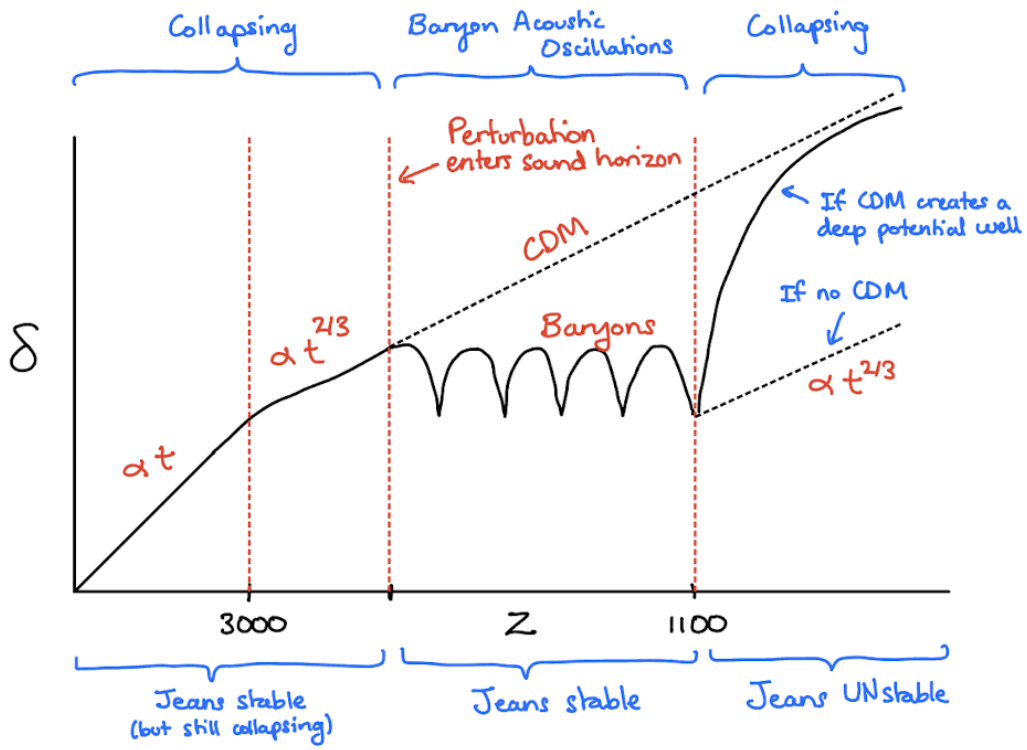


Figure 8.7: The evolution of an overdensity δ over redshift for both the baryonic matter and dark matter.

Note that **dark matter** is not coupled at all to photons, so it does not experience acoustic oscillations like baryonic matter does. In fact, it has no pressure to counteract gravity, so it will simply be collapsing from the beginning. However, it is possible for the dark matter to stagnate if it enters the horizon while the universe is still radiation dominated ($\rho_r \gg \rho_m$), since in this case the timescale for expansion t_{Hubble} is much shorter than the timescale for collapse t_{Jeans}:

$$t_{\text{Hubble}} \sim \frac{1}{\sqrt{G\rho_r}} \quad , \quad t_{\text{Jeans}} \sim \frac{1}{\sqrt{G\rho_m}} \tag{8.71}$$

However, it will quickly begin collapsing again once the universe becomes matter dominated.

Later still, we reach the epoch of recombination at $z = 1100$. At this point, the strong coupling between photons and electrons vanishes because electrons combine with protons to form atoms. Therefore, we lose pressure support from the photons and our pressure drops rapidly to that of an ideal gas

$$c_s = \sqrt{\frac{kT}{m}} \quad , \quad T = 2.7(1+z)\,\text{K} \quad , \quad P = \frac{\rho}{\mu m_p} kT \tag{8.72}$$

This causes the Jeans length to drop rapidly as well, meaning our overdensity will resume collapsing at this point. However, it collapses faster than usual because the baryons fall into the CDM well that has accumulated while the baryons were undergoing oscillations. We can actually use this as another test of dark matter![119,127]

132. What is the "spherical top hat" model for the formation of galaxies and clusters? What is the significance of the number $18\pi^2$?

This model assumes that an overdensity in the early universe can be described by a perfectly spherical region with a sharp edge that has a higher density than the surrounding environment (i.e. Figure 8.8).

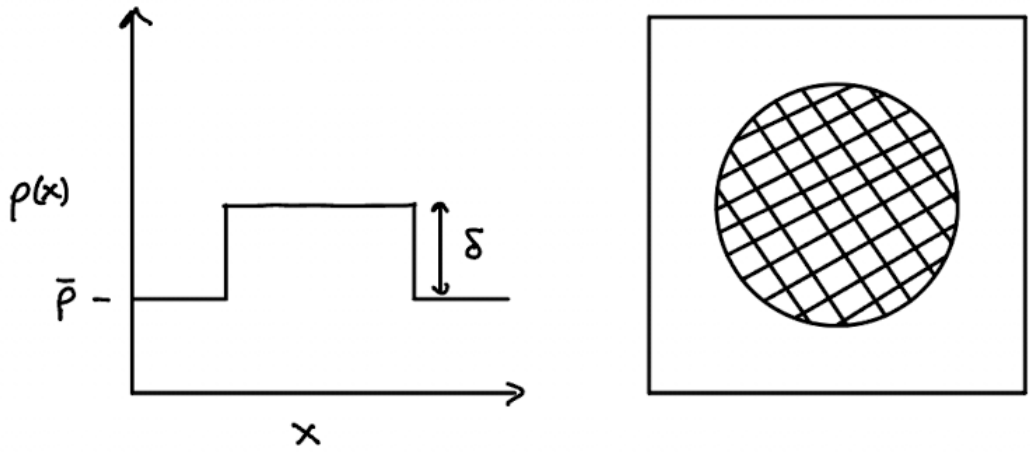


Figure 8.8: The spherical top hat model for an overdensity in the early universe. The left plot shows what the density would look like along one axis, while the right plot shows the spherical overdensity in space.

How does the overdensity $\delta(\mathbf{x}) = (\rho(\mathbf{x}) - \bar{\rho})/\bar{\rho}$ evolve with the expansion of the universe? To find out, we can use **Birkhoff's Theorem**, which states that the dynamics of a uniformly expanding, self-gravitating sphere are equivalent to a section of the universe as a whole. Therefore, all we have to do is apply the **Friedmann equations** (see §8.Q137) to this little spherical region.

First, we'll take the background "unperturbed" density $\bar{\rho} = \rho_u$ to be the critical density of the universe:

$$H^2(t) = \frac{8\pi G\rho_u(t)}{3} \quad \longrightarrow \quad \rho_u(t) = \frac{3H^2(t)}{8\pi G} \tag{8.73}$$

Then, for the spherical "perturbed" region, we have a density that's higher than the critical density, so the density parameter $\Omega > 1$, i.e. we can think of this as a little mini closed universe! This requires the curvature to be positive, so $K = +1$ in the Friedmann equation, giving us

$$H^2(t) = \frac{8\pi G\rho_p(t)}{3} - \frac{c^2}{a^2(t)} \quad \longrightarrow \quad \rho_p(t) = \frac{3}{8\pi G}\left(H^2(t) + \frac{c^2}{a^2(t)}\right) \tag{8.74}$$

We can solve for ρ in both cases and plug them in to find δ:

$$\delta(t) = \frac{\rho_p(t) - \rho_u(t)}{\rho_u(t)} = \frac{1}{\rho_u(t)}\frac{3c^2}{8\pi G a^2(t)} = \frac{c^2}{a^2(t)H^2(t)} \tag{8.75}$$

At early times when the universe is radiation dominated, $H^2 \propto a^{-4}$ so $\delta \propto a^2$. Later in the matter dominated regime we have $H^2 \propto a^{-3}$ and $\delta \propto a$.

We see that the density perturbation gets more significant over time. What is the turnaround time? That is, what is the time when the spherical perturbation stops getting bigger (due to the Hubble flow) and starts collapsing? This will happen when the sphere has its maximum radius, $r_{\max}$. We can use energy conservation to equate the energy at this time, which is purely gravitational potential energy, to the energy some time later when $r < r_{\max}$:

$$\mathscr{E} = -\frac{GM}{r_{\max}} = \frac{1}{2}\dot{r}^2 - \frac{GM}{r} \quad \longrightarrow \quad \dot{r} = \sqrt{\frac{2GM}{r}}\left(1 - \frac{r}{r_{\max}}\right)^{1/2} \tag{8.76}$$

Now for no apparent reason, let's make a trigonometric substitution

$$\frac{r}{r_{\max}} = \sin^2\theta \tag{8.77}$$

$$\frac{\mathrm{d}(r/r_{\max})}{\mathrm{d}\theta} = 2\sin\theta\cos\theta \quad \longrightarrow \quad \frac{\mathrm{d}r}{r_{\max}} = 2\sin\theta\cos\theta\,\mathrm{d}\theta \tag{8.78}$$

Then our equation for $\dot{r}$ becomes

$$2\sin\theta\cos\theta\,\mathrm{d}\theta = \sqrt{\frac{2GM}{r r_{\max}^2}}\sqrt{1-\sin^2\theta}\,\mathrm{d}t \tag{8.79}$$

$$= \sqrt{\frac{2GM}{r_{\max}^3\sin^2\theta}}\cos\theta\,\mathrm{d}t \tag{8.80}$$

$$= \sqrt{\frac{2GM}{r_{\max}^3}}\cot\theta\,\mathrm{d}t \tag{8.81}$$

Rearranging, we get

$$\int \mathrm{d}\theta\,\frac{2\sin\theta\cos\theta}{\cot\theta} = \sqrt{\frac{2GM}{r_{\max}^3}}\int \mathrm{d}t \tag{8.82}$$

$$2\int \mathrm{d}\theta\sin^2\theta = \sqrt{\frac{2GM}{r_{\max}^3}}\,t \tag{8.83}$$

Using the half-angle formula we get for the LHS

$$\int_0^\theta \mathrm{d}\theta'(1-\cos 2\theta') = \theta - \frac{1}{2}\sin 2\theta \tag{8.84}$$

Now if we replace $\eta = 2\theta$ and solve for t, we get

$$\boxed{t(\eta) = \frac{1}{2}\sqrt{\frac{3}{8\pi G\rho_{\max}}}(\eta - \sin\eta)} \tag{8.85}$$

$$\boxed{r(\eta) = \frac{1}{2}r_{\max}(1-\cos\eta)} \tag{8.86}$$

where the r equation just comes from the definition of our substitutions.

This is the equation of a cycloid! In other words, it is the path traced out by a point on the surface of a rolling circle (see Figure 8.9).

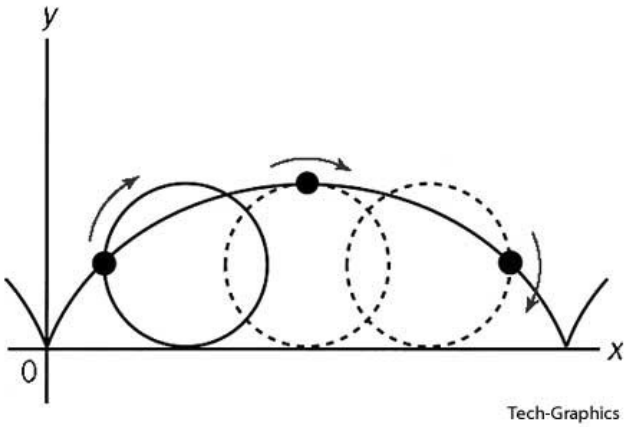


Figure 8.9: The path traced out by a point on a circle gives a cycloid[157].

It is easy to see from this that the maximum radius will be achieved when $\eta = \pi$, so we can solve the time equation to get the time when the spherical perturbation is at a maximum size:

$$t_{\max} = \frac{\pi}{2}\sqrt{\frac{3}{8\pi G\rho_{\max}}} \tag{8.87}$$

What do our densities look like at this time? First let's look at the unperturbed background density. We can use equation (8.146) to write ρ_u as a function of t:

$$\rho_u(t) = \frac{1}{6\pi G t^2} \quad \longrightarrow \quad \rho_{u,\max} = \frac{32}{18\pi^2}\rho_{\max} \quad \longrightarrow \quad \frac{\rho_{\max}}{\rho_{u,\max}} = \frac{9}{16}\pi^2 \tag{8.88}$$

In other words, at the turnover point the density is ∼5.6 times larger than the ambient density. After this point, the radius decays sinusoidally, so we have **nonlinear growth** of the density perturbation. Naïvely from our cycloid model we would expect $r \to 0$, but we know this doesn't happen because at some point galaxies **virialize**.

If we only care about the end point, we can just use the virial theorem in combination with energy conservation to get the density ratio at the end stage of evolution:

$$2T_f + U_f = 0 \ , \quad E = U_{\max} + T_{\max} = U_f + T_f \tag{8.89}$$

The "max" subscripts here denote the energies at the point when $r = r_{\max}$, not necessarily the point of maximum T or U! In fact, $T_{\max} = 0$ and $U_{\max} = -GM/r_{\max}$, so we have

$$T_f = -\frac{U_f}{2} \quad \longrightarrow \quad E = \frac{U_f}{2} = U_{\max} \tag{8.90}$$

Therefore

$$\frac{U_f}{U_{\max}} = 2 = \frac{GM}{r_f}\frac{r_{\max}}{GM} = \frac{r_{\max}}{r_f} \quad \longrightarrow \quad \frac{\rho_f}{\rho_{\max}} = \left(\frac{r_{\max}}{r_f}\right)^3 = 8 \tag{8.91}$$

where ρ_f is the density of the final stage of evolution. What we ultimately want here is a comparison between this final density ρ_f and the final density of the unperturbed background, $\rho_{u,f}$. So we need a relationship between $\rho_{u,f}$ and $\rho_{u,\max}$. This can be found relatively simply. Between the maximum of the cycloid and the end, we cover an angle $\eta = \pi \to 2\pi$, so our final time should be $t_f \sim 2t_{\max}$. Therefore, our $\rho_u \propto t^{-2}$ will change by a factor

$$\frac{\rho_{u,f}}{\rho_{u,\max}} = \frac{1}{4} \tag{8.92}$$

Finally, we have

$$\frac{\rho_f}{\rho_{u,f}} = \frac{\rho_f}{\rho_{\max}} \frac{\rho_{\max}}{\rho_{u,\max}} \frac{\rho_{u,\max}}{\rho_{u,f}} = 8 \times \frac{9}{16}\pi^2 \times 4 \tag{8.93}$$

$$\boxed{\frac{\rho_f}{\rho_{u,f}} = 18\pi^2} \tag{8.94}$$

This is the significance of the $18\pi^2$ figure in the spherical top-hat model. It describes the final density of a collapsed overdensity perturbation relative to the background density of the universe. In other words, galaxies in the universe today should have a characteristic overdensity factor of ~ 180. This is close enough to 200 that the astrophysics community has mostly settled on using 200 to calculate quantities like the "virial radius", which is the radius of a galaxy that encloses a predefined overdensity factor[119].

133. Summarize the observational evidence in favor of the standard "Big Bang" model of the universe.

1. **Expansion of the Universe**

   We see all galaxies (outside of the local group) moving away from us very fast, in a way that correlates with their distance. See §8.Q134 on Hubble's law: $v = H_0 d$. The only explanation that makes sense is that the universe is expanding. Then, if we extrapolate backward in time, we infer that the universe must have been much denser in the past, eventually all culminating in a single point. The initial expansion of this point is called the **Big Bang**.

2. **The Cosmic Microwave Background (CMB)**

   The Big Bang hypothesis implies that the universe used to be much denser, and thus much hotter. If we go back in time far enough, the ambient temperature should reach a point that prevents electrons and atomic nuclei from being bound to each other. During this time, the universe is "opaque" because photons are constantly scattering off of free electrons and cannot travel very far. The photons are also essentially a perfect blackbody at the ambient temperature of $T \sim 3000$ K.

   At the point of transition when the universe becomes cool enough for electrons to combine with nuclei, the photons suddenly can travel much further without scattering, eventually reaching us today. The last time a photon scatters before reaching us is pretty much the same distance from us in all directions, so this makes up a surface on the sky that we call the **surface of last scattering**. Since redshift expands the wavelengths like $1+z$ and temperature $T \propto \nu \propto 1/\lambda$, $T \propto (1+z)^{-1}$, so the temperature of this radiation today should be $T \sim 3000/1101 \sim 2.7$ K where $z = 1100$ is the redshift of the surface of last scattering (i.e. recombination).

   In fact, we do observe a nearly perfect blackbody at a temperature of ~ 2.7 K coming from all directions in the sky uniformly. This is called the **Cosmic Microwave Background (CMB)** radiation and is one of the strongest lines of evidence we have to support the Big Bang model. For more info on the CMB, see question §8.Q127.

3. **Big Bang Nucleosynthesis (BBN)**

   We can extrapolate back even further than the CMB epoch and surmise that the ambient temperature of the universe must have been hot enough to fuse Hydrogen into Helium (He), Deuterium (D), and Lithium (Li). This process is called **Big Bang Nucleosynthesis (BBN)**. By constraining how long the universe was in this phase, we can use nuclear physics to get estimates of the abundances of D, He, and Li relative to H.

   The measured abundances of these elements in the present day universe all match up pretty well to the expected estimates, lending further creedence to the Big Bang model. See §8.Q146 for more info on BBN.

4. **Galactic Evolution and Structure Formation**

   Galaxies that we observe further away are further back in time due to the finite speed of light. According to the Big Bang model then, these galaxies would have to be in earlier formation stages than the galaxies we see in the present-day universe, since they've only had a finite time to form since the Big Bang. We see evidence of these earlier formation stages in a number of ways: the furthest galaxies tend to be smaller and less massive (implying they have had less time to merge with other galaxies and grow larger) and have lower metallicities (implying their stellar populations have not reprocessed as much H and He into heavier elements).

134. What is the Hubble constant? Explain in detail at least three independent methods to measure it. How does it relate to the age of the universe?

What is the Hubble constant?[11,119]

In 1929, Edwin Hubble was measuring distances to Cepheid variables in other galaxies and noticed a correlation between these distances and the recessional velocities measured from the redshifts of these galaxies. This relationship is known as **Hubble's Law**:

$$\boxed{v = H_0 d} \tag{8.95}$$

where H_0 is the **Hubble constant**. The plot from Hubble's original paper is shown in Figure 8.10.

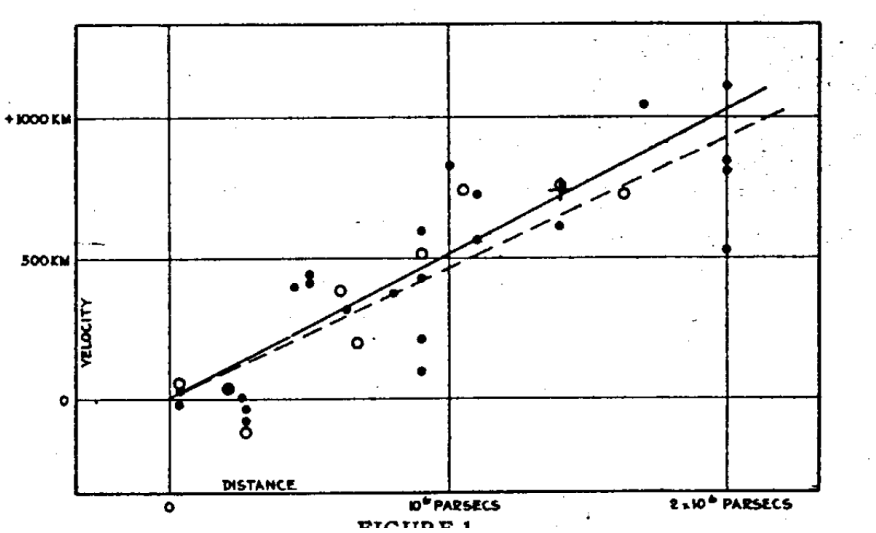


Figure 8.10: The relationship between distance and recessional velocity from Hubble's original paper[158].

It was noticed that the galaxies were not only moving away from us, but also from each other. In other words, this is not due to the intrinsic motions of the galaxies themselves, but the universe itself is expanding outwards in all directions, causing everything to move away from everything else. See the relationship between the Hubble constant and the expansion of the universe in Appendix section §12.12. In short, we get H in terms of the scale factor $a(t)$ to be

$$H(a) = \frac{\dot{a}}{a} \tag{8.96}$$

Modern estimates place the current value at around $H_0 \simeq 70$ km s^{-1} Mpc^{-1}.

How does it relate to the age of the universe?[119]

If we are to believe the Hubble law, then the universe is expanding. Then, if we go backward in time, we should eventually reach a point where the universe was all concentrated at a single point. If we assume the universe has always been expanding at the same rate that it is today, then this timescale is just the inverse of the Hubble constant! (Notice that the units of H_0 are just s^{-1} if we cancel out

the km with the Mpc). We call this the **Hubble time**, which gives an age estimate of

$$v t_H = d \quad \longrightarrow \quad t_H = \frac{1}{H_0} = 13.97\,\text{Gyr} \tag{8.97}$$

This is actually stupidly close to the true age of the universe considering the fact that we completely ignored how the expansion rate changes over time.

Ways to measure H_0

1. **Standard Candles**

   The most obvious way, and the way Hubble used in his original paper—we have a set of objects where we can independently measure the distance and the recessional velocity. The velocities are easy and can be obtained directly from the redshift ($\Delta\lambda/\lambda \simeq \Delta v/c$). The distances are more difficult. We can use **standard candles**, which are objects where the intrinsic luminosity is known from physics, so we can get the distance by comparing it to the observed flux ($L = 4\pi d^2 F$). Some examples of standard candles are:

   - **Cepheid variables**—relationship between intrinsic brightness and the pulsation period.
   - **RR Lyrae variables**—like Cepheids.
   - **Type Ia supernovae**—are thought to all have the same luminosity because they involve a white dwarf surpassing the Chandrasekhar mass, which is the same for all white dwarfs.
   - **Tip of the Red Giant Branch (TRGB)**—the brightest stars on the red giant branch undergo a Helium flash which has a known intrinsic luminosity (since the triple alpha process is *extremely* temperature sensitive, all stars that reach this point have an almost identical core temperature/pressure/density, and thus an almost identical helium core mass of $\sim$0.5 $M_\odot$).

2. **Tully-Fisher**

   See §7.Q110. In brief, the Tully-Fisher relationship can be used to get the intrinsic luminosity of a spiral galaxy, which can then be compared to its observed flux to obtain a distance. Then we just compare the distance to the recessional velocity as with the previous method.

3. **Gravitational Lensing**

   See §6.Q96. For strongly lensed quasars, time delays between two different images of the same object can be on measurable timescales (like seconds). The time delays occur because light travels along two paths of different lengths for the different images. The amount of time delay depends on the the distances from us to the lens, us to the source, and the lens to the source, thus we can use this method to obtain a distance measure independent of redshift. These measurements are primarily done by the H0LiCOW (H_0 Lenses in COSMOGRAIL's Wellspring) team[160]. See an example of these strongly lensed quasars in Figure 8.12.

4. **Standard sirens**

   Like standard candles, but with gravitational waves. As it turns out, *every* gravitational wave signal from the merger of two compact objects is a standard siren! From a gravitational wave signal of a binary merger, we can get a direct measurement of the chirp mass $\mathscr{M}$, i.e.

$$\mathscr{M} = \frac{c^3}{G}\left[\frac{5}{96}\pi^{-8/3} v^{-11/3} \dot{v}\right]^{3/5} \tag{8.98}$$

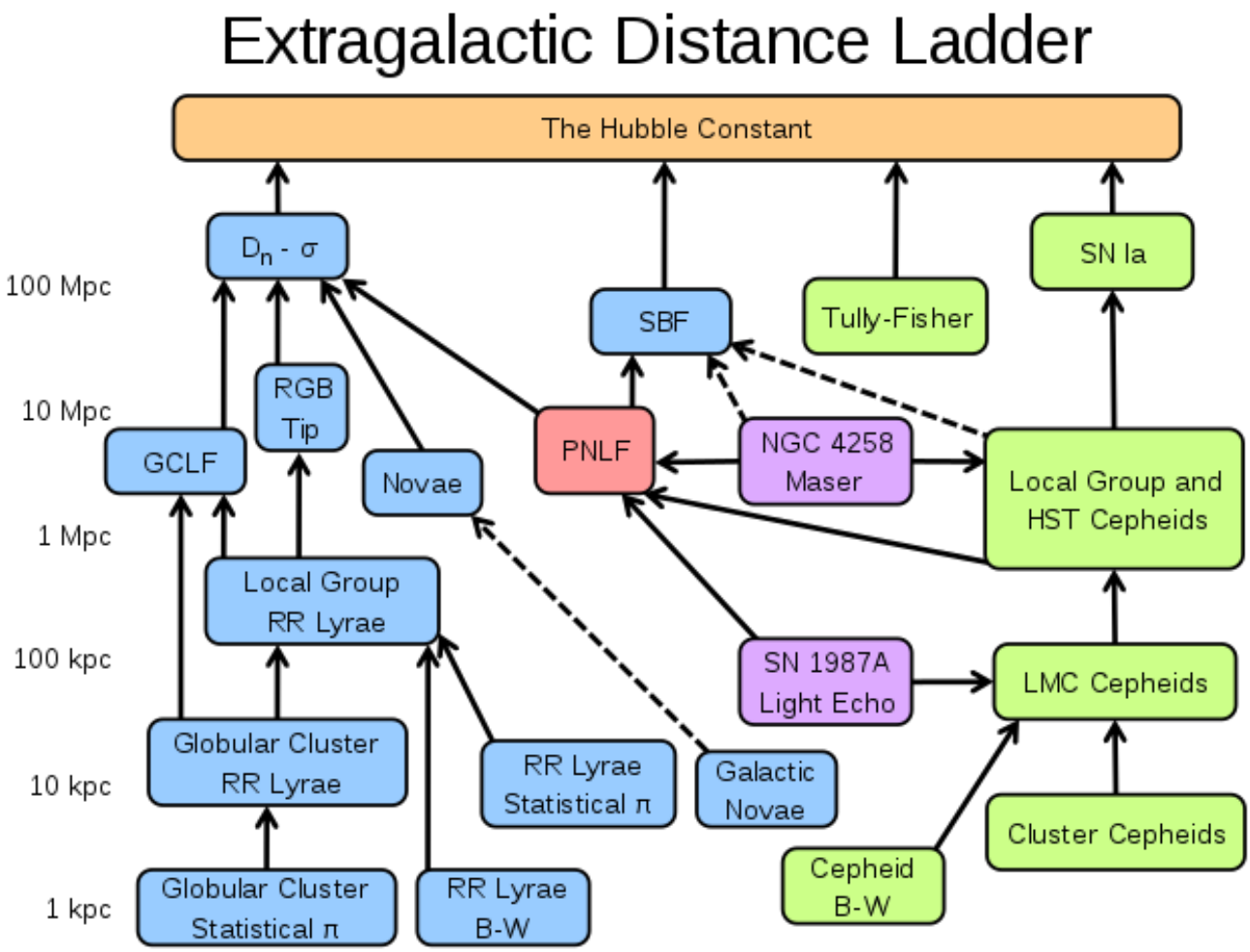


Figure 8.11: The distance ladder. See Ref. 159.

And the intrinsic luminosity of the signal is directly related to the chirp mass:

$$L = \frac{32G^{7/3}}{5c^5}(\pi \nu \mathscr{M})^{10/3} \tag{8.99}$$

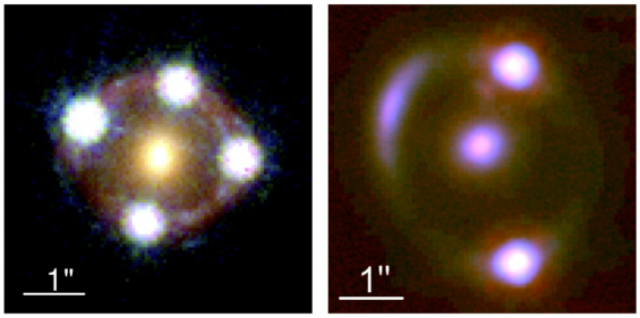
Gravitational waves are attenuated over distance in the same way that light waves are, so the observed power will be smaller by a factor of the luminosity distance d_L, which is related to the Hubble constant:

$$F = \frac{L}{4\pi d_L^2} \quad , \quad d_L = (1+z)\frac{c}{H_0}\int_0^z \frac{dz'}{E(z')} \tag{8.100}$$

(where, for simplicity, I took the expression for d_L assuming a universe with no curvature; see Ref. 161). This method was partially developed by Scott Hughes (MIT/MKI)!

5. **BAO Feature in the 2-Point Correlation Function**

   See §7.Q122. In brief, we can measure the angular diameter distance as a function of redshift by comparing the observed angular size of the BAO in galaxy 2-point correlation functions to



(c) HE 0435−1223 (d) SDSS 1206+4332

Figure 8.12: Two strongly lensed quasars with multiple images from the H0LiCOW sample[121].

the intrinsic size, which is constrained by physics (see §8.Q129). Then, $d_A = r/\theta$, which we can relate to the expression for d_A based on H_0 and the cosmological density parameters. The BAO shows up as a small peak in the 2-point correlation function because these density perturbations are "frozen in" at recombination and collapse into galaxies, retaining their separations.

6. **The CMB Power Spectrum**

   See §8.Q129. In brief, we can use the locations and amplitudes of the acoustic peaks in the CMB power spectrum to constrain cosmological parameters, including H_0, since these affect how overdensity perturbations in the early universe evolve.

There is a discrepancy between the values of the Hubble constant obtained using "early universe" measurements (5 and 6 above) and "late universe" measurements (1–4 above). This is called the **Hubble tension** (see Figure 8.13). In general, the late universe methods tend to converge on a value of $\sim 73 \pm 2$ km s^{-1} while the early universe methods get a value of $\sim 67 \pm 0.5$ km s^{-1}. This is a tension of $> 5\sigma$! The cause of this discrepancy is still unknown and is one of the biggest problems in modern astrophysics.

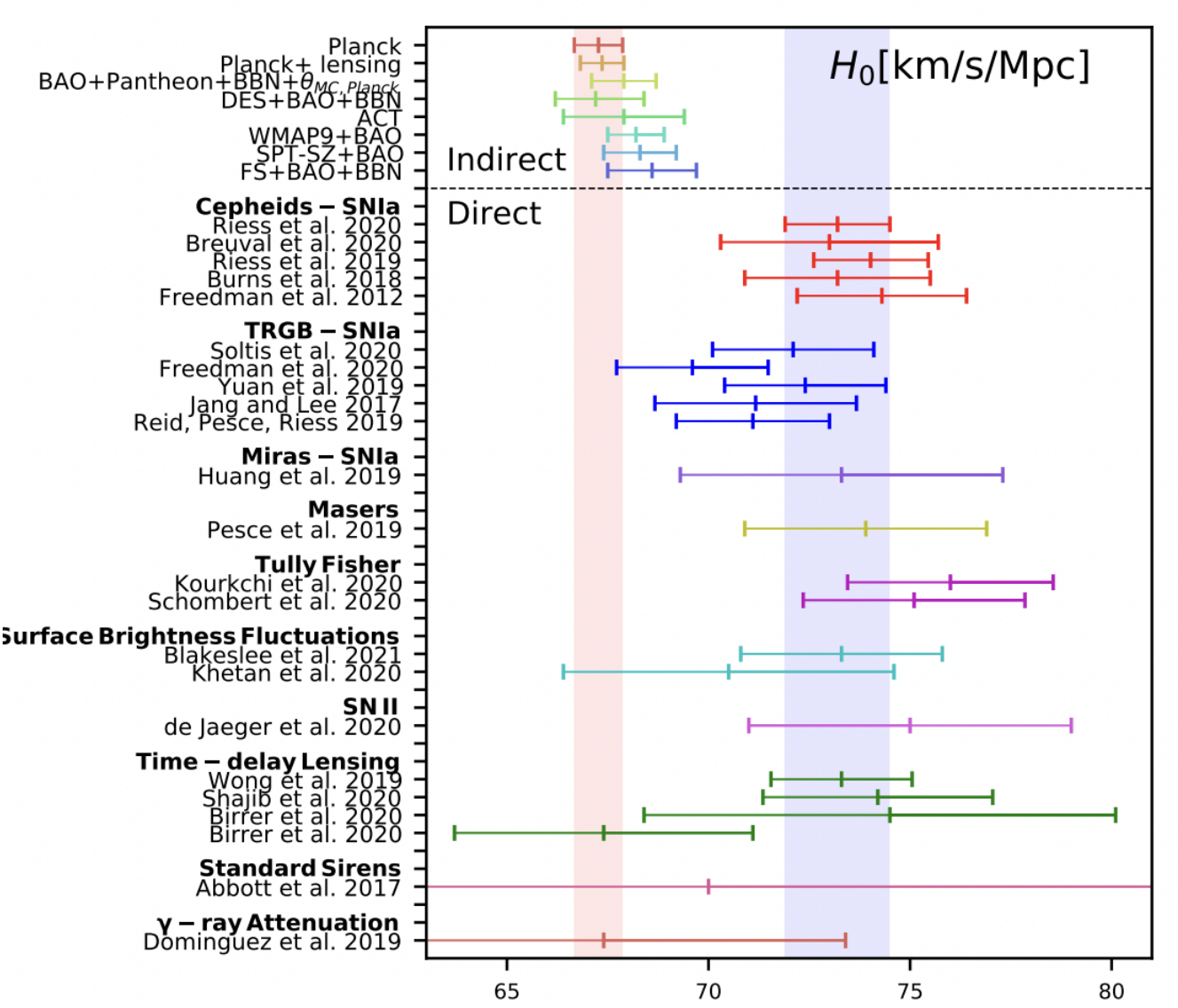
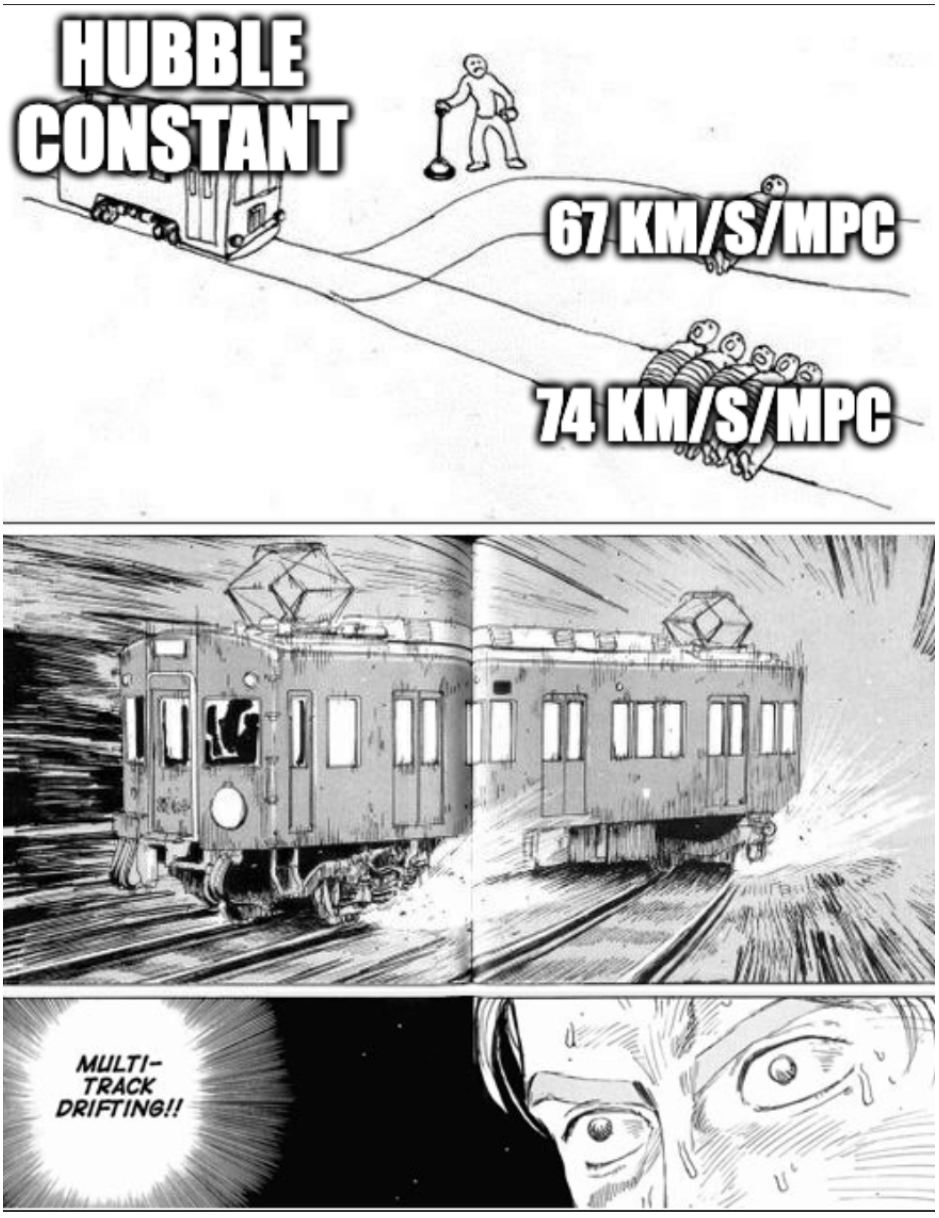


Figure 8.13: A chart showing different measurements of the Hubble constant using different techniques[162].


HUBBLE CONSTANT
67 KM/S/MPC
74 KM/S/MPC
MULTI-TRACK DRIFTING!!

135. How can the age of the universe be estimated empirically? Do the current estimates agree with that obtained from the Hubble constant?

The simplest $\mathcal{O}(1)$ estimate of the age of the universe from the Hubble constant is $t_H \sim 1/H_0$, which gives $\sim 13.97\,\text{Gyr}$ using a Hubble constant value of $H_0 = 70\,\text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1}$. This is often referred to as the **Hubble time**.

There are a few ways to measure this empirically. We can get estimates for the ages of objects in the universe and reason that the universe must be at least this old, so these methods all put lower bounds on the age of the universe.

Globular Cluster Ages[22]

By measuring the turn-off point from the main sequence in a globular cluster, its age can be estimated because the hotter stars burn up their fuel and evolve into red giants faster than the cooler stars. For more information on this, see §3.Q17.

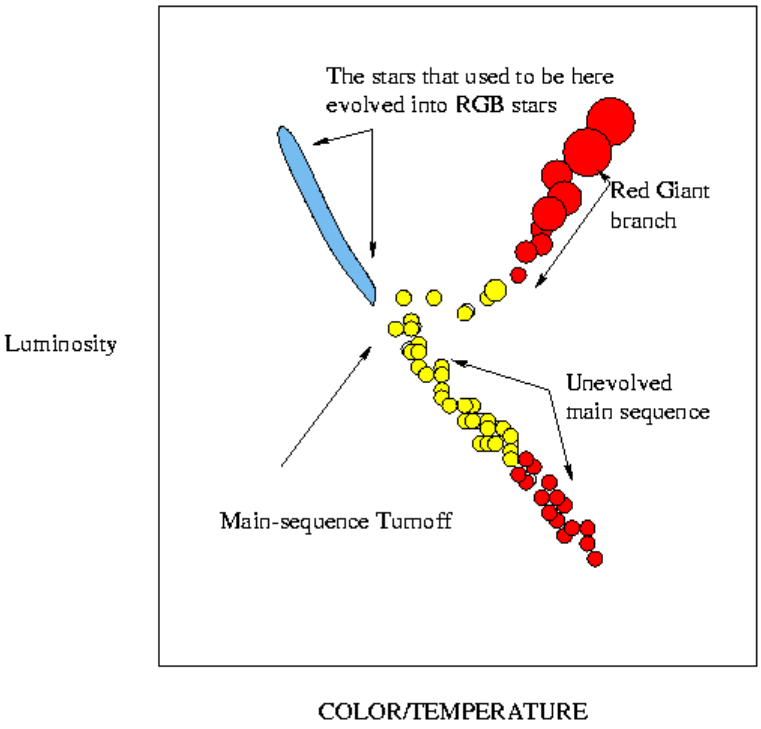


Figure 8.14: Example of measuring a globular cluster's age from the main sequence turn-off point[163].

The ages estimated with this method range from $\sim 12 - 13$ Gyr, roughly consistent with the Hubble time.

White Dwarf Cooling[22]

White dwarfs cool over time through radiation—this is a process that we understand and can model fairly well (luminosity scales with core temperature as $L \propto T_c^{7/2}$). If star formation extended back infinitely in time, this would have no effect on the observed luminosity function of white dwarfs. However, since the universe does have a finite age, we can observe a cutoff in the luminosity function of white dwarfs that corresponds to how much the lowest mass white dwarfs have been able to cool over the history of the universe (see Figure 8.15).

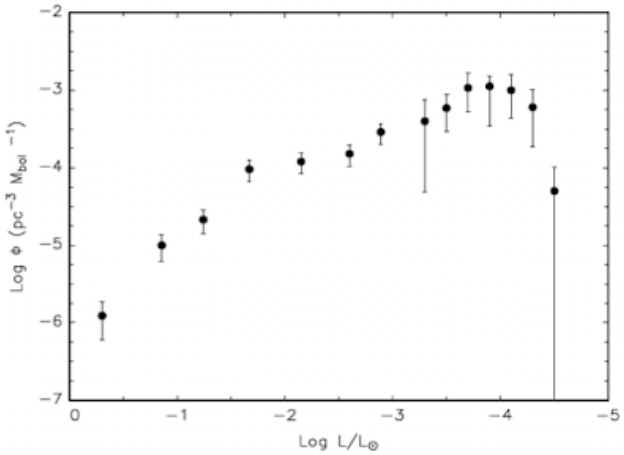


Figure 8.15: The white dwarf luminosity function in the Galactic disk[22].

This cutoff occurs at about $L \sim 10^{-4.4} L_\odot$, corresponding to $T \sim 1000$ K (compared to initial temperatures of order 10^5 K), from which we can get an age estimate of $\sim 12-13$ Gyr, which again agrees with the Hubble time.

The MW Stellar Halo[22]

The stellar halo of the MW is thought to contain individual stars (not in clusters) that are also old. However, this age determination is more difficult than with globular clusters because these halo stars are not a single population—they come from a history of mergers.

136. How would you compute the age of our universe using the Friedmann equation? What approximations can be made to simply this calculation?

Using the observer's version of the Friedmann equation, written in terms of the scale factor a, we have

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2(\Omega_m a^{-3} + \Omega_r a^{-4} + \Omega_K a^{-2} + \Omega_\Lambda) \tag{8.101}$$

See the derivation in §8.Q137. This is a separable differential equation, so we can solve for the age at any given a by rearranging and integrating over time

$$\int_0^t \mathrm{d}t' = \frac{1}{H_0}\int_0^a \frac{\mathrm{d}a'}{\sqrt{\Omega_m(a')^{-1} + \Omega_r(a')^{-2} + \Omega_K + \Omega_\Lambda(a')^2}} \tag{8.102}$$

I put this integral into Mathematica and I think I saw its soul die a little. So let's try to make this a little bit easier on ourselves.

Right away we can ignore the radiation density Ω_r and the curvature Ω_K because their values are negligible. At this point we could technically solve the integral analytically. If we did so, we would obtain

$$t \simeq \frac{1}{H_0}\int_0^1 \frac{\mathrm{d}a}{\sqrt{\Omega_m a^{-1} + \Omega_\Lambda a^2}} = \frac{2}{3H_0\sqrt{\Omega_\Lambda}}\operatorname{arcsinh}\left(\sqrt{\frac{\Omega_\Lambda}{\Omega_m}}\right) = 13.8\,\mathrm{Gyr} \tag{8.103}$$

But I don't like integrals that require trig substitutions because my tiny pea brain can't handle it. So let's make one more approximation: the universe is matter-dominated for most of its history since $\Omega_m a^{-3} \gg \Omega_\Lambda$ for $a < 1$. Therefore, let's just ignore the Ω_Λ term too (which requires us to set $\Omega_m = 1$). In this case, we just recover the results for a matter-dominated universe from §8.Q137

$$t \simeq \frac{1}{H_0\sqrt{\Omega_m}}\int_0^1 a^{1/2}\mathrm{d}a = \frac{2}{3H_0\sqrt{\Omega_m}}a^{3/2}\bigg|_0^1 \simeq \frac{2}{3}\frac{1}{H(a)} \tag{8.104}$$

All we have to do is plug in $a = 1$ to get the age today, which gives an estimate of ~ 9.3 Gyr. This is a bit of an underestimate from the typically accepted value of 13.8 Gyr because the universe at the current epoch is actually dominated by Ω_Λ, which drives additional expansion and causes $H(a)$ to be larger than expected for a purely matter-dominated universe[119].

137. What is the Robertson-Walker metric? What are the Friedmann equations? How does the universe expand if it is (i) radiation dominated? (ii) matter dominated with $\Omega \ll 1$, $\Omega = 1$, $\Omega \gg 1$? (iii) dominated by a cosmological constant?

The **Friedmann-Lemaître-Robertson-Walker (FLRW) Metric** is given by[11,119,127]

$$\boxed{\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2\right]} \tag{8.105}$$

For a full derivation of this, and a full walkthrough of setting up different cosmological distances in an expanding universe, see the Appendix §12.12.

The Friedmann Equations[11,119]:

Now we're going to derive Friedmann's equation using a purely classical argument. To do this, we can use **Birkhoff's Theorem** again (see §8.Q132). Then, for simplicity let's take the radius of a spherical region in the universe to have a comoving distance of $\mathbf{r} = 1$ such that $R(t) = a(t)$. Then, if the density of this sphere is ρ, the mass enclosed in the sphere is

$$M(<a) = \frac{4}{3}\pi a^3\rho \tag{8.106}$$

The self-gravitational force this sphere experiences is

$$\frac{F}{m} = \ddot{a} = -\frac{GM(<a)}{a^2} = -\frac{4}{3}G\pi a\rho \tag{8.107}$$

Let's multiply by $\dot{a}$:

$$\dot{a}\ddot{a} = -\frac{4}{3}G\pi\rho a\dot{a} \tag{8.108}$$

Now we can use conservation of mass $\rho a^3 = \rho_0 a_0^3$ to obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\dot{a}^2\right) = -\frac{4}{3}\pi G\rho_0\frac{a_0^3}{a^2}\dot{a} = -\frac{4}{3}\pi G\rho_0 a_0^3\frac{\mathrm{d}}{\mathrm{d}t}\left(-\frac{1}{a}\right) \tag{8.109}$$

Integrating over time and calling our constant of integration $-Kc^2$

$$\frac{1}{2}\dot{a}^2 = \frac{4}{3}\pi G\rho_0\frac{a_0^3}{a} - Kc^2 = \frac{4}{3}\pi G\rho a^2 - Kc^2 \quad\longrightarrow\quad \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3} - \frac{Kc^2}{a^2} \tag{8.110}$$

This gets us to the first Friedmann equation if the cosmological constant $\Lambda = 0$. If we did a full proper derivation with GR, we would've obtained

$$\boxed{\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3} - \frac{Kc^2}{a^2} + \frac{\Lambda c^2}{3}} \tag{8.111}$$

$$\boxed{\left(\frac{\ddot{a}}{a}\right) = -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right) + \frac{\Lambda c^2}{3}} \tag{8.112}$$

These are the **Friedmann Equations** in all of their glory. Let's quickly go over how one *would* go about this derivation using GR, without going through any of the actual algebra:

1. Assume the universe is homogeneous and isotropic on the largest (spatial) scales, such that the FLRW metric can be taken as our metric of spacetime.
2. Construct the Einstein curvature tensor $G_{\mu\nu}$ from the FLRW metric and its (second-order) derivatives.
3. Assume that our source term can be represented by a perfect fluid, $T_{\mu\nu} = (\rho + P/c^2)u_\mu u_\nu + P g_{\mu\nu}$, or $T^\mu_\nu = \text{diag}(-\rho c^2, P, P, P)$.
4. Solve the Einstein field equations, $G_{\mu\nu} + \Lambda g_{\mu\nu} = (8\pi G/c^4)T_{\mu\nu}$.

For the sake of my observer pals out there, let's reparametrize this in terms of things that are measurable. First, we define the **expansion rate**

$$H(t) \equiv \frac{\dot{a}(t)}{a(t)} \quad , \quad H_0 \equiv H(t_0) \tag{8.113}$$

Next, let's look at the density ρ. Normally we would think this is just due to matter, but this actually includes the density from the energy of photons as well (i.e. $\rho = \mathscr{E}/c^2$ where $\mathscr{E}$ is the energy density of the universe). So, we can break this up into $\rho = \rho_m + \rho_r$ where ρ_m is the contribution from matter and ρ_r is the contribution from photons. We can also extend this concept to the curvature and the cosmological constant, defining their "densities" to be

$$\rho_K(t) \equiv -\frac{3}{8\pi G}\frac{Kc^2}{a^2(t)} \quad , \quad \rho_\Lambda \equiv \frac{\Lambda c^2}{8\pi G} \tag{8.114}$$

Notice that $\rho_\Lambda = \text{const}$. This allows us to write the first Friedmann equation as

$$H^2(t) = \frac{8\pi G}{3}(\rho_m + \rho_r + \rho_K + \rho_\Lambda) \tag{8.115}$$

Then we can define the **critical density** ρ_{crit} as the density needed to make the universe closed (or flat). By "making the universe closed" we mean making $K = 0$. We can see that this must happen if we have

$$\rho_{\text{crit}}(t) \equiv \frac{3H^2(t)}{8\pi G} \quad , \quad \rho_{\text{crit},0} \equiv \rho_{\text{crit}}(t_0) = \frac{3H_0^2}{8\pi G} \tag{8.116}$$

since this would require the contribution from the ρ_K term to go to 0. Then, we can define the cosmological **density parameters** relative to this critical density:

$$\Omega_i(t) \equiv \frac{\rho_i(t)}{\rho_{\text{crit}}(t)} = \frac{8\pi G}{3H^2(t)}\rho_i(t) \tag{8.117}$$

And the value today is

$$\Omega_{i,0} \equiv \Omega_i(t_0) = \frac{8\pi G}{3H_0^2}\rho_{i,0} \tag{8.118}$$

Now our goal is to write everything in terms of their **present-day** values, since these are easily

observable. This gives density parameters of

$$\Omega_{m,0} = \frac{\rho_{m,0}}{\rho_{\text{crit},0}} = \frac{8\pi G}{3H_0^2}\rho_{m,0} \tag{8.119}$$

$$\Omega_{r,0} = \frac{\rho_{r,0}}{\rho_{\text{crit},0}} = \frac{8\pi G}{3H_0^2}\rho_{r,0} \tag{8.120}$$

$$\Omega_{K,0} = \frac{\rho_{K,0}}{\rho_{\text{crit},0}} = -\frac{Kc^2}{H_0^2} \tag{8.121}$$

$$\Omega_{\Lambda,0} = \frac{\rho_{\Lambda,0}}{\rho_{\text{crit},0}} = \frac{\Lambda c^2}{3H_0^2} \tag{8.122}$$

We need to know how each of them evolves with respect to the scale parameter $a(t)$. To do that, we need an additional constraint: an **equation of state** that relates the pressure and density of our perfect fluid. A general form of this equation of state is taken as

$$\boxed{P_i = w_i \rho_i c^2} \tag{8.123}$$

where w_i are constants. Let's assume that our fluid is at a constant temperature, so there is no heat transfer, then by the first law of thermodynamics

$$\frac{\mathrm{d}U}{\mathrm{d}t} = -\frac{\mathrm{d}W}{\mathrm{d}t} = -P\frac{\mathrm{d}V}{\mathrm{d}t} \tag{8.124}$$

Replacing the energy with the energy density, and using $V = (4/3)\pi R^3$, we get

$$\frac{\mathrm{d}(R^3 u)}{\mathrm{d}t} = -P\frac{\mathrm{d}(R^3)}{\mathrm{d}t} \tag{8.125}$$

Then we substitute the density $\rho c^2 = u$ and the scale factor $a\chi = R$ (where χ is a comoving distance, but it doesn't matter as it cancels)

$$\frac{\mathrm{d}(\rho a^3)}{\mathrm{d}t} = -\frac{P}{c^2}\frac{\mathrm{d}(a^3)}{\mathrm{d}t} \tag{8.126}$$

Plugging in our equation of state, we obtain

$$\frac{\dot{\rho}}{\rho} = -3(1+w)\frac{\dot{a}}{a} \tag{8.127}$$

Then, integrating

$$\boxed{\rho = \rho_0 a^{-3(1+w)}} \tag{8.128}$$

where I took $a_0 = a(\text{now}) = 1$. There are now three cases of interest to see how each of our density contributions scales with a (curvature is excluded, as it is a special case):

1. **Matter**: $P = 0$ for all ρ. In other words, on the largest spatial scales, matter acts like "dust". Therefore, $w = 0$ and $\rho_m \propto a^{-3}$. Conceptually, this makes sense: we have 3 spatial dimensions, so if matter is not created or destroyed, then as the universe expands, the density scales inversely to the volume a^3.

2. **Radiation**: $P_{\text{rad}} = \rho_r c^2/3$ (radiation pressure). Therefore, $w = 1/3$ and $\rho_r \propto a^{-4}$. Conceptually, radiation density decreases with a^{-3} just like matter, but there is an additional a^{-1} factor corresponding to the energy of individual photons being stretched out and decreasing due to the expansion. $E_\gamma = h\nu_{\text{emit}} = h\nu_{\text{obs}}/a$.
3. **Cosmological Constant**: We know the density here is constant, so the only possible solution is $w = -1$, i.e. $P = -\rho_\Lambda c^2$. Conceptually, this corresponds to some kind of vacuum energy density, which has a negative pressure that accelerates the expansion. See §8.Q140 for more details.

For the curvature, we have a^{-2} directly from the Friedmann equation. Therefore:

$$\rho_m(t) = \rho_{m,0}a^{-3}(t) = \Omega_{m,0}\rho_{\text{crit},0}a^{-3}(t) \tag{8.129}$$
$$\rho_r(t) = \rho_{r,0}a^{-4}(t) = \Omega_{r,0}\rho_{\text{crit},0}a^{-4}(t) \tag{8.130}$$
$$\rho_K(t) = \rho_{K,0}a^{-2}(t) = \Omega_{K,0}\rho_{\text{crit},0}a^{-2}(t) \tag{8.131}$$
$$\rho_\Lambda(t) = \Omega_{\Lambda,0}\rho_{\text{crit},0} = \text{const} \tag{8.132}$$

From now on I'm going to drop the 0 index and just assume all of the density parameters are measured at the current epoch. Putting this all back into the Friedmann equation, we recover

$$H^2(a) = \frac{8\pi G\rho_{\text{crit}}}{3}(\Omega_m a^{-3} + \Omega_r a^{-4} + \Omega_K a^{-2} + \Omega_\Lambda) \tag{8.133}$$

We notice that the prefactor is just H_0^2. We're almost done. There is one more thing to do: the scale parameter a is not an easy observable, but the redshift of photons z is. The relationship between the scale factor and the redshift is $1/a = 1+z$. For a derivation of this, see the Appendix §12.12. Now we can write the Friedmann equation fully in terms of observable quantities:

$$\boxed{H^2(z) = H_0^2[\Omega_m(1+z)^3 + \Omega_r(1+z)^4 + \Omega_K(1+z)^2 + \Omega_\Lambda]} \tag{8.134}$$

Oftentimes the part in brackets is given its own function:

$$E(z) \equiv \sqrt{\Omega_m(1+z)^3 + \Omega_r(1+z)^4 + \Omega_K(1+z)^2 + \Omega_\Lambda} \quad\longrightarrow\quad H(z) = H_0E(z) \tag{8.135}$$

Note that by the way we defined the density parameters, they must all sum to 1 (such that for $z = 0$, $H = H_0$). Often we take advantage of this by writing the *total* density parameter, which includes all of the density parameters *except for curvature* (since curvature is a "fake" density parameter):

$$\Omega \equiv \Omega_m + \Omega_r + \Omega_\Lambda \quad , \quad \Omega_K = 1 - \Omega \tag{8.136}$$

Recent values of these cosmological parameters from the WMAP mission are

$$\boxed{\Omega_m = 0.27 \pm 0.04} \tag{8.137}$$
$$\boxed{\Omega_\Lambda = 0.73 \pm 0.04} \tag{8.138}$$
$$\boxed{\Omega_r = 8.24 \times 10^{-5}} \tag{8.139}$$
$$\boxed{\Omega_b = 0.045 \pm 0.003} \tag{8.140}$$

Where Ω_b is the baryonic matter density.

How does the universe expand in different regimes?[119]

The observer's form of the Friedmann equation makes this painfully obvious since we explicitly separated out the dependences on the matter, radiation, curvature, and cosmological constant densities.

(i) Radiation Dominated:

Take Ω_r to be the only significant density parameter. Then,

$$H(a) = \frac{\dot{a}}{a} \simeq H_0\sqrt{\Omega_r a^{-4}} = H_0\Omega_r^{1/2}a^{-2} \quad \longrightarrow \quad \dot{a} = H_0\Omega_r^{1/2}a^{-1} \tag{8.141}$$

We integrate to obtain

$$\int \mathrm{d}a\, a = H_0\Omega_r^{1/2}\int \mathrm{d}t \quad \longrightarrow \quad \frac{1}{2}a^2 = H_0\Omega_r^{1/2}t \quad \longrightarrow \quad \boxed{a(t) = (2H_0\Omega_r^{1/2}t)^{1/2}} \tag{8.142}$$

We see that $a \propto t^{1/2}$, the expansion is slower than linear. The reverse relationship gives the age in a radiation-dominated regime

$$t = \frac{1}{2}\frac{1}{H_0\sqrt{\Omega_r a^{-4}}} \quad \longrightarrow \quad \boxed{t \simeq \frac{1}{2}\frac{1}{H(z)}} \tag{8.143}$$

(ii) Matter Dominated:

Take Ω_m and Ω_K to be the only significant density parameters. Then,

$$H(a) = \frac{\dot{a}}{a} \simeq H_0\sqrt{\Omega_m a^{-3} + \Omega_K a^{-2}} \quad \longrightarrow \quad \dot{a} = H_0\sqrt{\Omega_m a^{-1} + \Omega_K} \tag{8.144}$$

If $\Omega = 1$, then $\rho = \rho_{\text{crit}}$ and we can ignore the curvature term which simplifies things greatly

$$\int \mathrm{d}a\, a^{1/2} = H_0\Omega_m^{1/2}\int \mathrm{d}t \quad \longrightarrow \quad \frac{2}{3}a^{3/2} = H_0\Omega_m^{1/2}t \quad \longrightarrow \quad \boxed{a(t) = \left(\frac{3}{2}H_0\Omega_m^{1/2}t\right)^{2/3}} \tag{8.145}$$

We have $a \propto t^{2/3}$. The expansion is faster than the radiation regime, but still slower than linear. Once again the age can be estimated from

$$t = \frac{2}{3}\frac{1}{H_0\sqrt{\Omega_m a^{-3}}} \quad \longrightarrow \quad \boxed{t \simeq \frac{2}{3}\frac{1}{H(z)}} \tag{8.146}$$

In the other cases where $\Omega \neq 1$, we have

$$\int \frac{\mathrm{d}a}{\sqrt{\Omega_m a^{-1} + \Omega_K}} = H_0 t \tag{8.147}$$

If $\Omega \gg 1$, the universe must be closed, i.e. $\Omega_K = 1 - \Omega < 0$. This means the expansion will eventually slow down and reverse, causing the universe to collapse in on itself.

If $\Omega \ll 1$, we have the opposite situation, so the universe must be open, i.e. $\Omega_K = 1 - \Omega > 0$. In the

extreme case where Ω_K dominates over Ω_m, we have

$$\boxed{a(t) \simeq H_0 \sqrt{\Omega_K}\, t} \tag{8.148}$$

So $a \propto t$ and

$$\boxed{t \simeq \frac{1}{H(z)}} \tag{8.149}$$

(iii) Λ Dominated

Take Ω_Λ to be the only significant density parameter. Then,

$$H(a) = \frac{\dot{a}}{a} \simeq H_0 \sqrt{\Omega_\Lambda} \tag{8.150}$$

And integrate

$$\int \frac{\mathrm{d}a}{a} = H_0 \sqrt{\Omega_\Lambda} \int \mathrm{d}t \quad \longrightarrow \quad \ln \frac{a}{a_0} = H_0 \Omega_\Lambda^{1/2} (t - t_0) \quad \longrightarrow \quad \boxed{a(t) = a_0 \exp\left[H_0 \Omega_\Lambda^{1/2} (t - t_0)\right]} \tag{8.151}$$

We see that we have $a \propto \exp(t)$ exponential growth. Then the time relative to today is

$$\boxed{t - t_0 \simeq \frac{1}{H(z)} \ln\left(\frac{1}{1+z}\right)} \tag{8.152}$$

For a general view of how the universe expands over time for different density parameters, see Figure 8.16.

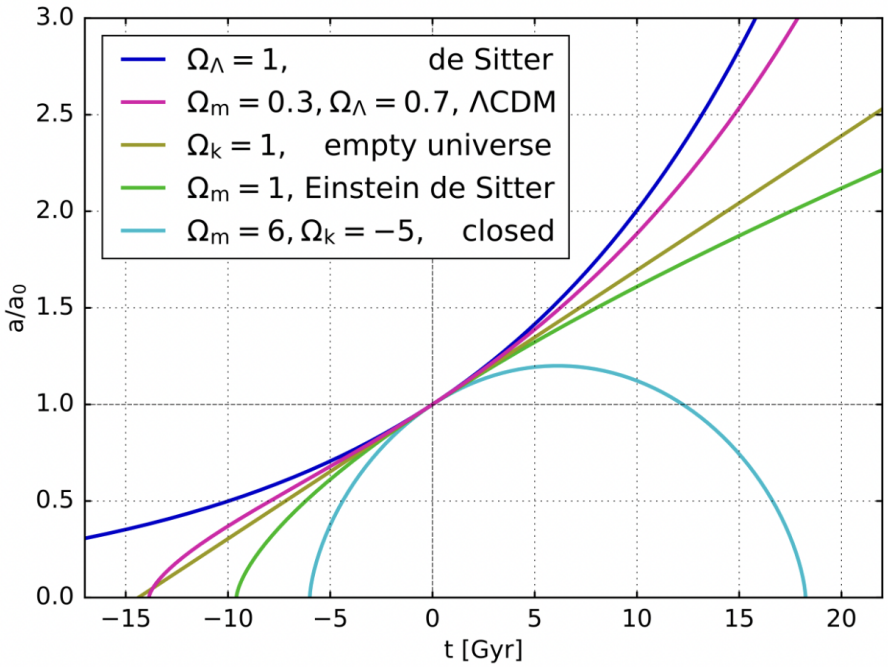


Figure 8.16: The expansion of the universe in different regimes[164].

138. What was the approximate temperature of the universe at recombination, and why did recombination happen then?

The early universe has a high ambient temperature, so much so that there are a lot of photons with energies $\gtrsim 13.6$ eV, the ionization potential of Hydrogen. Therefore, we just have a soup of free electrons and nuclei until the universe cools enough that the electrons can combine with the nuclei and form bound atoms. This process is called **recombination**, and we define the specific point of recombination to be when 50% of the electrons in the universe have been bound to nuclei[119].

Recombination is also the point we see CMB photons from, since this is the first time photons can free-stream without scattering off of electrons constantly. Therefore, the temperature of the universe at recombination should just be the temperature of the CMB—but not the temperature *today*, the temperature *at the epoch of recombination*. Essentially, we have

$$\rho_{\text{rad}} \propto a^{-4} \propto (1+z)^4 \quad \text{Cosmological equation of state + Radiation pressure} \tag{8.153}$$

$$\rho_{\text{rad}} \propto T^4 \quad \text{Stefan-Boltzmann Law + Radiation pressure} \tag{8.154}$$

$$\therefore\ T \propto (1+z) \tag{8.155}$$

We can find what the ambient temperature is at recombination using the **Saha equation** (§5.Q77)

$$\frac{n_p n_e}{n_H} = \frac{2g_p}{g_H}\left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} e^{-\chi/kT} \tag{8.156}$$

with $g_p = 2$, $g_H = 4$, $\chi = 13.6$ eV, and the condition for recombination which is $n_p = n_e = n_H = n_b/2$ where n_b is the total number density of baryons (recall that electrons are leptons, not baryons, so $n_b = n_p + n_H$):

$$\frac{1}{2}n_b = \left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} e^{-\chi/kT} \tag{8.157}$$

We can get n_b from the observed baryon density $\Omega_b = 0.045$ and parametrize T in terms of the redshift and observed CMB temperature $T = T_0(1+z)$ where $T_0 = 2.7$ K:

$$\frac{1}{2}\frac{\Omega_b \rho_{\text{crit}}}{m_p}(1+z)^3 = \left(\frac{2\pi m_e kT_0(1+z)}{h^2}\right)^{3/2} e^{-\chi/kT_0(1+z)} \tag{8.158}$$

Solving numerically gives a redshift of $z_{\text{recomb}} \simeq 1400$, but this is inaccurate for a couple of reasons:

- He will recombine earlier than H because it has a higher ionization energy
- We assume electrons immediately enter the ground state when they recombine with protons, whereas in reality they would start in an excited state before cascading down to the ground state. This has the effect of reducing the "effective" ionization potential from 13.6 eV, so we have to cool down more than expected to reach 50% recombination.

More accurate estimates put the recombination epoch at $\boxed{z_{\text{recomb}} \simeq 1100}$[11,119].

We therefore have $\boxed{T = 2.7(1100) \sim 3000\,\text{K}}$[11,119]

Possible point of confusion:

Just calculating the thermal energy $kT = \chi$ gives you a temperature of $T \sim 10^5$ K, much larger

than 3000 K. Normally we'd expect Hydrogen to start recombining at much hotter temperatures. What's going on here? For $T \sim 3000$ K, the ionizing energy 10^5 K is far into the tail of the photon distribution. However, there are *many* more photons than baryons ($n_b/n_\gamma \sim 10^{-10}$, see §8.Q145), so even in the tail of the distribution, we still have more than enough photons to keep H ionized. The baryon-to-photon ratio is encoded implicitly on the LHS of equation (8.158) within ρ_{crit}. We can write this more explicitly by substituting $n_b = \eta n_\gamma$ where we use the expression for n_γ from §8.Q145:

$$\frac{1}{2}\eta\frac{2\zeta(3)}{\pi^2}\left(\frac{kT}{\hbar c}\right)^3\left(\frac{2\pi m_e kT}{h^2}\right)^{-3/2} e^{\chi/kT} = 1 \tag{8.159}$$

Rearranging, we get

$$2 \approx 3.84\eta\left(\frac{kT}{m_e c^2}\right)^{3/2} e^{\chi/kT} \tag{8.160}$$

Note: to solve this numerically, it's helpful to do a log transform to avoid running into errors with machine precision.

139. When (or at what redshift) did the universe become (i) matter dominated? (ii) optically thin to electron scattering? What temperature was the CMB at these times? What significance did these events have for the CMB?

(i) Matter-Radiation Transition[11,119]

We want to find when the Ω_m and Ω_r terms in the Friedmann equation become equal to each other (Ω_K and Ω_Λ are negligible in this era). So

$$\Omega_m(1+z)^3 = \Omega_r(1+z)^4 \tag{8.161}$$

$$z = \frac{\Omega_m}{\Omega_r} - 1 \tag{8.162}$$

With current values of $\Omega_m \simeq 0.3$ and $\Omega_r \simeq 10^{-4}$ we get

$$\boxed{z_{\text{mr}} \simeq 3000} \tag{8.163}$$

The current CMB temperature is $T_0 \sim 3$ K, so this corresponds to a CMB temperature of

$$\boxed{T_{\text{CMB,mr}} = T_0(1+z_{\text{mr}}) \simeq 9000\,\text{K}} \tag{8.164}$$

The significance of this transition regarding the CMB has to do with how an overdensity perturbation enters the sound horizon. We saw in §8.Q131 that the baryons will start oscillating due to pressure support from photons when this happens. However, the dark matter has two options. If the perturbation enters the sound horizon before the epoch of matter-radiation equality, the dark matter will stagnate since the timescale for expansion ($\propto \rho_r^{-1/2}$) is much shorter than the timescale for collapse ($\propto \rho_m^{-1/2}$). Otherwise, if the perturbation enters the horizon after the epoch of matter-radiation equality, the dark matter will just continue collapsing like normal.

Thus, the difference in structure formation leaves an imprint on the CMB power spectrum through the SW and ISW effects (§8.Q128) which depend on the time-varying gravitational wells that photons pass through. The redshift z_{mr} also depends on Ω_m, which affects the amplitude of the CMB power spectrum as a whole.

(ii) Optically Thin Transition[119]

This occurs at **recombination**, which is at a redshift/temperature of

$$\boxed{z_{\text{recomb}} \simeq 1100} \tag{8.165}$$

$$\boxed{T_{\text{CMB,recomb}} = T_0(1+z_{\text{recomb}}) \simeq 3000\,\text{K}} \tag{8.166}$$

For a derivation of these values, see §8.Q138.

The significance of this transition regarding the CMB is that this sets the **surface of last scattering**, i.e. we see (approximately) all CMB photons coming from this redshift, since this is the last time they can scatter before the universe becomes optically thin.

This also sets the scale at which we see the BAO peaks in the CMB power spectrum, since these depend on the length of the sound horizon at recombination (which is when density perturbations stop oscillating and collapse). See §8.Q129 and §8.Q131.

(iii) BONUS ROUND Matter-Λ Transition

We can use the same process we did for (i) and apply it to the matter-cosmological constant transition:

$$\Omega_\Lambda = \Omega_m(1+z)^3 \tag{8.167}$$

$$z = \left(\frac{\Omega_\Lambda}{\Omega_m}\right)^{1/3} - 1 \tag{8.168}$$

For $\Omega_\Lambda \simeq 0.7$ and $\Omega_m \simeq 0.3$, this gives

$$\boxed{z_{\mathrm{m}\Lambda} \simeq 0.3} \tag{8.169}$$

In other words, this transition only happened relatively recently!

140. What evidence is there for dark energy and what might it be?

Dark energy is a spoopy name for the cosmological constant in the Friedmann equations. It is natural to attribute this phenomenon to the **vacuum energy density** of space, since if this existed it would produce a **negative pressure**, pushing things apart from each other. To see why, recall the discussion in §8.Q137 about the cosmological equation of state and the fluid equation:

$$P = w\rho c^2 \quad , \quad \frac{\mathrm{d}(a^3\rho)}{\mathrm{d}t} = -\frac{P}{c^2}\frac{\mathrm{d}(a^3)}{\mathrm{d}t} \tag{8.170}$$

The cosmological constant, by definition, has a constant density ρ_Λ that does not change with the scale factor of the universe, so the above equation requires $w = -1$:

$$P = -\rho_\Lambda c^2 \tag{8.171}$$

Due to this negative pressure, dark energy drives the **acceleration** of the expansion of the universe. As the universe expands, it fills more volume, but since ρ_{vac} is constant, this means more dark energy must continuously appear to fill the increasing volume, driving the expansion even faster. Some more general models consider a Λ that can change over time, replacing dark energy with *quintessence* (a time-dependent energy density). Some more exotic models consider dark energy might be due to modified gravity.

There is actually a recent result from the Dark Energy Spectroscopic Instrument (DESI) that suggests that dark energy may have a value of w that evolves over time, i.e.

$$w = w_0 + w_a(1-a) \tag{8.172}$$

where it reaches $w = -1$ somewhere between a redshift of $z \sim 0.2$–0.3 and has since evolved such that the value at $z = 0$ is between -0.8 and -0.6. The physical significance of this is difficult to interpret, but it boils down to an evolution in the nature of dark energy over time,[165] since when $w \neq -1$, the energy density is no longer constant with a / z, since $\rho \propto a^{-3(1+w)}$.

Calculations of the vacuum energy density from particle physics predict that it should scale with the inverse of the Planck time squared:

$$\Lambda c^2 \sim \frac{1}{t^2_{\text{Planck}}} \sim 10^{88}\,\text{s}^{-2} \quad \longrightarrow \quad \Omega_\Lambda = \frac{\Lambda c^2}{3H_0^2} \sim 10^{120} \tag{8.173}$$

This is 120 orders of magnitude off from the observed value of $\Omega_\Lambda \sim 0.7$. So, basically perfect agreement if you ask me[8]!

Evidence for Dark Energy

1. **CMB Power Spectrum**

   We can infer a value for Ω_Λ by fitting the CMB power spectrum, which gives $\Omega_\Lambda \simeq 0.7$. See §8.Q129. This comes from the measurements of Ω_m and Ω_K from the overall amplitude and the locations of the acoustic peaks, from which we can get Ω_Λ using

   $$1 - \Omega_m - \Omega_K = \Omega_\Lambda \tag{8.174}$$

   Which assumes $\Omega_r \simeq 0$.

[8] for legal reasons, this is a joke

2. **Type Ia Supernovae**

Type Ia Supernovae are standard candles that allow us to accurate get their distances. The distance that we derive from these measurements is the **luminosity distance** d_L, which depends not only on the hubble constant, but also on the cosmological density parameters. For low redshifts in a flat universe ($\Omega_K = \Omega_r = 0$):

$$d_L(z) = (1+z)\frac{c}{H_0}\int_0^z \frac{\mathrm{d}z'}{\sqrt{\Omega_m(1+z)^3+\Omega_\Lambda}} \tag{8.175}$$

We can see that this has a slight dependence on Ω_Λ (see §12.12 for a generalized definition of d_L). Thus, one can model the luminosity distances obtained from Type Ia supernovae as a function of z and fit models using different cosmological parameters to see what fits best. See what this looks like in Figure 8.17.

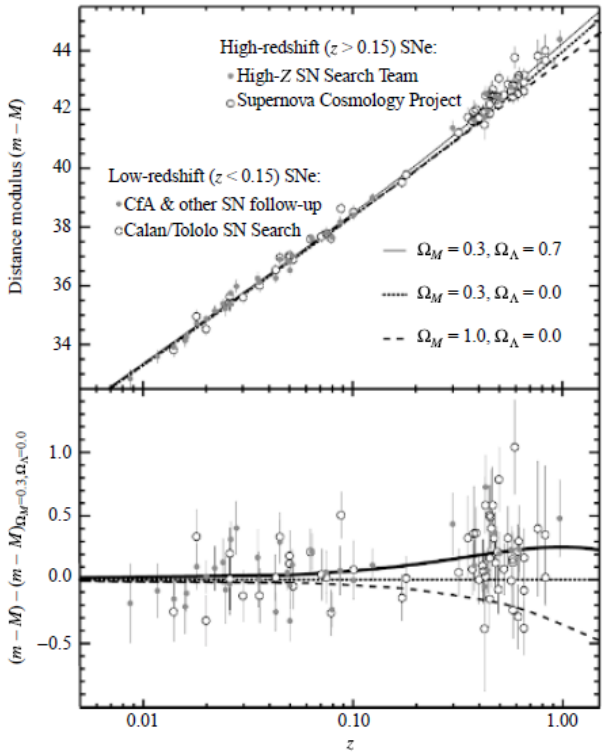
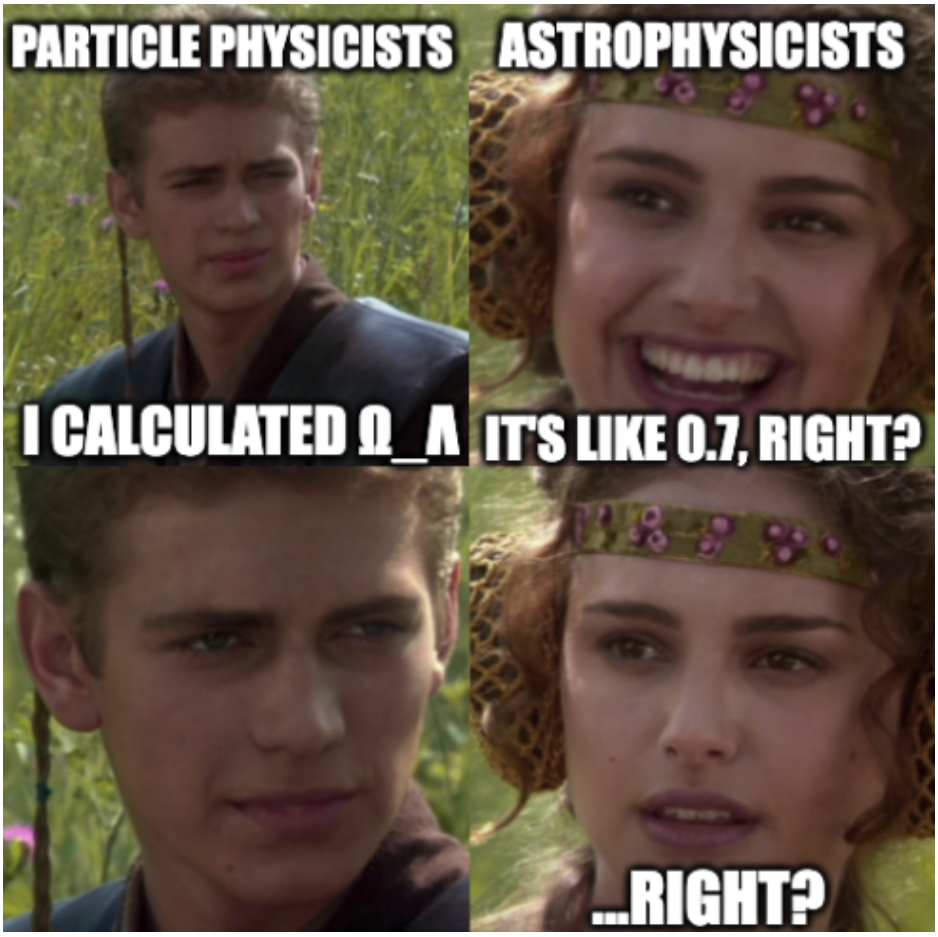


Figure 8.17: A model that fits the distance modulus $m-M$ as a function of redshift z, compared to a sample of Type Ia supernovae[11].

Looking at the fits in the Figure, we can see that the supernovae at larger redshifts notably were at higher distances than expected based on a model with no dark energy ($\Omega_\Lambda = 0$). We can see this by looking at (8.175) and imagining a model with $(\Omega_m, \Omega_\Lambda) = (1,0)$ vs. $(0.3, 0.7)$. At $z = 1$, the first model has $1/\sqrt{8}$ and the second has $1/\sqrt{0.3(2)^3+0.7} = 1/\sqrt{3.1}$, so the second is obviously larger. (If we want to imagine models with a nonzero curvature we have to use a more complicated expression for d_L, so I've omitted that here).

In other words, at a given distance, the supernovae had been redshifted less than expected. For a conceptual argument: a lower redshift corresponds to a higher scale factor, meaning

the expansion of the universe must have been *slower* in the past than it is today. Thus, the expansion of the universe is *accelerating*. This is a clear sign of dark energy[11]!



141. Is most of the hydrogen in the universe neutral or ionized? Describe the approximate ionization history of the Universe, and at what redshift the major phase transitions occurred.

Currently most of the hydrogen in the universe is **ionized**. For the thermal history of the Universe, refer to Appendix section §12.3.

In short:

- For times $t \lesssim 10^{-5}$ s, the universe is too hot even for quarks to combine to form hadrons (protons and neutrons)
- At $t \sim 10^{-5}$ (corresponding to $z \sim 10^{15}$), quarks and gluons form protons and neutrons. At this point, the universe is still too hot for leptons (electrons) to combine with the hadrons to form atomic nuclei, so the universe is completely **ionized**.
- At $t \sim 400$ kyr (corresponding to $z \sim 1100$), the universe has finally cooled down enough to allow electrons to combine with protons and neutrons to form atoms in a process called **recombination**. At this point the universe becomes almost completely **neutral**.
- Much later, at $t \sim$ Gyr (corresponding to $z \sim 6$), the first stars and galaxies form and start producing ionizing radiation. This eventually produces enough to almost completely reionize the universe in a process called **reionization**. The actual exact redshift this occurs at is still pretty uncertain and could be anywhere between ∼5–15.
- Today, $z = 0$, the Universe is still almost completely ionized.

A plot of the approximate ionization fraction of Hydrogen X_{HI} against redshift is shown in Figure 8.18.

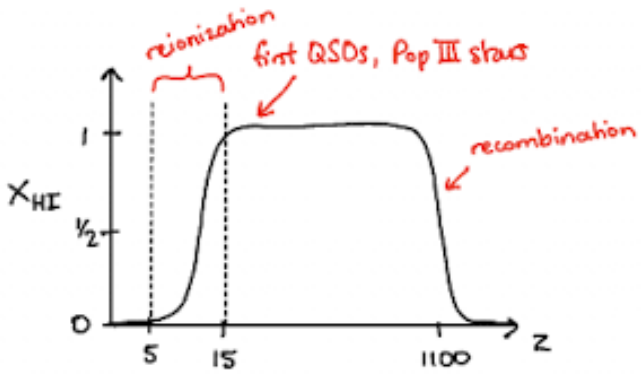


Figure 8.18: The ionization of Hydrogen in the universe as a function of redshift.

The epoch of reionization is probed with H I 21 cm cosmology (see §8.Q126) and the Lyα forest (§8.Q142) in the spectra of high-redshift quasars. The epoch of recombination is tightly constrained by the Saha equation (§8.Q138) and the CMB power spectrum (§8.Q129).

142. What is the Ly alpha forest? Why are cosmologists interested in it?

The **Lyα Forest** refers to a series of absorption lines shown in the spectra of high-redshift quasars that affect the UV continuum to the left of the Lyα line (shorter wavelengths). The idea is that, on its way back to us, the light from the quasar passes through a number of different intergalatic clouds of H and He at different redshifts that will then produce Lyα absorption lines at these different redshifts. A typical quasar spectrum might have something like 50 of these Lyα absorption lines at shorter redshifts. Absorption lines from He and metals can also sometimes be seen[119].

For an example of what this looks like, see Figure 8.19.

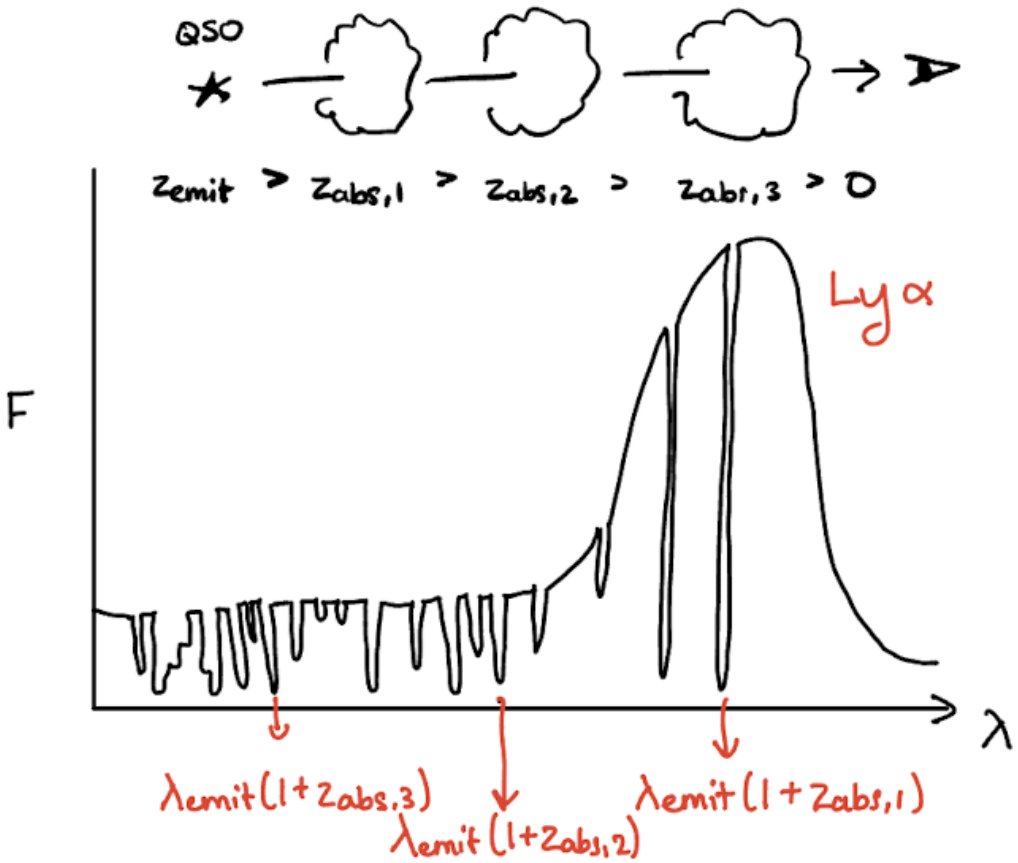


Figure 8.19: The Lyman-alpha forest of a typical quasar, with a diagram showing the path of the light above the spectrum.

Note: bound-bound absorption around the Lyα line is NOT the only important attenuation process going on here! There is also a "line blanketing" effect at wavelengths shorter than Lyα where the continuum is attenuated by a superposition of many absorption lines on top of each other as well as resonant scattering. This causes a staircase-like attenuation pattern at the wavelengths of each Lyman-series line.

There is also bound-free absorption (also sometimes called **photoelectric absorption**) for photons above the Lyman limit ($\lambda_{obs} < 912(1+z_{abs,i})$Å). The amount of absorption from this process is *strongly* dependent on the redshift, since the amount of neutral Hydrogen in the Universe that's even available to be ionized in the first place is also a strong function of redshift.

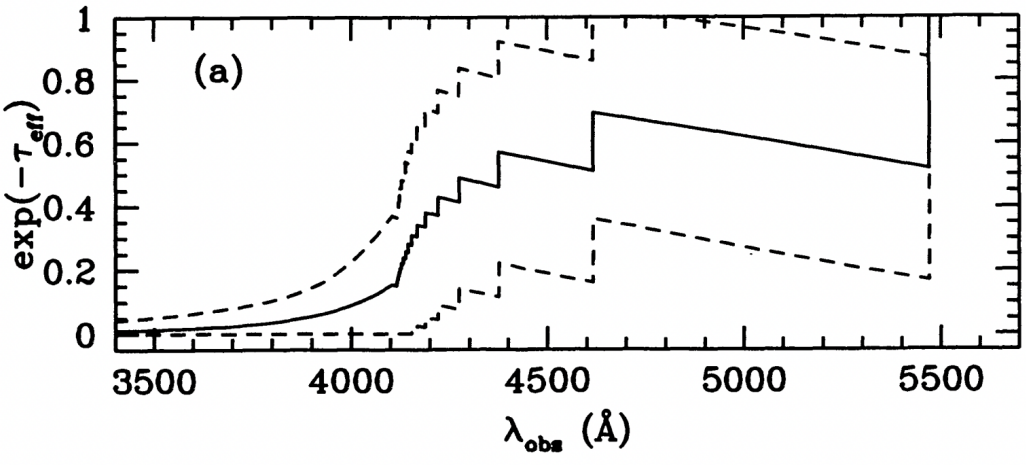


Figure 8.20: The attenuation fraction of a typical quasar at a redshift of $z = 3.5$. Notice the staircase drops that correspond to the Lyman series. Lyα: $1216(1+3.5) = 5472$ Å, Lyβ: $1026(1+3.5) = 4617$ Å, Lyγ: $972(1+3.5) = 4374$ Å, etc. These are caused by the blanketing effect. Above the Lyman limit, $912(1+3.5) = 4104$ Å, we have a drop-off due to bound-free absorption, but it does not go completely to 0 since the universe is already ionized at $z = 3.5$[166]

The combined effects of these processes are often referred to as **IGM absorption** since they originate from H in the IGM. If we take this to the logical extreme, we can find quasars at redshifts $z \gtrsim 6$, from before the reionization of the Universe. In this case, the UV continuum to the left of Lyα will drop almost completely to 0. This is because the vast majority of Hydrogen in the IGM is still neutral and available to be ionized. This is called a **Gunn-Peterson Trough** (see Figure 8.21).

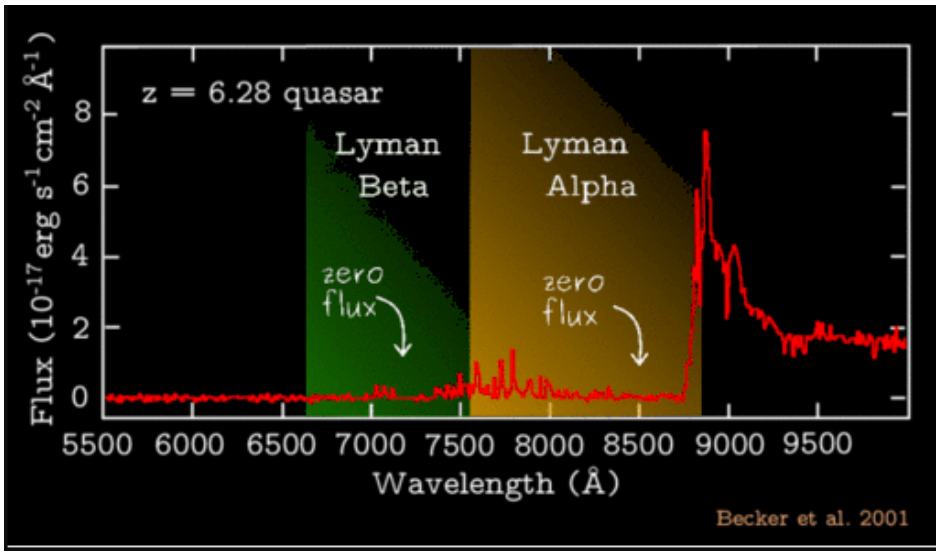


Figure 8.21: The Gunn-Peterson trough in an actual observed quasar[167].

Why are cosmologists interested? The absorption mechanisms are strongly dependent on the amount of neutral vs. ionized Hydrogen in the Universe as a function of redshift. Thus, the Lyα forests of high-redshift quasars can be used in tandem with 21 cm cosmology to constrain the epoch of reionization. We can also use the optical depth to get a measurement of the total amount of neutral H I in the IGM along a line of sight, which can help constrain DM structure models.

143. What are the "flatness problem" and the "isotropy problem" in standard Big-Bang cosmology? How does the inflationary model resolve these problems?

Flatness Problem[119]

This is a "fine tuning" problem related to how flat the Universe is today ($\Omega_m + \Omega_\Lambda + \Omega_r = 1$).

To see why this is a problem, start with the Friedmann equation and substitute in the critical density, ignoring Λ since it is negligible in the early universe

$$H^2 = \frac{8\pi G\rho}{3} - \frac{Kc^2}{a^2} + \frac{\Lambda c^2}{3} \quad \longrightarrow \quad \rho_{\text{crit}} a^2 = \rho a^2 - \frac{3Kc^2}{8\pi G} \tag{8.176}$$

Recall that $\rho = \Omega\rho_{\text{crit}}$:

$$\rho a^2\left(\frac{1}{\Omega} - 1\right) = -\frac{3Kc^2}{8\pi G} = \text{const} \tag{8.177}$$

Notice that the RHS is just a constant. In the matter-dominated regime, $\rho \propto a^{-3}$, so we have

$$\boxed{a \propto \left(\frac{1}{\Omega} - 1\right)} \tag{8.178}$$

Since the Planck time, the scale factor $a(t)$ has increased by a factor of $\sim 10^{60}$. In other words, $1/\Omega(t) - 1$ must have also increased by a factor of 10^{60}. This means that if $\Omega(t_{\text{Planck}})$ was even *slightly* off from 1, its value today would be *way off* from 1. But we observe today that $\Omega(t_0) \sim 1$, so at the Planck time it must have been within $\Omega(t_{\text{Planck}}) = 1 \pm 10^{-60}$. What are the chances of that, eh? It seems like the initial parameters of the Universe would have to be really finely tuned.

Isotropy Problem (Horizon Problem)[119]

This is a problem related to the seemingly paradoxical scale at which the CMB is causally connected.

In this context, the "horizon" is the boundary of the largest causally connected region. So, at the time when the CMB starts free streaming (recombination), what is the horizon size? Recall that in question §8.Q129 we derived the *sound* horizon size at recombination for the baryon acoustic oscillations. The *light* horizon size at recombination will follow the same steps, just replacing the speed of sound with the speed of light $c_s \to c_s\sqrt{3} = c$. In other words, our angular size is

$$\boxed{\theta_h = \frac{d_h}{d_A} = \frac{1}{\sqrt{z}} \simeq 1.7^\circ} \tag{8.179}$$

This is about 1 pixel in the WMAP mission, not even close to the whole sky. No 2 pixels in WMAP are causally connected...and yet, the CMB temperature is uniform over the whole sky to 1 part in 10^5. How is this possible?

Inflation[119,168]

Inflation is a theory that solves both of these problems simultaneously. It does so by proposing that in the very early history of the Universe, between $\sim 10^{-36} - 10^{-32}$ s, the Universe underwent extremely rapid expansion where the scale factor changed by a factor of $\sim 10^{30}$. See, for example, Figure 8.22.

This expansion is caused by some homogeneous scalar field ϕ (read: $\nabla\phi \equiv 0$) which produces a

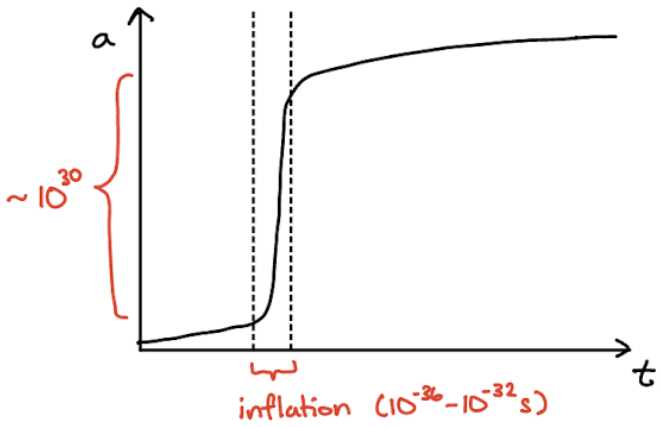


Figure 8.22: The scale factor as a function of time in the inflation model.

density and pressure

$$\rho(\phi) = \frac{1}{2}\dot{\phi}^2 + V(\phi) \tag{8.180}$$

$$P(\phi) = \frac{1}{2}\dot{\phi}^2 - V(\phi) \tag{8.181}$$

Then, the potential is assumed to dominate, $\dot{\phi}/2 \ll V(\phi)$, so ρ is constant in time. Notice that this has the same effect as the cosmological constant by causing exponential expansion. Plugging into the Friedmann equation:

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3} = \frac{8\pi G}{3}V(\phi) = \text{const.} \quad \longrightarrow \quad \frac{\mathrm{d}a}{a} = \text{const.} \quad \longrightarrow \quad a \propto \exp(t) \tag{8.182}$$

The other Friedmann equation gives us the equation of motion for the scalar field itself:

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0 \tag{8.183}$$

Notice that this is just a damped harmonic oscillator with a damping term $3H\dot{\phi}$. Usually here it is assumed that $\ddot{\phi} \ll 3H\dot{\phi}$ so that the potential, which we say starts on the top of a hill (unstable equilibrium), does a "slow roll" down into a potential minimum. Once it reaches the minimum, it no longer dominates the density ρ, so inflation ends. An example of what this potential might look like (in 1D) in shown in Figure 8.23. Note that this is not how the original inflation models proposed by Alan Guth in 1981[169] worked—they instead assumed that ϕ started out in a local minimum and then quantum tunneled to the global minimum. However, these models are disfavored today because they would produce large-scale inhomogeneities which are not observed.

We don't really have any interpretation for what this scalar field *actually is*...it's really just a mathematical device that we use to get this model where the Universe expands rapidly before calming down.

Now let's look at how this solves the flatness and isotropy problems. For flatness, look back at equation (8.177). Now instead of $\rho \propto a^{-3}$ we have $\rho \propto$ const., so

$$\left(\frac{1}{\Omega} - 1\right) \propto a^{-2} \tag{8.184}$$

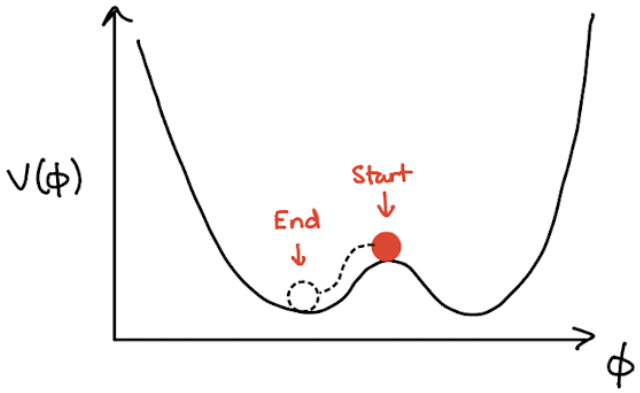


Figure 8.23: An example of a slow-roll inflation potential in 1D.

So now, a increasing by a factor of 10^{30} must cause $1/\Omega - 1$ to *decrease* by a factor of 10^{60}. So instead of driving Ω rapidly *away* from 1, we rapidly force Ω to *approach* 1 to minimize the LHS. In this model, having a flat Universe today is a natural consequence of the rapid inflation of the early Universe, and in fact we could not have any other kind of curvature. We can think of this conceptually like the inflation "stretching out" any curvature until it becomes flat

For the isotropy problem, we have to recalculate the horizon distance d_h in the ϕ-dominated regime:

$$d_h(t) = a(t)\int_0^t \frac{c\mathrm{d}t'}{a(t')} = a(t)\int_0^a \frac{c\mathrm{d}a'}{a'\dot{a}'} = a(t)\int_0^a \frac{c\mathrm{d}a'}{(a')^2 H} \tag{8.185}$$

Remember, $H = \text{const}$ here, so we get

$$d_h(a) = a\frac{c}{H}\left(\frac{1}{0} - \frac{1}{a}\right) \to \infty \tag{8.186}$$

The interpretation here is that before inflation the Universe is small enough that everything is causally connected. Then, when inflation rapidly increases the size of the Universe, it causes these large regions to have causal connections that they otherwise wouldn't have.

There is an additional problem not mentioned in the question that inflation also solves: the **magnetic monopole problem**. In short, there is no reason for us to believe that magnetic monopoles should not or cannot exist in our universe (we observe *electric* monopoles, after all). As such, according to the Big Bang model, there should have been many stable magnetic monopoles produced in the early universe. Inflation solves this by allowing magnetic monopoles to exist prior to the inflationary epoch. Then, after inflation, magnetic monopoles can no longer be produced, and the ones that *were* produced have had their densities diluted so strongly (expansion by a factor $\sim 10^{60}$) that we wouldn't expect to see any of them today.

144. What is the baryonic contribution to the cosmological mass density and how is it determined? Where are the bulk of the baryons in our Universe?

The baryonic content to the cosmological mass density is approximately

$$\boxed{\Omega_b \simeq 0.045 \pm 0.003} \tag{8.187}$$

Compared to an $\Omega_m \simeq 0.3$, the baryons account for only $\sim$ 15% of the total matter density of the Universe. This is obtained by two methods[119]:

1. **CMB**: Fitting the acoustic peaks in the CMB power spectrum and measuring the strength of the baryonic mass loading effect, which causes the relative amplitudes of the even and odd peaks to change. For details, see §8.Q129.
2. **BBN**: Measuring the abundances of ^{2}H, ^{3}He, ^{4}He, and ^{7}Li produced by BBN. The primordial abundances of these elements obviously depends on the amount of baryons in the Universe. For details, see §8.Q146.

Where are the bulk of the baryons in our Universe?

Notice that both of the above methods for measuring the baryon density probe the early epochs of the Universe. For late Universe methods of probing the current baryon density, this is mostly pretty easy since baryonic matter is luminous (read: we can observe it directly from its emitted light), but for dim objects like intergalactic H I clouds and black holes, we can use other methods like the **Lyman alpha forest** to probe H I column densities (§8.Q142) and **gravitational microlensing** to probe planets, black holes, and other dark objects (not dark matter).

Here we run into a problem: the measured amount of baryons in the late/current Universe is different from the amount in the early Universe—about 40% of the early Universe amount is unaccounted for in the current Universe. This is called the **missing baryons problem**. It is thought that most of these missing baryons reside in the Warm/Hot Intergalactic Medium (WHIM). We observe only ~7% in stars and galaxies, 5% in the CGM, 4% in the ICM, and only 2% in cold gas. The Lyman-alpha forest (i.e. the cold IGM) contains a whopping 28%, while the warm WHIM contains about 15%. This would make the current breakdown shown in Figure 8.24.

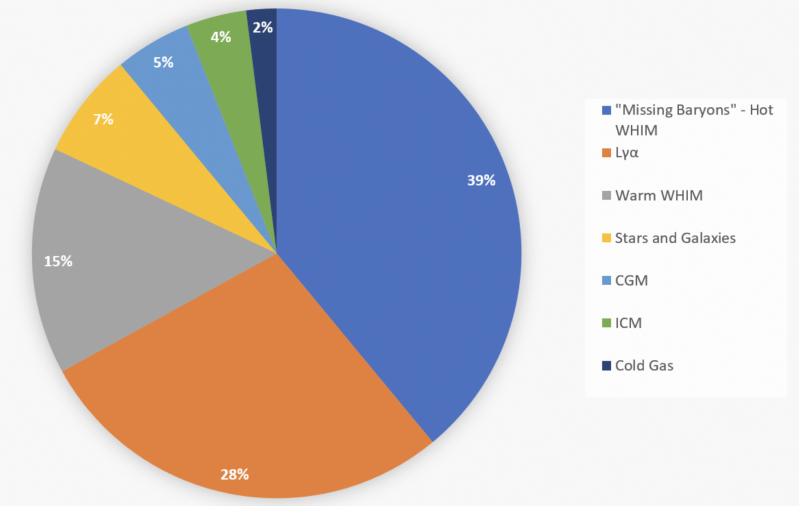


Figure 8.24: The baryon breakdown of the current Universe[170].

145. How did the particle content of the universe evolve? What is the most abundant particle in the universe today?

See Appendix section §12.3 on the thermal history of the Universe. In brief, the particle content evolves from being dominated by a quark-gluon plasma with neutrinos, to a soup of protons/neutrons and photons/leptons constantly exchanging with each other through pair creation and annihilation, to neutral atoms and photons, to mostly ionized nuclei with free electrons, photons, and neutrinos.

In the current universe, the dominant particles **by far** are **photons and neutrinos**. We can estimate by comparing their density parameters (Ω) and dividing by the mass-energy per particle to get a rough particle number. For baryons, this is easy:

$$n_b = \frac{\Omega_b \rho_{\text{crit}}}{m_p} \tag{8.188}$$

Where $\Omega_b \approx 0.045$ and $\rho_{\text{crit}} \approx 10^{-29}\,\text{cm}^{-3}$. This gives

$$\boxed{n_b \approx 3 \times 10^{-7}\,\text{cm}^{-3}} \tag{8.189}$$

For relativistic particles, this becomes a bit harder since we have to replace m_p with E/c^2, but E can have a wide range of values. If we assume particles are in a thermal distribution, we can use statistical arguments. Let's start with a simple dimensional analysis. In a thermal distribution of temperature T, particles tend to have energies around kT, which corresponds to a flux of kT/h particles/second, or a density of kT/hc particles/cm. In 3 dimensions, this becomes

$$n_\gamma \sim n_\nu \sim \left(\frac{kT}{hc}\right)^3 \tag{8.190}$$

Using $k \sim 10^{-16}$ erg/K, $T \sim 1$ K, $h \sim 10^{-27}$ erg s, and $c \sim 10^{10}$ cm/s, we get a surprisingly close value of

$$n_\gamma \sim n_\nu \sim 10^3\,\text{cm}^{-3} \quad \longrightarrow \quad \eta = \frac{n_b}{n_\gamma} \sim 10^{-10} \tag{8.191}$$

Now, being a little bit more careful, the fraction of particles in a thermal distribution with energy E is

$$f(E) = \frac{1}{\exp[(E-\mu)/kT] \pm 1} \tag{8.192}$$

where the $+1$ is taken for particles that obey Fermi-Dirac statistics and the -1 is taken for particles that obey Bose-Einstein statistics. We then integrate this over a 6D phase space $\text{d}^3\mathbf{x}\text{d}^3\mathbf{p}$ where each cell has a volume h^3 and we have a degeneracy g. We can get the number density by dividing out the volume, so

$$n = \frac{g}{h^3} \int \text{d}^3\mathbf{p}\, f(E(\mathbf{p})) \tag{8.193}$$

And recall $E(\mathbf{p}) = \sqrt{(pc)^2 + (mc^2)^2}$. Since we're interested in relativistic particles, $kT \gg mc^2$ and $E \approx pc$. We also assume the chemical potential $\mu \approx 0$ since photons can be easily created and

destroyed, and the momentum is isotropic, so $\mathrm{d}^3\mathbf{p} = 4\pi p^2 \mathrm{d}p$

$$n_{\mathrm{rel}} = \frac{4\pi g}{h^3} \int_0^\infty \frac{p^2 \mathrm{d}p}{\exp(pc/kT) \pm 1} \tag{8.194}$$

The integral can be evaluated analytically for the brave, which gives the relativistic particle number density. If we use the Bose-Einstein statistics, this is for photons, whereas Fermi-Dirac statistics gives neutrinos:

$$n_\gamma = \frac{\zeta(3)}{\pi^2} g \left(\frac{kT}{\hbar c} \right)^3 \quad , \quad n_\nu = \frac{3}{4} \frac{\zeta(3)}{\pi^2} g \left(\frac{kT}{\hbar c} \right)^3 \tag{8.195}$$

where $\zeta(x)$ is the Riemann-Zeta function and $\zeta(3) = 1.202$. Using $T = 2.7$ K for the CMB and $g = 2$, we get a relativistic number density of

$$\boxed{n_\gamma \approx 400\,\mathrm{cm}^{-3}} \quad , \quad \boxed{n_\nu \approx 300\,\mathrm{cm}^{-3}} \tag{8.196}$$

These are both **much** higher than the baryon number density, showing that the current universe is dominated by photons and neutrinos[24].

We also get a baryon-to-photon ratio when fitting BBN models, which give a value consistent with the above analysis of

$$\boxed{\frac{n_b}{n_\gamma} \approx 5 \times 10^{-10} \frac{\Omega_b h^2}{0.01}} \tag{8.197}$$

where the $\Omega_b h^2$ dependence comes from $\Omega_b \rho_{\mathrm{crit}}$ in the expression for n_b.

146. Describe, qualitatively, the synthesis of light elements in the Big Bang. What is the deuterium bottleneck, and how does it help produce the right helium abundance? What role does radiation play in Big Bang nucleosynthesis?

When the universe is between 1–100 s old, it has cooled down enough to form protons and neutrons, but is still hot enough to allow nuclear reactions. It is still radiation-dominated at this point.

The first thing that happens is the **"freeze-out"** of protons and neutrons themselves. By **freeze-out** we mean that the ratio between protons and neutrons in the universe becomes fixed because the reactions that exchange between them are no longer able to happen. Those reactions being:

$$\begin{aligned} \mathrm{n} + \nu_e &\longleftrightarrow \mathrm{p} + \mathrm{e}^- \\ \mathrm{n} + \mathrm{e}^+ &\longleftrightarrow \mathrm{p} + \bar{\nu}_e \end{aligned}$$

These are both mediated by the weak nuclear force. If we assume that before freeze-out the protons and neutrons were able to reach equilibrium, then their ratio should be determined by Boltzmann statistics:

$$n_i = g_i \left(\frac{m_i kT}{2\pi\hbar^2} \right)^{3/2} e^{-m_i c^2/kT} \quad \longrightarrow \quad \frac{n_n}{n_p} = \left(\frac{m_n}{m_p} \right)^{3/2} e^{-(m_n - m_p)c^2/kT} \tag{8.198}$$

This works because the neutron is slightly more massive than the proton $(m_n - m_p)c^2 = 1.29$ MeV, so it can be thought of as an "excited state" of the proton. Based on the reaction rates of the weak interactions, we know that freeze-out occurs at roughly $kT \sim 0.8$ MeV, which gives a neutron-to-proton ratio of $n_n/n_p \approx 0.20$. This freeze-out occurs roughly 1 second after the big bang.

Once the neutrons are frozen out, they can react with the protons to synthesize heavier nuclei. In general, reactions will always prefer making more stable nuclei. And of the nuclei with low atomic numbers, ^{4}He is the most stable by far. This can be seen by considering the possible reactions. First we can make deuterium ($^2\mathrm{H} = \mathrm{D}$) by

$$\begin{aligned} \mathrm{p} + \mathrm{n} &\longrightarrow \mathrm{D} + \gamma \\ \mathrm{p} + \mathrm{p} &\longrightarrow \mathrm{D} + \mathrm{e}^+ + \nu_e \\ \mathrm{n} + \mathrm{n} &\longrightarrow \mathrm{D} + \mathrm{e}^- + \bar{\nu}_e \end{aligned}$$

Here, the first reaction is mediated by the strong force while the latter two are mediated by the weak force. Therefore, the first once will be much more important, and we can pretty much ignore the other two. Then we can fuse deuterium to make

$$\begin{aligned} \mathrm{D} + \mathrm{p} &\longrightarrow {}^3\mathrm{He} + \gamma & {}^3\mathrm{He} + \mathrm{n} &\longrightarrow {}^4\mathrm{He} + \gamma \\ \mathrm{D} + \mathrm{D} &\longrightarrow {}^3\mathrm{He} + \mathrm{n} & {}^3\mathrm{He} + \mathrm{D} &\longrightarrow {}^4\mathrm{He} + \mathrm{p} \\ \mathrm{D} + \mathrm{D} &\longrightarrow {}^3\mathrm{H} + \mathrm{p} & {}^3\mathrm{H} + \mathrm{p} &\longrightarrow {}^4\mathrm{He} + \gamma \\ \mathrm{D} + \mathrm{n} &\longrightarrow {}^3\mathrm{H} + \gamma & {}^3\mathrm{H} + \mathrm{D} &\longrightarrow {}^4\mathrm{He} + \mathrm{n} \\ \mathrm{D} + \mathrm{D} &\longrightarrow {}^4\mathrm{He} + \gamma \end{aligned}$$

Notice, as we expected, all paths end in ^{4}He because it is more stable than anything else around it. Therefore, we can safely assume that all the neutrons eventually form ^{4}He nuclei, leaving the protons to form ^{1}H (i.e. just staying as protons) or ^{2}H (D) nuclei.

An ^{4}He nucleus contains 2 protons and 2 neutrons. So say that n_n neutrons combine with n_n protons to form $n_n/2$ helium nuclei each weighing 4 amu, leaving the remaining $n_p - n_n$ protons to form hydrogen each weighing 1 amu. Then our helium mass fraction would be

$$\frac{^4\text{He}}{\text{H} + ^4\text{He}} = \frac{4(n_n/2)}{(n_p - n_n) + 4(n_n/2)} = \frac{2(n_n/n_p)}{1 + (n_n/n_p)} = 0.33 \tag{8.199}$$

However, this is inconsistent with observations! Where did we go wrong? Well, **free neutrons are not stable** and they will decay into protons with a lifetime of about 15 minutes. Once protons and neutrons fuse to form deuterium (D) nuclei, which are much more stable, these decays will stop. But there is a period of time between when the protons and neutrons freeze-out and when deuterium fusion begins when some neutrons will decay, causing our initial ratio of 0.2 to decrease. This is known as the **deuterium bottleneck**.

To know how many neutrons decay, we need to know how long it takes for deuterium fusion to begin, so we need to know how at what temperature exactly it is able to start. For this, we can use the **Saha equation** (modified to account for the different masses of the particles):

$$\frac{n_D}{n_p n_n} = \frac{g_D}{g_p g_n}\left(\frac{m_D}{m_p m_n}\right)^{3/2}\left(\frac{kT}{2\pi\hbar^2}\right)^{-3/2} e^{2.22\,\text{MeV}/kT} \tag{8.200}$$

Where we have $g_D = 3$, $g_p = g_n = 2$, and $m_n \approx m_p \approx m_D/2$:

$$\frac{n_D}{n_n} = 6n_p\left(\frac{m_D kT}{2\pi\hbar^2}\right)^{-3/2} e^{2.22\,\text{MeV}/kT} \tag{8.201}$$

The proton number n_p is often expressed in terms of the baryon to photon ratio $\eta \equiv n_b/n_\gamma$ with $n_p \sim (5/6)n_b$. This is done because we know the primordial photon number well—see equation (8.195). Then we just substitute $n_p = (5/6)\eta n_\gamma$:

$$\frac{n_D}{n_n} = 5\eta \times \frac{2\zeta(3)}{\pi^2}\left(\frac{kT}{\hbar c}\right)^3\left(\frac{m_n kT}{\pi\hbar^2}\right)^{-3/2} e^{2.22\,\text{MeV}/kT} \tag{8.202}$$

We'll say the deuterium bottleneck ends when there are equal amounts of deuterium and neutrons, so

$$1 = 6.7\eta\left(\frac{kT}{m_n c^2}\right)^{3/2} e^{2.22\,\text{MeV}/kT} \tag{8.203}$$

Given we know $\eta \sim 5 \times 10^{-10}$ (§8.Q145), this can be solved numerically for T, giving $T \approx 69$ keV or 8×10^8 K. Since the universe is **radiation dominated**, this sets the timescale for how long it takes to cool from $kT = 0.8$ MeV to 69 keV, which corresponds to an age of $t = 200$ s. With a neutron half-life of 890 s, we can find the neutron to proton ratio at deuterium fusion to be

$$\frac{n_n}{n_p} = \frac{e^{-t/\tau}}{5 + (1 - e^{-t/\tau})} \approx 0.15 \tag{8.204}$$

Thus, we can update our helium mass ratio to be

$$\boxed{\frac{^4\text{He}}{\text{H} + ^4\text{He}} = \frac{2(n_n/n_p)}{1+(n_n/n_p)} = 0.26} \tag{8.205}$$

This does agree with observations, which get ~ 0.25, quite well! See in Figure 8.25 people have fit models to the observed ratios of ^{3}He/H, Y $\equiv$ ^{4}He/H, D/H, and ^{7}Li/H using η as the fitting parameter[24,119].

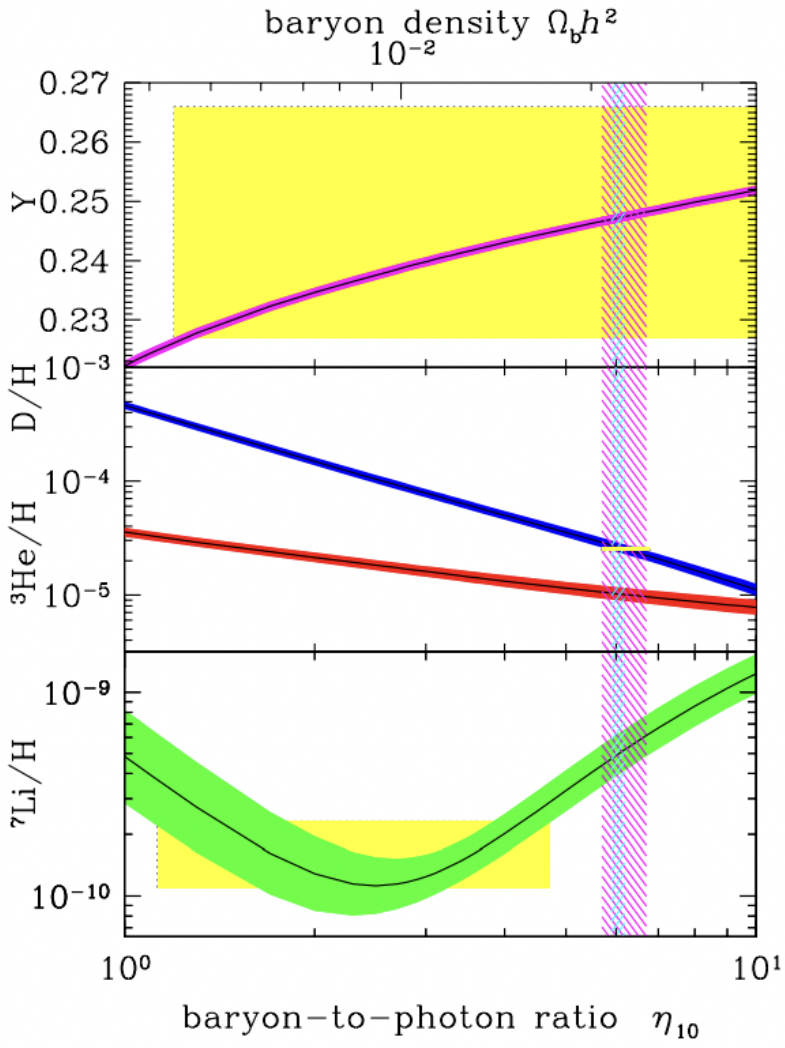


Figure 8.25: Big Bang Nucleosynthesis models[171].

147. What are the thermal and kinetic Sunyaev-Zel'dovich effects? What are they useful for?

When photons from the CMB reach the surface of last scattering, they can usually free stream to us without interacting with any more free electrons. However, if they happen to pass through the hot ICM of a galaxy cluster, which *does* have a lot of free electrons, they can be upscattered up to higher energies via **inverse Compton scattering** (because the electrons have significantly higher energies than the microwave photons). This is called the **Sunyaev-Zel'dovich (SZ) Effect**.

Note: The scattering of photons in different directions **does not** decrease the amount of CMB photons we receive coming from the direction of a galaxy cluster. Because the CMB is isotropic, statistically speaking, for every photon scattered *out of* our line of sight, there is another photon that is scattered *into* our line of sight. The difference in the CMB temperature observed over galaxy clusters is purely due to the *change in energy* of the CMB photons and not their scattering in different directions. The shift in the observed intensity is depicted in Figure 8.26.

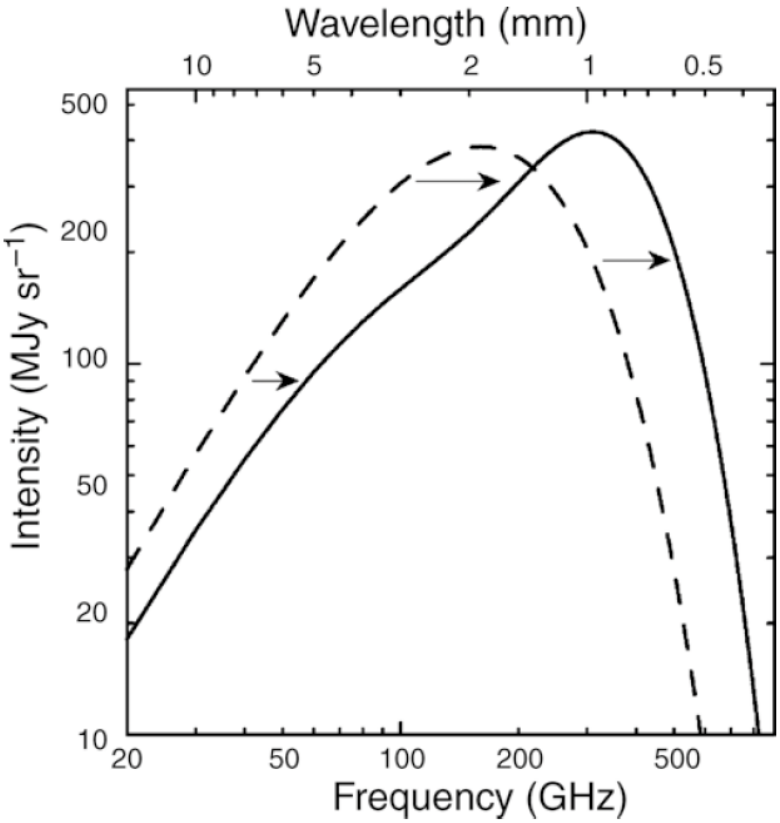


Figure 8.26: The shift in the observed intensity towards higher frequencies (and thus higher energies) as a result of the SZ effect[11].

Depending on how the electrons scatter the photons, this effect can be subdivided into two categories:

1. **Thermal SZ Effect**[119,172]

   The motions of electrons come from the **random thermal energy of the ICM**. To get the average change in energy of a photon in a single scattering event, we can use a clever trick. We take equation (5.33), which gives the average power radiated by an electron in a uniform magnetic

field, and just replace u_B with $u_{\rm rad}$ for the incident radiation field. Then, if we have a number density n electrons, we will have on average $N = n\sigma_T c$ collisions per second. Then,

$$\langle \Delta E \rangle = \frac{P}{N} = \frac{4}{3}\frac{\sigma_T}{N} c\gamma^2\beta^2 u_{\rm rad} = \frac{4}{3}\frac{1}{n}\gamma^2\beta^2 u_{\rm rad} \tag{8.206}$$

And the average energy of photons in a radiation field $u_{\rm rad}$ is

$$\langle E \rangle = \frac{u_{\rm rad}}{n} \tag{8.207}$$

Therefore

$$\left\langle \frac{\Delta E}{E} \right\rangle = \frac{4}{3}\gamma^2\beta^2 \tag{8.208}$$

In a thermal distribution of electrons, $m_e v_e^2/2 = 3kT_e/2$. Plugging in for β and assuming $\gamma \approx 1$:

$$\left\langle \frac{\Delta E}{E} \right\rangle = \frac{4kT_e}{m_e c^2} \tag{8.209}$$

We often define the Compton y-parameter based on this ratio as a dimensionless measure of the amount the energy changes from repeated Compton scatterings in a distribution of electrons:

$$\boxed{y \equiv \int {\rm d}\ell \, \frac{kT_e}{m_e c^2}\sigma_T n_e(\ell)} \tag{8.210}$$

Notice that y is independent of the frequency, so all frequencies will shift in energy fractionally by the same amount. This is in effect the same thing as shifting the temperature of the blackbody spectrum. Now we can recast the change in energy into a change in the observed temperature. In the Rayleigh-Jeans limit we have $I_\nu \propto \nu^2 T_{\rm bb}$ where $T_{\rm bb} \neq T_e$ so

$$\frac{\partial I_\nu}{I_\nu} \approx \frac{\partial T_{\rm bb}}{T_{\rm bb}} \approx 2\frac{\partial \nu}{\nu} \approx 2\frac{\partial E}{E} \quad \longrightarrow \quad \frac{\partial I_\nu}{I_\nu} \approx \frac{\partial T_{\rm bb}}{T_{\rm bb}} \sim -2y \tag{8.211}$$

Doing a more careful analysis would yield a frequency dependence which we can absorb into a generic function $f(\nu)$:

$$\boxed{\frac{\Delta T_{\rm bb}}{T_{\rm bb}} = f(\nu)y = f(\nu)\frac{kT_e}{m_e c^2}\sigma_T N_{e,{\rm col}}} \tag{8.212}$$

where

$$f(\nu) = \frac{h\nu}{kT_{\rm bb}}\frac{e^{h\nu/kT_{\rm bb}}+1}{e^{h\nu/kT_{\rm bb}}-1} - 4 \tag{8.213}$$

This confirms that in the Rayleigh-Jeans limit, $f(\nu) \to -2$. Typically $\Delta T_{\rm bb}/T_{\rm bb} \sim 10^{-3}$.

2. **Kinetic SZ Effect**[173]

The motions of the electrons come from the **bulk motion of the ICM**, which tends to drift towards other clusters. In this case, the bulk velocity will impart an observed Doppler shift on

the CMB photons. Recall that in the limit of $v \ll c$ the Doppler shift is given by

$$\frac{\Delta\lambda}{\lambda} = \frac{\Delta\nu}{\nu} = \frac{\Delta v}{c} \tag{8.214}$$

Then this imparts a change in the observed temperature

$$\boxed{\frac{\Delta T_{\rm bb}}{T_{\rm bb}} = -\frac{v_{\rm pec}}{c}\tau_e} \tag{8.215}$$

Where $\tau_e = \sigma_T N_{e,\rm col}$ is the optical depth, and the peculiar velocity $v_{\rm pec}$ is measured with respect to the Hubble flow. Thus, this can shift the observed energies in either direction (lower or higher) depending on whether the peculiar velocity is away from or towards us. This is typically a smaller effect than the thermal SZ effect. See the change in intensity and tempearture due to both effects in Figure 8.27.

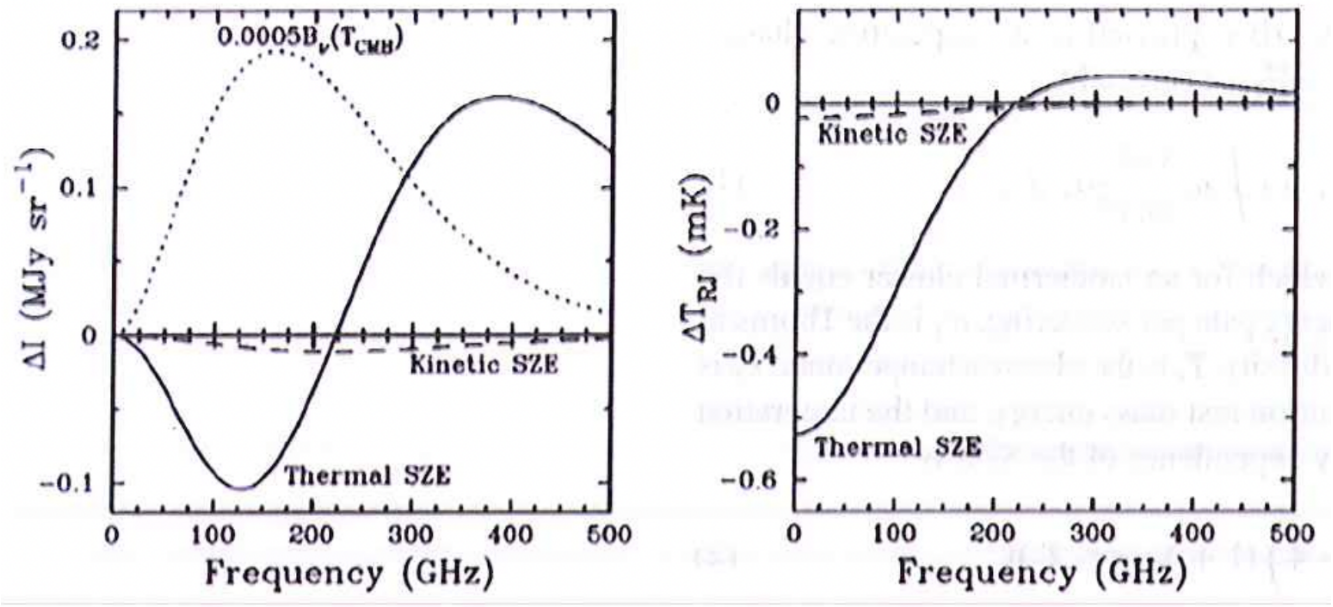


Figure 8.27: The change in intensity plotted as ΔI_ν (left) and temperature ΔT (right) for both the thermal (solid) and kinetic (dashed) SZ effects[173].

The SZ effect has a number of uses

- Can be used to discover galaxy clusters by looking for decrements in the CMB. Since the SZ effect is independent of redshift z, this is a great technique to find high-z clusters.
- Can probe the pressure profiles of clusters since the thermal effect is proportional to $n_e k T_e$ along different lines of sight.
- Can probe the masses of clusters since both effects are proportional to the column density of the ICM $N_{e,\rm col}$ along different lines of sight.
- Can probe peculiar velocities of clusters with the kinetic SZ effect to see how their motions are correlated with each other (though this is hard since the kinetic SZ effect is much smaller and hard to disentangle from the thermal SZ effect)

148. What range of physical scales are being probed by current and future CMB experiments? Are the anisotropies related to galaxy formation? What is the current observational status?

Physical Scales

The angular scales of past, present, and future CMB experiments are:

- COBE (1989–1993)[174,156]: $\gtrsim 5^\circ$, $\ell \sim 2-40$
- WMAP (2001–2010)[175,176]: $\gtrsim 10'$, $\ell \sim 2-1200$
- Planck (2009–2013)[177,178]: $\gtrsim 5'$, $\ell \sim 2-2000$
- SPT (2011–Present)[179,180]: $\gtrsim 4'$, $\ell \sim 300-3000$
- ACT (2010–Present)[181,182]: $\gtrsim 1'$, $\ell \sim 300-8000$
- CMB-S4 (Planned)[183]: $\gtrsim 1'$, $\ell \sim 2-8000$
  - Plans to get both low and high ℓ to constrain B-mode polarization

Notice, the ground-based telescopes (SPT and ACT) can get up to a higher maximum ℓ because they can be built with larger diameters than the space-based telescopes. But they cannot probe the low-ℓ modes because they can only see half of the sky.

Are the anisotropies related to galaxy formation?

In short, yes. Anisotropies are due to matter over- and under-densities in the early universe that we believe eventually grow into the large scale structure of the universe (galaxy clusters, groups, and voids) that we observe today. The differences in gravitational potential lead to the SW and ISW effects, and acoustic oscillations that occur while these density perturbations evolve imprint their signature wavy pattern on the CMB power spectrum. However, the Silk damping tends to smooth out these anisotropies on scales smaller than the Silk mass, $M_{\rm silk} \sim 10^{13}\,M_\odot$. For details, see §8.Q128.

What is the current observational status?

- Planck data has excellent quality down to an $\ell \sim 2000$, and modern analyses are reaching the statistical limit of what can be done with the CMB power spectrum.
- There is a fundamental limit on the uncertainties we can obtain for the low-ℓ end of the CMB, since these values require averaging over large portions of the sky. We can only get a handful of data points to contribute to these measurements (for example, $\ell = 1$ corresponds to $\theta = 180^\circ$, so we only get 2 data points to work with).
- More work and higher resolution measurements *are* needed for CMB polarization measurements. The future BICEP3[184] and CMB-S4 observatories will be working on this.

149. What is the "Lyman limit", and how does it relate to observations of high-redshift galaxies?

This was mentioned in §8.Q142 on the Lyα forest, but to go into a bit more detail, the **Lyman limit** (sometimes called the **Lyman break**) is the wavelength/frequency/energy of the Lyman series ($n_1 = 1$) of Hydrogen atoms at the limit where the initial state n_2 goes to infinity. Using the Rydberg formula:

$$E = 13.6\,\text{eV}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right) \tag{8.216}$$

We can see that using $n_1 = 1$ and $n_2 = \infty$ just gives us $E = 13.6$ eV, the ionization potential of Hydrogen. This corresponds to a wavelength of ~ 912 Å. In other words, photons with wavelengths shorter than this will be able to ionize neutral Hydrogen.

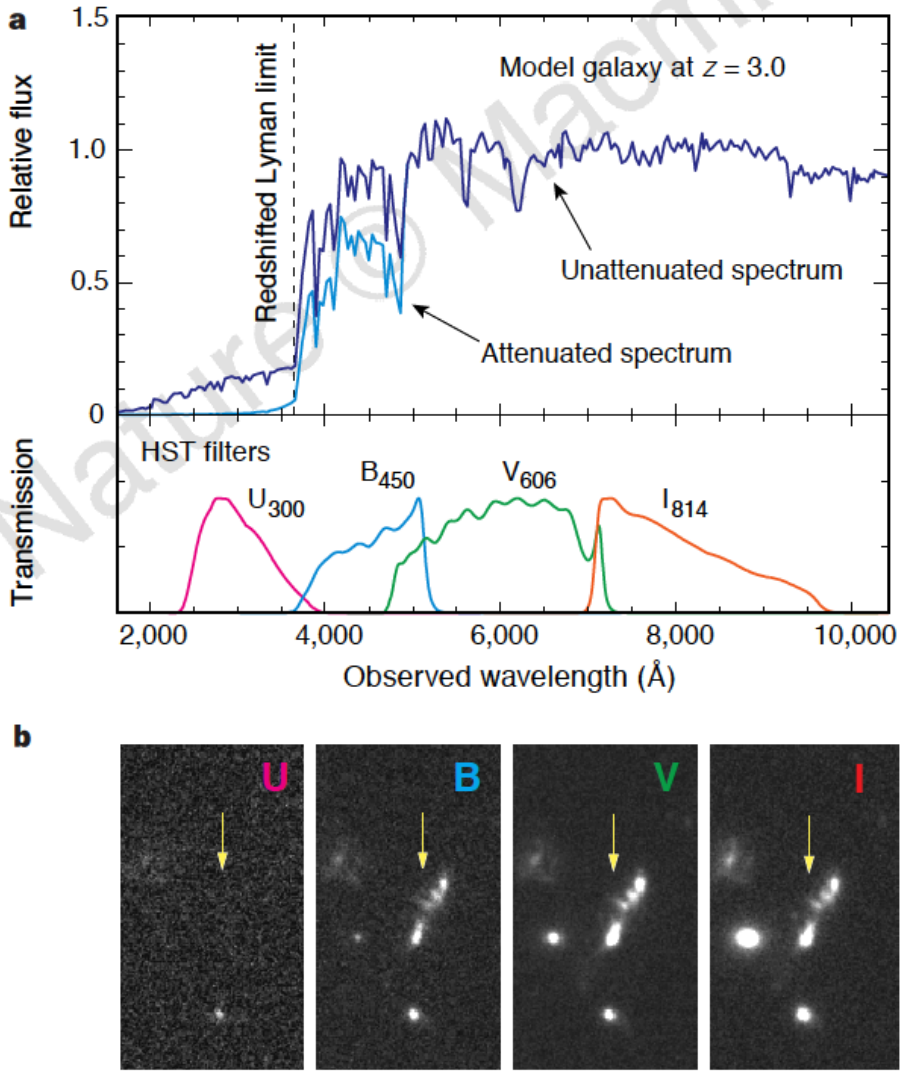


Figure 8.28: An example spectrum of a Lyman-limit galaxy and corresponding images in different filters showcasing the photometric redshift[185]. Sorry I coudln't find a version without the shitty watermark.

This is an important cutoff because it allows us to probe the neutral Hydrogen content of the Universe as a function of redshift using high-redshift galaxies. When there is a lot of neutral Hydrogen, the Lyman continuum will drop to 0 above the Lyman limit, whereas if most of the Hydrogen is already ionized, the continuum will be much less attenuated. The continuum actually starts getting attenuated *before* the Lyman limit due to **IGM absorption** (see §8.Q142 for a discussion), but it doesn't completely drop to 0 until the Lyman limit. See Figure 8.28.

The steep dependence of the spectrum on redshift also allows for the use of the **photometric redshift** technique, where we can estimate the redshift of a source just by looking at it in a few different filters and noting when the galaxy completely fades from view. This obviously isn't as precise as spectroscopic redshifts, but it can be an alright substitute if there is a clear drop-off in intensity and we don't have any spectroscopic data to work with. Rohan Naidu at MIT/MKI uses this technique.

150. What mass should a neutrino have to close the universe? Compare it with current upper bounds (or detections).

The density required to close the universe (i.e. make it flat with $k = \Lambda = 0$) is, by definition, the critical density

$$\rho_{\text{crit}} = \frac{3H_0^2}{8\pi G} \sim 10^{-29}\,\text{g}\,\text{cm}^{-3} \tag{8.217}$$

(see §8.Q137). Comparing this to the number density of neutrinos found in §8.Q145, we get

$$m_\nu = \frac{\rho_{\text{crit}}}{n_\nu} \approx \frac{10^{-29}}{300} \approx 10^{-32}\,\text{g}\,\text{cm}^{-3} \approx \boxed{19\,\text{eV}} \tag{8.218}$$

However, if we only need the neutrinos to be dark matter, we can instead say

$$m_\nu = \frac{\Omega_c \rho_{\text{crit}}}{n_\nu} \approx \frac{0.27 \times 10^{-29}}{300} \approx \boxed{5\,\text{eV}} \tag{8.219}$$

Compared to current upper bounds from the KATRIN[186] experiment are

$$m_\nu \leqslant 0.8\,\text{eV} \tag{8.220}$$

Clearly the neutrino does not seem to be the sole answer to dark matter. It can also be argued that neutrinos in fact *shouldn't* be most of the dark matter since they are "hot" (relativistic), which would wipe out small scale structure formation. There is one theory of dark matter that posits there is a fourth "sterile" neutrino that is heavier than the other neutrinos and doesn't participate in the weak interaction, but this hypothesis has yet to see any observational evidence. See neutrino mass constrains in Figure 8.29.

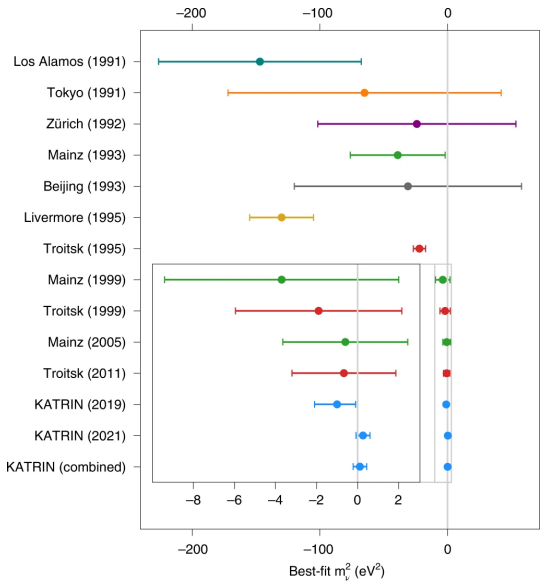


Figure 8.29: A history of neutrino mass constrains, with the most recent estimates being near the bottom. A zoomed-in panel shows the most recent results. Note that the x axis here is m^2. See Ref. 187.

151. What is the meaning and purpose of a log N–log S curve? Explain the current interpretation of this curve for gamma-ray bursts.

This is essentially a test for us to see if the distribution of some sources on the sky are homogeneous and isotropic (or not) without directly knowing the distance to each source. The time-integrated flux S (sometimes called "fluence") is defined as

$$S = \frac{E}{4\pi d^2} \tag{8.221}$$

If we assume every source has the same intrinsic energy E, and these sources have some minimum detectable fluence S_{min} corresponding to a maximum distance

$$d_{\text{max}} = \sqrt{\frac{E}{4\pi S_{\text{min}}}} \tag{8.222}$$

Then the total volume in which we can detect sources is

$$V = \frac{4}{3}\pi d_{\text{max}}^3 \tag{8.223}$$

And if there is a number density n of sources, then the total number of detectable sources is

$$N = nV = \frac{4}{3}\pi n\left(\frac{E}{4\pi S_{\text{min}}}\right)^{3/2} \longrightarrow \boxed{N \propto S_{\text{min}}^{-3/2}} \tag{8.224}$$

Therefore, we can plot the number of bursts N with integrated fluxes above some minimum threshold S_{min} (so N is a function of S_{min}), and if the distribution is homogeneous and isotropic, we should expect to see a slope of -1.5 (see Ref. 188).

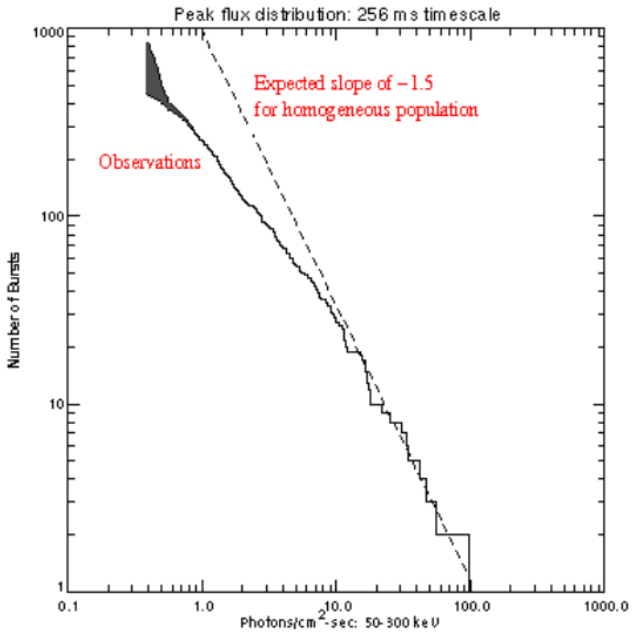


Figure 8.30: The log N–log S plot for gamma-ray bursts[189].

Doing this plot for gamma-ray bursts (Figure 8.30), we see a perfect match at the bright end of the

distribution, but a deviation from the expected slope for lower values of $S_{\min}$. The fact that the bright end of the distribution is isotropic means we cannot localize the GRBs to our own galaxy, otherwise we would expect an anisotropic distribution (since we are offset from the galactic center)—they must be extragalactic in origin. The deviations at the lower end means that we're seeing the edge of the distribution—there is a hard cutoff at some large distance/redshift where we no longer see GRBs. What could this mean?

- The Universe has a finite age, so at very large distance we're seeing far enough back in the Universe's history that galaxies had not yet formed, so there can be no GRBs.
- The rate of GRBs was probably different in the past than it is today.
- Could in principle be explained by non-Euclidean (curved) geometry, but this is not backed up by other cosmology studies.

152. In our Universe, at roughly what scale does perturbation theory break down and become nonperturbative?

Perturbation theory intrinsically assumes that the perturbations are small (i.e. $\delta \ll 1$). When the density constrast grows to order unity, this assumption no longer holds. We can see from the perturbation studied in §8.Q132, the density contrast when the radius is at a maximum is, from (8.88),

$$\delta_{\max} = \frac{\rho_{\max} - \rho_{u,\max}}{\rho_{u,\max}} = \frac{9}{16}\pi^2 - 1 \approx 4.6 \tag{8.225}$$

Thus, we can *very roughly* consider the density's growth up until this point to be perturbative, but after it reaches its maximum size and decouples from the Hubble flow, we have to treat it nonlinearly. In a matter-dominated universe, this corresponds to

$$\rho(z) \approx 5.6\rho_u(z) \approx 5.6\frac{3H_0^2}{8\pi G}\Omega_m(1+z)^3 \tag{8.226}$$

If we wanted to be more pedantic, we could set $\delta = 1$, corresponding to $\rho(z) = 2\rho_u(z)$, which just changes our $\mathcal{O}(1)$ constant out front from $5.6 \to 2$. This just means the density perturbation actually becomes nonlinear before it reaches maximum expansion. I'll use this from now on (it doesn't have a huge effect on the results). Plugging in $\rho = 3M/4\pi r^3$ and solving for r:

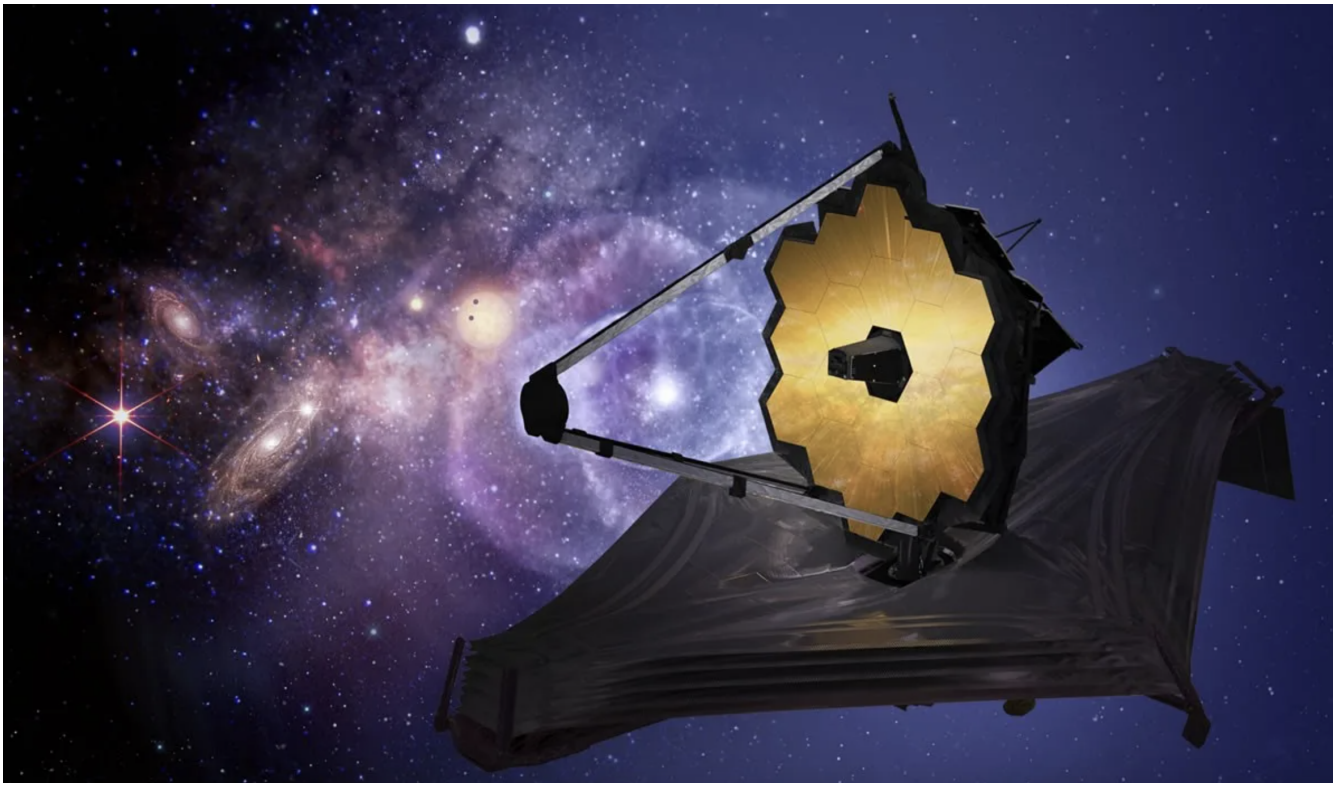
$$r(z) = \left(\frac{GM}{H_0^2\Omega_m}\right)^{1/3}\left(\frac{1}{1+z}\right) \tag{8.227}$$

Rewriting the constants in a more convenient form

$$\boxed{r(z) = \left(\frac{M}{10^{11}\,M_\odot}\right)^{1/3}\left(\frac{1}{1+z}\right)\text{Mpc}} \tag{8.228}$$

Any systems above this scale will evolve non-linearly. Generally speaking, the largest systems like galaxies and galaxy clusters will go through this nonlinear phase where they eventually virialize, which prevents them from collapsing any further. There is more power at smaller scales, so smaller scales tend to evolve nonlinearly first, creating hierarchical structure formation[190].

9 Instrumentation & Astronomical Methods



153. Explain how a radio interferometer manages to produce sub-arcsecond images when none of the constituent dishes has an angular resolution of better than 1 arcmin.

The basic principle behind **radio interferometry** is to use multiple radio dishes to increase the angular resolution of the observations. Since the angular resolution goes like

$$\Delta\theta \propto \frac{\lambda}{D} \tag{9.1}$$

Increasing the effective diameter D allows us to resolve smaller angular differences $\Delta\theta$. But how does this actually work?

Aperture synthesis

Each individual radio dish has an antenna that measures a voltage V_i as a function of time. The actual voltage measured will be a sum of the voltages received from different locations on the sky. For example, if we have coordinates (ℓ, m) that parametrize the location on the sky, then

$$V_i = \iint V_i(\ell, m)\mathrm{d}\ell\,\mathrm{d}m \tag{9.2}$$

Now, image we have two physically separated radio dishes looking at the same part of the sky in a monochromatic band at frequency ν. They will receive the signal with a relative time delay based on how far apart they are. If two telescopes are separated by a distance b and looking at an object at a colatitude θ, then the extra path length taken for one dish will be $b\sin\theta$. See the physical geometry in Figure 9.1. This extra path length, shown in yellow, will be constant over the field of view, so we can add an extra delay in the signal for the telescope 2 to allow them to arrive at the signal processing computer (shown with the blue box) at the same time.

However, there will be an additional delay (shown in green) that *is* dependent on the location in the field of view. At an angle α away from the center, the additional delay will be $u\sin\alpha$, where $u = b\cos\theta$ (note that this is only a delay when α is in the same direction as θ, as is shown in the Figure; if α is in the other direction, then it becomes a slight lead time). Now, the sky has 2 dimensions, so we can do the exact same geometry in the orthogonal direction (into the page), which adds another component to the delay/lead time. We define the other coordinates to be ν and β, such that our delay is $\nu\sin\beta$. Then, we define $\ell = \sin\alpha$ and $m = \sin\beta$ such that the total delay time is $u\ell + \nu m$. It is also customary to normalize u and ν by the wavelength λ such that they represent a number of cycles. Then we can write the voltage in telescope 2 relative to that in telescope 1 simply by

$$V_2 = V_1 e^{-2\pi i(u\ell+\nu m)} \tag{9.3}$$

Now, when the signals from the two telescopes arrive at the computer, they are sent through a **correlator**, which multiplies them and takes their time-average. Why this is done may not seem obvious at first, but if we go through the math, we'll see the consequences:

$$C = \langle V_i V_j \rangle = \left\langle \iint V_i(\ell_i, m_i)\mathrm{d}\ell_i\mathrm{d}m_i \iint V_j(\ell_j, m_j)\mathrm{d}\ell_j\mathrm{d}m_j \right\rangle \tag{9.4}$$

Now, if the two sky locations are different in the two telescopes (i.e. $\ell_i \neq \ell_j$ and/or $m_i \neq m_j$),

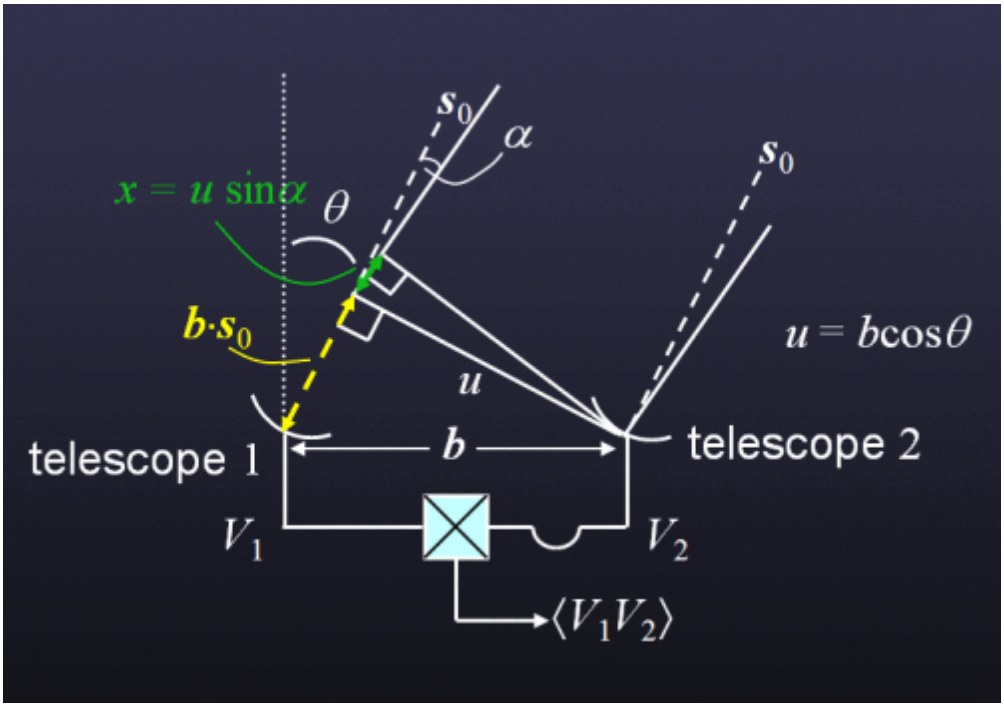


Figure 9.1: The setup for radio interferometry between two telescopes.

obviously the signals will be uncorrelated, so $C = 0$. However, if they're the same, we get something interesting

$$C = \langle V_i V_j \rangle = \left\langle \iint V_i(\ell, m) V_j(\ell, m) \mathrm{d}\ell \mathrm{d}m \right\rangle \tag{9.5}$$

$$= \iint \langle V_i(\ell, m) V_j(\ell, m) \rangle \mathrm{d}\ell \mathrm{d}m \tag{9.6}$$

Using (9.3), we have

$$C = \langle V_i V_j \rangle = \iint \langle V_i(\ell, m)^2 \rangle e^{-2\pi i(u\ell + vm)} \mathrm{d}\ell \mathrm{d}m \tag{9.7}$$

$$\longrightarrow \quad \boxed{\mathcal{V}(u, v) = \iint I_\nu(\ell, m) e^{-2\pi i(u\ell + vm)} \mathrm{d}\ell \mathrm{d}m} \tag{9.8}$$

Well look at that, since the observations are in a monochromatic wave band, the average of V^2 is proportional to the intensity I_ν, and the correlation signal is just proportional to the **Fourier transform of the intensity**! We often call $\mathcal{V}(u, v)$ the **complex visibility function**, where the u and v coordinates can be thought of as spatial frequencies in the EW and NS directions. As we see, one pair of telescopes gives us **one sample** of $\mathcal{V}(u, v)$ for one set of coordinates (u, v). But to reconstruct the 2D intensity $I(\ell, m)$ perfectly from $\mathcal{V}(u, v)$, we would need to know $\mathcal{V}(u, v)$ for all possible values of u and v:

$$\boxed{I_\nu(\ell, m) = \iint \mathcal{V}(u, v) e^{2\pi i(u\ell + vm)} \mathrm{d}u \mathrm{d}v} \tag{9.9}$$

Note that, in practice, since the elements of the interferometer (the telescopes) have a finite size, they have a response function $\mathscr{A}(\ell, m)$ called the **primary beam**, which gives the normalized reception pattern of the individual elements. So the actual correlation measured will be

$$C = \iint \mathscr{A}(\ell, m) I(\ell, m) e^{-2\pi i(u\ell + vm)} \mathrm{d}\ell \mathrm{d}m \tag{9.10}$$

Therefore, after doing the Fourier transform, we must divide by $\mathscr{A}(\ell, m)$ to recover the true $I(\ell, m)$.

Obviously it's impossible to get every single value of (u, v), but we can scale this up so that with N telescopes we have $N(N-1)/2$ pairs from which we have unique values of (u, v). We can also take advantage of the rotation of the Earth, which changes the angle θ from the zenith that our source lies and therefore changes the (u, v) values for each pair. An example of this for the SMA is shown in Figure 9.2.

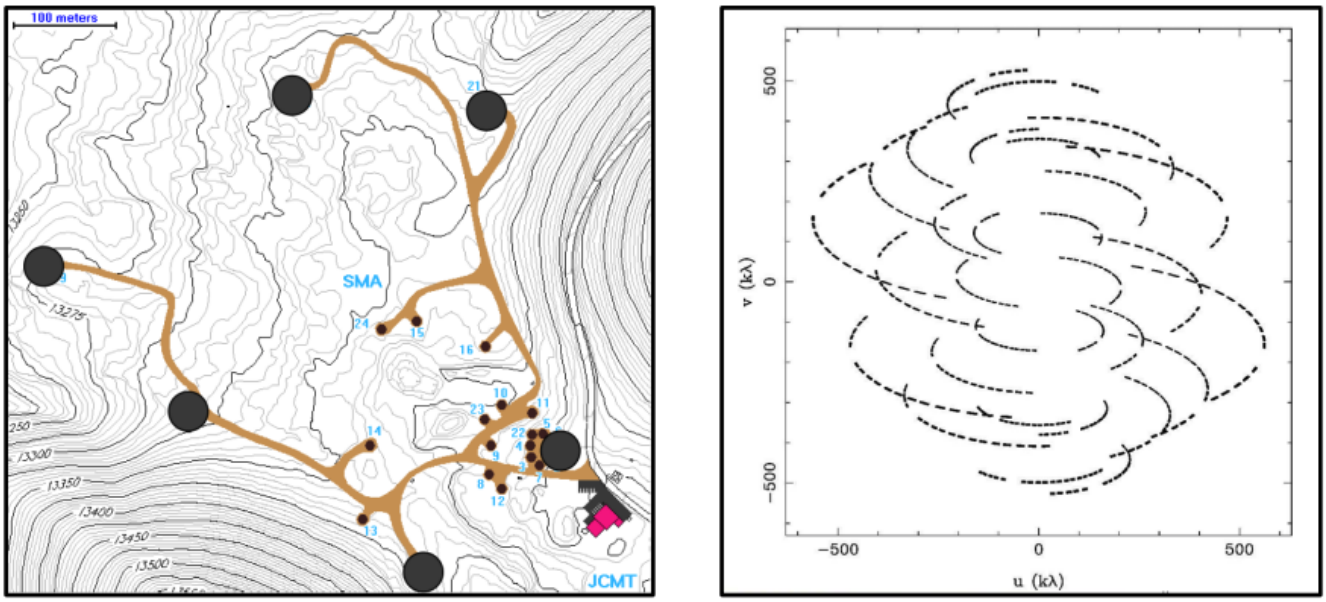


Figure 9.2: The u and v values for each pair of telescopes in the SMA for its "extended" configuration.

Not having full coverage in (u, v) space creates noise when we try to estimate the Fourier transform, so there are a lot of signal processing and deconvolution techniques that radio astronomers use to try to reduce this niose. An example of what happens when we miss coverage in different parts of (u, v) space is shown in Figure 9.3.

The *maximum* separation between two telescopes, B_{max}, determines the angular resolution of the resultant image

$$\Delta\theta_{\mathrm{min}} \sim \frac{\lambda}{B_{\mathrm{max}}} \tag{9.11}$$

And the *minimum* separation between two telescopes, B_{min}, determines the field of view of the resultant image

$$\Delta\theta_{\mathrm{max}} \sim \frac{\lambda}{B_{\mathrm{min}}} \tag{9.12}$$

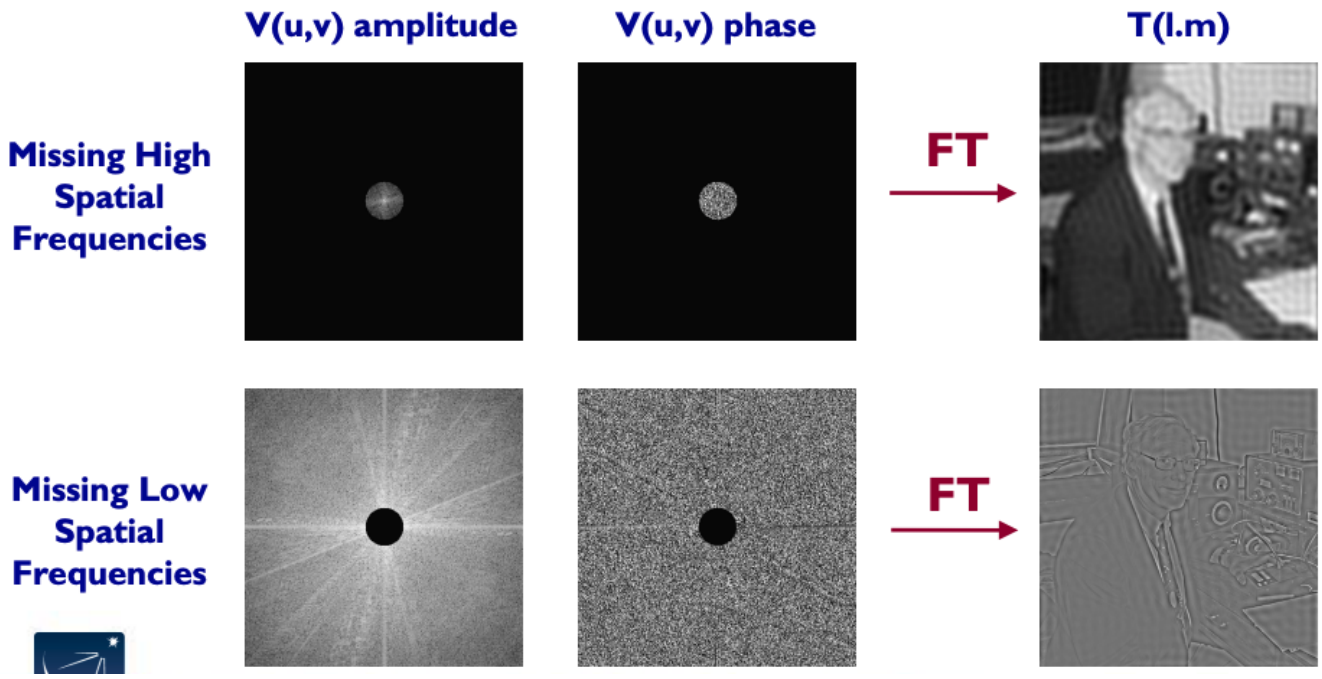


Figure 9.3: The effects of missing low and high spatial frequencies when reconstructing the original image with a Fourier transform.

In this way, a radio interferometer sort of acts like a spatial filter. The full range of angular sensitivities is

$$\frac{\lambda}{B_{\text{max}}} < \theta < \frac{\lambda}{B_{\text{min}}} \tag{9.13}$$

Any spatial scales smaller than the $\Delta\theta_{\text{min}}$ cannot be resolved, and any spatial scales larger than $\Delta\theta_{\text{max}}$ cannot be imaged[191,192,193].

Looking at an example, if we were observing in the $\lambda = 21.1$ cm band, then to produce images with a resolution of $\theta = 1''$, we would need a maximum baseline of

$$B_{\text{max}} \sim \frac{\lambda}{\theta} \sim 44\,\text{km} \tag{9.14}$$

Taken to the extreme, the Event Horizon Telescope (EHT) coordinates radio telescopes all across the globe to produce an effective maximum baseline the size of the entire Earth! It can achieve resolutions as low as 35 μas[194].

154. How does an X-ray telescope image? Why is it necessary to have the reflection take place at grazing angles of incidence?

How are X-ray photons collected?

For telescopes in the optical, we use mirrors to reflect incoming light rays and converge them into a single point, creating an image. However, this is not possible with X-rays because they are very energetic and will just pass straight through normal mirrors. Antiquated telescopes used lenses to refract the light into a single point, but this is also not possible with X-rays since their index of refraction in glass is very close to 1, meaning the lens would have to be *very* long for the X-rays to converge.

If the X-rays happen to strike the mirror at a *grazing incidence*, then they can still be reflected. There are a few different analogies one can use to understand this. Think of a rock and a lake. Drop the rock straight down, and it will drop straight into the water and sink down. But if you throw the rock parallel to the water's surface, it can skip a few times before falling in. Alternatively, think of a bullet and a wall. Shoot the bullet straight at the wall and it will punch through and leave a deep hole. Shoot it at an angle to the wall, and it will graze off and ricochet in another direction.

Quantitatively, the reflection probability depends on the angle of incidence θ_i and the indices of refraction of the two materials the light is being reflected off of (n_1 and n_2). The probability is given by the **Fresnel equations**

$$R_{\text{TE}} = \left| \frac{n_1 \cos\theta_i - n_2\sqrt{1-((n_1/n_2)\sin\theta_i)^2}}{n_1 \cos\theta_i + n_2\sqrt{1-((n_1/n_2)\sin\theta_i)^2}} \right|^2 \tag{9.15}$$

$$R_{\text{TM}} = \left| \frac{n_1\sqrt{1-((n_1/n_2)\sin\theta_i)^2} - n_2 \cos\theta_i}{n_1\sqrt{1-((n_1/n_2)\sin\theta_i)^2} + n_2 \cos\theta_i} \right|^2 \tag{9.16}$$

where "TE" and "TM" refer to the transverse electric and transverse magnetic polarizations. Solving for when R_{TE} and R_{TM} are 1 gives the *critical angle*[195]

$$\theta_c = \arcsin\left(\frac{n_2}{n_1}\right) \tag{9.17}$$

If $\theta \geqslant \theta_c$, then we will *only* have reflection. We can already see that since the refractive indices in the X-ray regime are both close or equal to 1, the critical angle for X-rays will have to be close to 90°. Let's say $n_1 = 1$ and $n_2 = 1-\delta$ where δ is small. Then let's call $\theta_r = \pi/2 - \theta_c$ the small angle complement to the critical angle. Then, using the small angle formula, we have

$$\sin\theta_c = 1-\delta \tag{9.18}$$

$$\cos\theta_r = 1-\delta \tag{9.19}$$

$$1 - \frac{\theta_r^2}{2} \approx 1-\delta \tag{9.20}$$

$$\theta_r \approx \sqrt{2\delta} \tag{9.21}$$

And δ can be estimated from dispersion theory, giving us

$$\boxed{\theta_r \approx \sqrt{4\pi r_e n_e \lambda^2}} \tag{9.22}$$

where r_e is the classical electron radius and n_e is the electron number density of the material[196].

As a consequence, X-ray telescopes are typically built with a series of mirrors that successively reflect X-rays into a convergence point. This type of telescope is called a **Wolter telescope**. See the setup in Figure 9.4. Having multiple reflecting mirrors in sequence like this allows the field of view to be increased without making the telescope too long. The mirrors are also arranged in a cylindrical pattern (imagine rotating the mirrors in the Figure along a horizontal axis that goes through the center of the configuration).

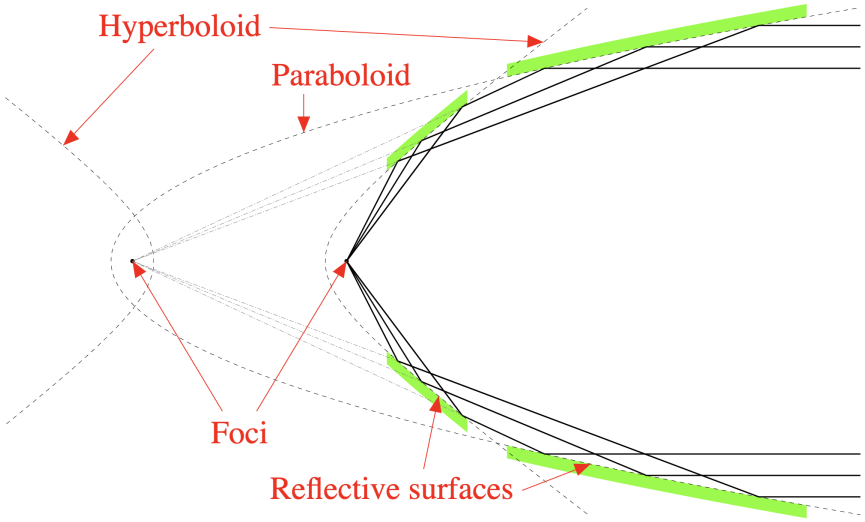


Figure 9.4: The mirror setup for a Wolter telescope of type 1.

In addition, since the grazing incidence is so shallow, mirrors can be stacked on top of each other in an annular configuration to further increase the amount of light that can be collected. In the *Chandra* X-ray telescope, for instance, there are 4 layers of these mirrors. See Figure 9.5.

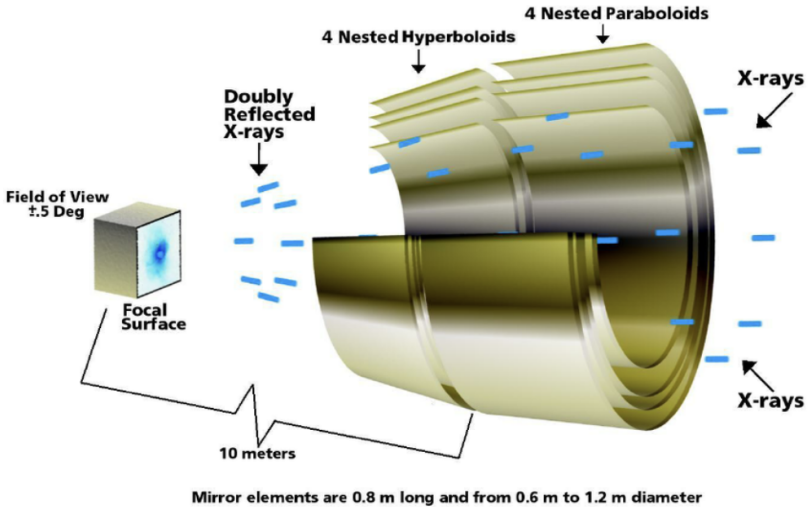


Figure 9.5: The *Chandra* X-ray telescope's mirror configuration.

Notice that the maximum angle for reflection $\theta_r \propto \sqrt{n_e}$. In other words, materials with higher densities will be more reflective. Thus, it is a common choice to use high-Z metals such as gold or iridium to coat the mirrors of X-ray telescopes[196].

The X-rays are typically measured using CCDs (see §9.Q157). For higher energy resolution, one can also use **grating spectrometers**, where a transmission or a reflection grating diffracts X-rays. These are used on *Chandra* (LETGS, METGS, and HETGS—PI: Claude Canizares), and *XMM-Newton* (RGS), with grating typically having $\sim 10^3$–10^4 lines/mm. This is good for studying narrow spectral features like lines.

Another option is **microcalorimeters**. These detect a change in temperature due to the arrival of a single X-ray photon. This works by using a material with a resistance that changes rapidly near some critical temperature where the pixel becomes a superconductor. This provides excellent energy resolution of a few eV. These are used on the recently launched *XRISM* mission[197].

155. How would you calculate the "signal-to-noise ratio" for a standard ground-based telescope performing CCD imaging? What terms dominate at optical versus infrared wavelengths? Why, when observing objects that are fainter than physical foregrounds like the night sky, does SNR increase as sqrt(time)?

The **signal-to-noise ratio**, as the name suggests, is just the ratio of the signal over the noise. Say we observe a source with N_* photon counts. Instead of looking at the total counts N_* over the whole exposure, let's look at the count *rate* R_* such that $N_* = R_* t$ with an exposure time of t. Then, we have the following sources of noise:

1. **Source noise**: The poisson noise from the source

   Poisson noise goes like $\sqrt{N}$, so $\sigma_* = \sqrt{R_* t}$.
2. **Sky noise**: The poisson noise from the sky background

   If each pixel has a sky count rate R_{sky} and our aperture covers n_{pix} pixels, then $\sigma_{\text{sky}} = \sqrt{n_{\text{pix}} R_{\text{sky}} t}$.
3. **Dark current**: The poisson noise from the CCD dark current (thermally excited electrons)

   If each pixel has a dark current of R_{dark}, then $\sigma_{\text{dark}} = \sqrt{n_{\text{pix}} R_{\text{dark}} t}$.
4. **Read noise**: The noise from reading out the CCD, independent of exposure time

   We'll call the read noise per pixel σ_{RN} such that the total read noise is $\sigma_{\text{RN,tot}} = \sqrt{n_{\text{pix}} \sigma_{\text{RN}}^2}$.

We add all of the errors in quadrature, so the total signal to noise ratio is

$$\boxed{\text{SNR} = \frac{R_* t}{\sqrt{R_* t + n_{\text{pix}} R_{\text{sky}} t + n_{\text{pix}} R_{\text{dark}} t + n_{\text{pix}} \sigma_{\text{RN}}^2}}} \tag{9.23}$$

In the optical, the limiting noise term is usually the source noise, which is the ideal scenario one wants to be in since we inherently cannot get the noise any lower than this. In the IR, however, the sky background becomes significant, as well as the dark current. The read noise should never be dominant. Let's look at some of these limiting cases[198]

- Source-dominated:

$$\boxed{\text{SNR} \simeq \sqrt{R_* t} \propto \sqrt{t}} \tag{9.24}$$

- Sky/dark-dominated ($R_{\text{sky}} \sim R_{\text{dark}} \gg R_*$):

$$\boxed{\text{SNR} \simeq \frac{R_* t}{\sqrt{n_{\text{pix}} (R_{\text{sky}} + R_{\text{dark}}) t}} \propto \sqrt{t}} \tag{9.25}$$

- Read noise-dominated:

$$\boxed{\text{SNR} \simeq \frac{R_* t}{\sqrt{n_{\text{pix}} \sigma_{\text{RN}}^2}} \propto t} \tag{9.26}$$

156. What is the diffraction limit for a telescope? Why does the signal-to-noise ratio scale like D^4 for diffraction limited imaging on ground-based telescopes with diameter D at IR wavelengths?

The Rayleigh Criterion[199]

The **diffraction limit** for a telescope is set by the size of the *diffraction pattern* created when light from a point source passes through a circular aperture. First, let's look at the simplified case of a 1D slit rather than a circular aperture. Light coming into the slit is essentially all parallel, but then on the other side of the slit it spreads out in all directions equally. Then, if we look at some point P on a screen on the other side of the aperture, at an angle θ away from the center, two light rays will be out of phase by the length difference between their two paths, given by $x\sin\theta$ (see the geometry in Figure 9.6). Then, all we have to do is integrate this over all x from $-D/2$ to $D/2$ to get the total electric field at angle θ:

$$E(\theta) = Ae^{-i\omega t}\int_{-D/2}^{D/2} \mathrm{d}x\, e^{ikx\sin\theta} \tag{9.27}$$

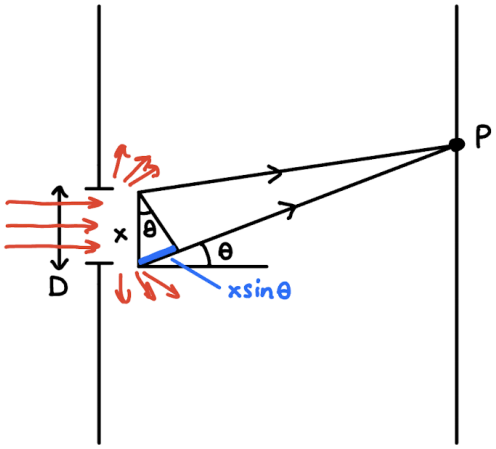


Figure 9.6: The interference of two light beams incident on a small slit of width $b = D$ at an angle θ.

Actually doing the integration reveals that the electric field is a sinc function, and the intensity, $I = \langle E^2\rangle$, is

$$I(\theta) = I_0 \mathrm{sinc}^2\left(\frac{1}{2}kb\sin\theta\right) \tag{9.28}$$

Now let's upgrade to the case of a 2D circular aperture. We now have to integrate over the whole area of the aperture. At some separation x, we cover an area $\mathrm{d}A = 2\sqrt{R^2-x^2}\mathrm{d}x$, so the integral becomes

$$E(\theta) = Ae^{-i\omega t}\int_{\text{area}} \mathrm{d}A\, e^{ikx\sin\theta} = 2Ae^{-i\omega t}\int_{-R}^{R} \mathrm{d}x\,\sqrt{R^2-x^2}e^{ikx\sin\theta} \tag{9.29}$$

See the new geometry of the aperture in Figure 9.7.

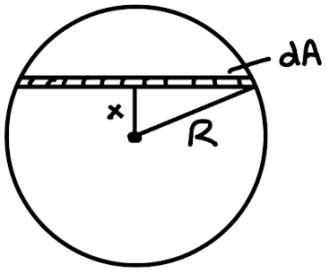


Figure 9.7: The geometry of diffraction in the case of a circular aperture.

Doing the new integral reveals that the electric field looks like a *Bessel function of the first kind*, and the intensity becomes

$$I(\theta) = I_0\left(\frac{2J_1(\frac{1}{2}kD\sin\theta)}{\frac{1}{2}kD\sin\theta}\right)^2 \tag{9.30}$$

This intensity profile is known as an **Airy disk**. So, how much does the light from a point source get spread out by the diffraction? We can find out by looking at where $I(\theta)$ goes to 0, which will happen when the bessel function J_1 goes to 0. The zeros of the Bessel functions are well know, and the first one happens at 3.8317. Therefore, this corresponds to an angle of

$$\frac{1}{2}kD\sin\theta = 3.8317 \quad\longrightarrow\quad \sin\theta = 2\frac{3.8317}{kD} \tag{9.31}$$

Replacing $k = 2\pi/\lambda$, we obtain

$$\sin\theta \approx 1.22\frac{\lambda}{D} \tag{9.32}$$

In the small angle limit, this becomes

$$\boxed{\theta \approx 1.22\frac{\lambda}{D}} \tag{9.33}$$

This is the **Rayleigh criterion**, otherwise known as the **diffraction limit**. Any two point sources separated by an angle smaller than this θ will not be resolvable. Note that in the real world, other effects such as atmospheric turbulence cause point sources to be smeared out further, increasing the minimum resolvable angular difference above the diffraction limit. The actual minimum resolvable θ is often referred to as the **seeing**, which depends on many factors like the humidity, temperature, airmass (how much of the atmosphere are you looking through—higher for lower altitudes), etc. One usually needs to go to outer space to obtain diffraction-limited seeing.

Scaling of the SNR

At IR wavelengths for faint sources, as discussed in §9.Q155, the sky noise term dominates the SNR, so we can write the SNR as approximately

$$\text{SNR} \simeq \frac{R_* t}{\sqrt{n_{\text{pix}} R_{\text{sky}} t}} = \frac{R_*}{\sqrt{n_{\text{pix}} R_{\text{sky}}}}\sqrt{t} \tag{9.34}$$

The count rates R_* and R_{sky} (counts per unit time) are flux × area, so they increase with the collecting area of the telescope, i.e. D^2. However, if our seeing is diffraction-limited, then the FWHM of the PSF *decreases* according to (9.33). This means that the number of pixels covered by our source decreases like θ^2, i.e. $n_{\text{pix}} \propto D^{-2}$. This exactly cancels out with the D^2 scaling of R_{sky} in the denominator, leaving our SNR to scale like

$$\boxed{\text{SNR} \propto D^2\sqrt{t}} \tag{9.35}$$

So, if we have some target SNR we are trying to reach, the *exposure time* that is required to reach this SNR scales inversely with the diameter like

$$\boxed{t \propto \text{SNR}^2 D^{-4}} \tag{9.36}$$

The question is confusingly worded and seems to imply that the SNR itself should scale with D^4, but this is not true—as we have seen, it scales like D^2. I have confirmed this with multiple sources: Chris Layden and Kevin Burdge both agree, as well as Ref. 200 (see the solution to problem 2b).

Note: if we're *seeing limited* rather than diffraction limited, then θ does not change as a function of D, so the number of pixels our source covers n_{pix} remains constant. In this case, the scaling becomes only linear with D

$$\text{SNR} \propto D\sqrt{t} \tag{9.37}$$

157. Briefly describe the following kinds of astronomical instruments and their purpose: Schmidt camera, Ritchy-Chrétien telescope, multi-object spectrograph, echelle spectrograph, CCD, coronagraph, laser interferometer, adaptive optics

I. Schmidt Cameras[199,201]

Also known as a **Schmidt telescope**, the Schmidt camera is an efficient telescope design that uses a spherical primary mirror in combination with a corrective lens called a Schmidt lens that removes the *spherical abberation* from the primary mirror.

When incoming light rays are all parallel onto a mirror, in order to focus them all into a single focal point, the mirror shape actually has to be *parabolic*. If the mirror is spherical, the light rays will not all hit the same focal point, and this is what introduces the spherical abberation. So, why not just build telescopes with parabolic mirrors? Well, the problem is that parabolic mirrors are much harder to manufacture than spherical mirrors. So, the Schmidt telescope takes a different approach by using a lens on the incoming light rays which bends them in such a way that, once they are reflected off the spherical mirror, they *do* converge onto a single focal point. The setup is shown in Figure 9.8.

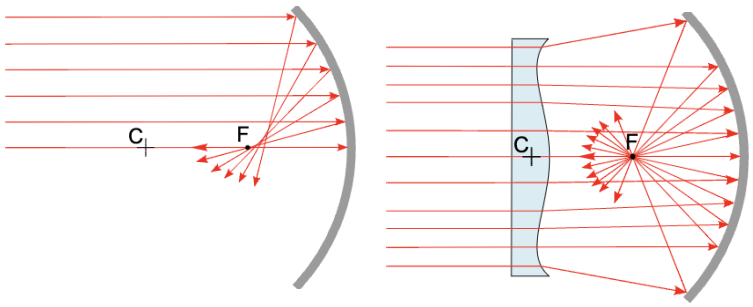


Figure 9.8: *Left*: Incoming parallel light rays onto a spherical mirror don't converge on a single focal point, producing spherical abberatin. *Right*: The corrective Schmidt lens bends the light in such a way that it cancels out with the spherical abberation.

The ability to use spherical mirrors allows these telescopes to be built with larger mirror sizes and larger fields of view. However, the focal plane is typically strongly curved, so any photographic plates or CCDs used must be likewise curved. A variant design, the **Schmidt-Cassegrain telescope**, puts a secondary mirror at this focal plain that focuses the light through a hole at the center of the primary mirror, reducing the need for a curved detector. See these designs in Figure 9.9.

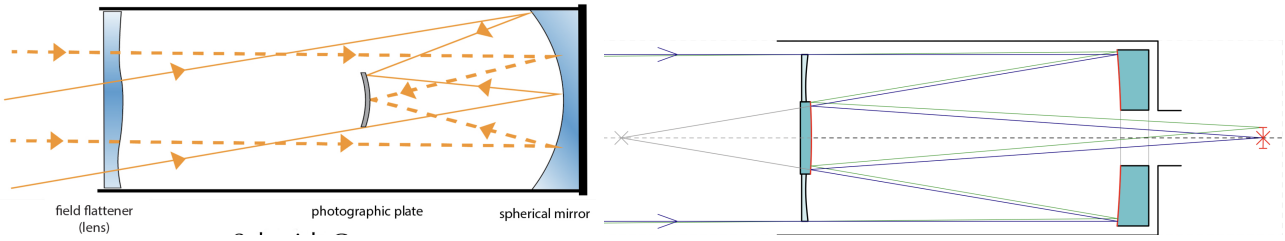


Figure 9.9: *Left*: A typical design for a Schmidt telescope with a curved photographic plate. *Right*: A typical design for a Schmidt-Cassegrain telescope with a hole in the primary mirror.

Examples: *Kepler*, Palomar observatory (ZTF), and Calar Alto observatory.

II. Ritchey-Chrétien Telescopes[202]

The Ritchey-Chrétien telescope is a telescope design in the Cassegrain family that used a hyperbolic primary and secondary mirror that are designed to reduce off-axis abberation (called **coma**).

As mentioned in the previous section, a *parabolic* primary mirror shape is required to focus parallel light rays into a single focal point with no abberation. However, this is only true if the parallel light rays are also parallel to the axis of the parabola. If they are coming in at an angle, they are no longer reflected into a single point. This causes point sources that are off-centered to appear to have tails like a comet (a coma). See Figure 9.10 for a representation of this effect—it is shown for a lens, but the principle is the same with a mirror.

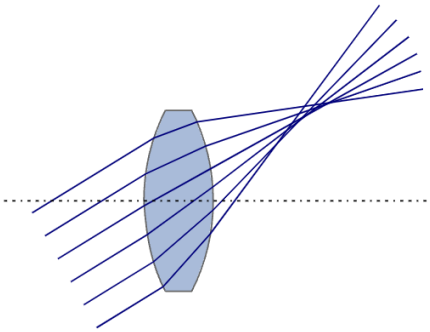


Figure 9.10: The coma abberation for parallel light rays coming in at an angle to a parabolic lens.

This effect can be corrected by using hyperbolic mirrors with specific curvatures, as shown in Figure 9.11.

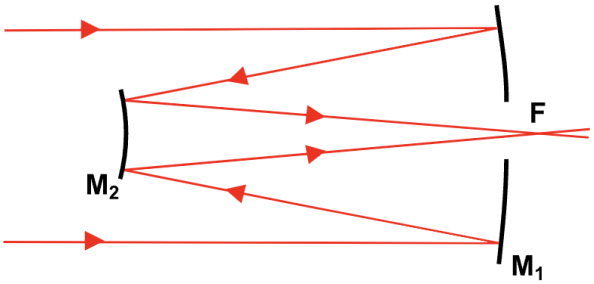


Figure 9.11: The design of a Ritchey-Chrétien telescope.

These telescopes are also often built with a Cassegrain configuration (see the section on the Schmidt camera). This design also allows telescopes to be built with larger and heavier mirrors, since it becomes easier to balance them by putting all of the heavy detectors and machinery underneath the primary mirror.

Examples: *HST*, Keck observatory, the SDSS telescope at Apache Point observatory, and the George Mason University telescope.

III. Spectrographs[203,204,205,206]

The question specifically asks about Multi-Object Spectrographs and Echelle Spectrographs, but first let's just go over what exactly a spectrograph is and what it does, because there are many different types.

Long-Slit Spectrographs

A long-slit spectrograph using a 1-dimensional slit aperture to collect light. The light is collected from a single point along the slit at a time—first it is reflected through a mirror that collimates the light coming from this point. Then, the collimated light hits a diffraction grating that disperses it into its individual wavelengths (the light reflects and interferes with itself in such a way that it cancels out all but one wavelength). Then, the dispersed light is collected with a CCD to create a 1D spectrum. Then, to get spectra for different points along the slit, we can rotate the diffraction grating to cause it to collect light from different points. This is all shown visually in Figure 9.12.

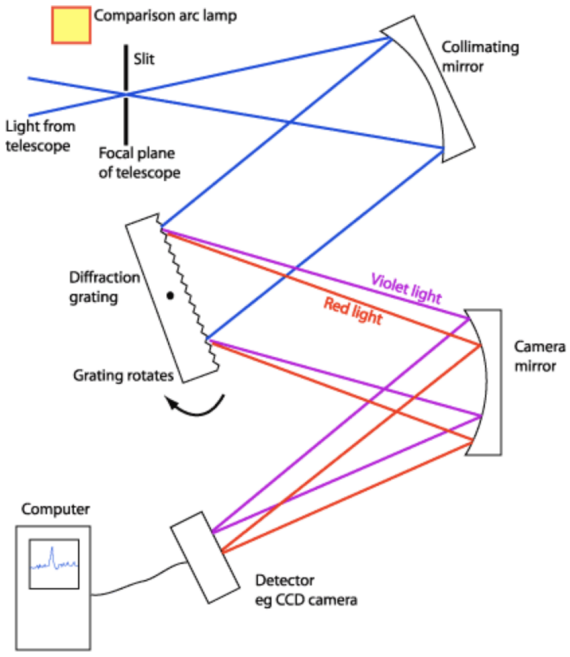


Figure 9.12: The setup of a long-slit spectrograph.

The resulting data is a 2D image—one dimension corresponds to the single spatial dimension of the slit, and the other dimension corresponds to the wavelength of light.

Examples: Kitt Peak's RC spectrograph, WHT's ISIS

Slitless Spectrographs

You may ask, why don't we just make our aperture 2D like a normal photometric observation to get spectral information over the whole field of view? The problem with this is that, without the restrictions of the slit, the dispersed light from the diffraction grating ends up getting blended together over different spatial locations (this is mostly a problem for extended sources or crowded fields of view). However, slitless spectrographs like this *do* exist and do have some use cases.

Examples: *JWST*'s NIRISS

Multi-Object Spectrographs

Another technique used to obtain spectra makes use of *optical fibers* at the focal plane instead of a slit. This allows us to isolate individual points in the image and obtain 1D spectra from them. We just place an optical fiber at the location of each object we want a spectrum from, and feed them all into the spectrograph along 1 dimension, just like with the slit. With this technique, we can obtain

spectra from *many* objects all simultaneously! However, we lose the spatial information—we just obtain a single 1D spectrum for each object.

Examples: SDSS, *JWST*'s NIRSpec, *HST*'s NICMOS, Gemini's GMOS

Integral Field Spectrographs

Looking for a way to get spectral information over a full 2D field of view, without compromising the blending that comes with slitless spectrographs? Integral Field Spectrographs have got you covered. These basically take a 2D image and split it up into smaller chunks—whether that be a series of 1D slits, or a grid of individual points—before feeding each chunk into a spectrograph. There are a few different approaches people take to do this. The most common one is to use an image slicer, which is basically a series of mirrors at slightly different angles that splits the image up into 1D slices. Each slice acts like a slit and can be fed into a long-slit spectrograph. See a diagram in Figure 9.13. The image slicing optics are typically referred to as an **Integral Field Unit (IFU)**.

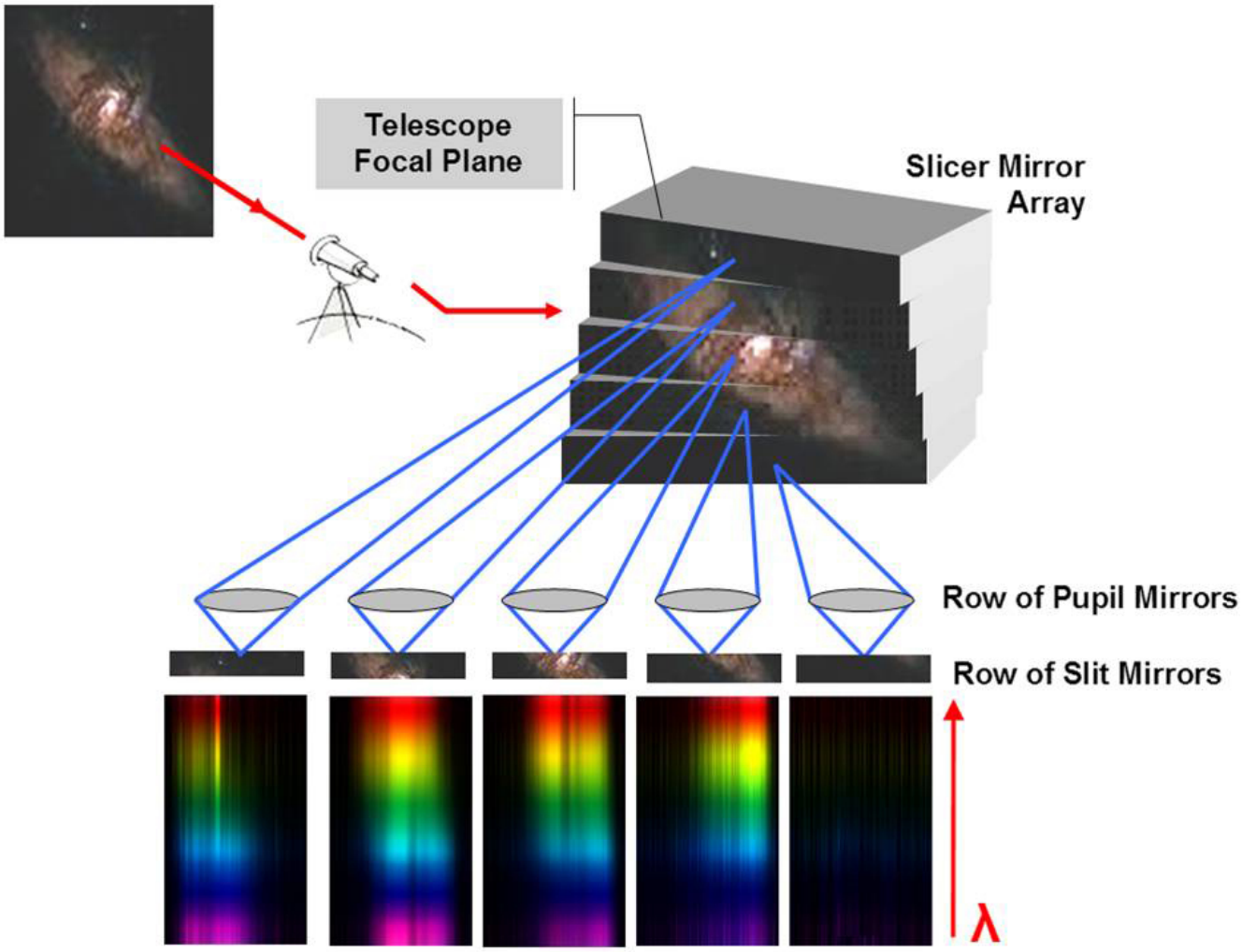


Figure 9.13: The instrumentation for an integral field spectrograph.

The result is a 3D data cube with 2 spatial dimensions and 1 wavelength dimension. The complex image slicing optics typically requires us to sacrifice spatial resolution for spectral resolution. For example, an IFU using this technique will rarely have more than a 50x50 pixel grid. However, another

technique uses an array of optical fibers for each pixel instead of an image slicer, which allows for a finer spatial resolution (this technique is used by MUSE).

Examples: *JWST*'s MIRI/MRS, Gemini's GMOS, Keck's KCWI, ESO/VLT's MUSE

Echelle Spectrographs

An echelle spectrograph uses an echelle (French for "ladder") diffraction grating, which has a low groove density and is optimized for high incidence angles and therefore high diffraction orders. At these high orders, the orders significantly overlap with each other (for example, at one point you might have order 1 light at 800 nm, order 2 light at 400 nm, order 3 light at 266.6 nm, etc.). To separate these orders, we then disperse the light in an *orthogonal* direction, typically using a prism (which disperses the light by using the fact that the index of refraction is wavelength-dependent). The result is that the spectral orders are stacked on top of each other, and we get an extremely high-resolution spectrum. The trade-off is that this design reduces the throughput of the spectrograph—less light is going onto each individual pixel on the detector. See the setup in Figure 9.14.

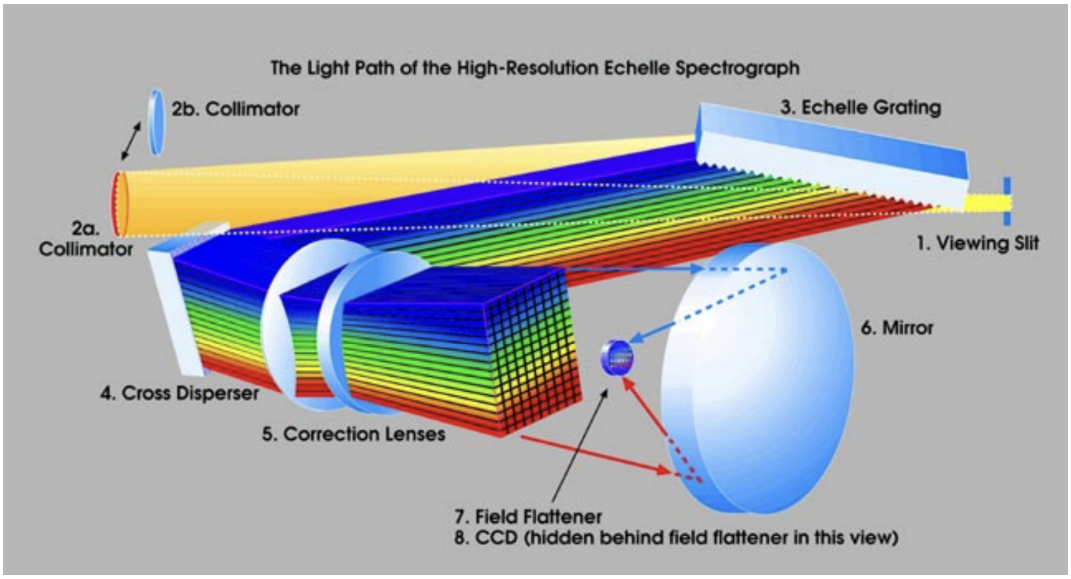


Figure 9.14: The setup of an echelle spectrograph.

Examples: La Silla's HARPS, Keck's HIRES and iSHELL

IV. CCDs[197]

Once the photons are collected, they are usually measured with a **charge-coupled device (CCD)**—these are used in the optical, UV, IR, and X-ray. This device essentially works by converting photons into electrons.

In a solid, electrons don't have distinct energy levels but rather they have allowed and forbidden *bands* of energy. In a conductor, the lower energy band is not completely filled, so electrons can freely move and conduct electricity. In an insulator, the lower energy band is filled, so electrons cannot move and cannot conduct electricity. A CCD is made of a **semiconductor** where the lower energy band is filled, but the band gap (AKA the forbidden region) between the lower and higher energy bands is small enough that an electron can jump up to the unfilled energy band by absorbing an incoming photon. See Figure 9.15.

CCDs made with Silicon don't work for wavelengths $\gtrsim 1.1$ μm because photons above this limit don't

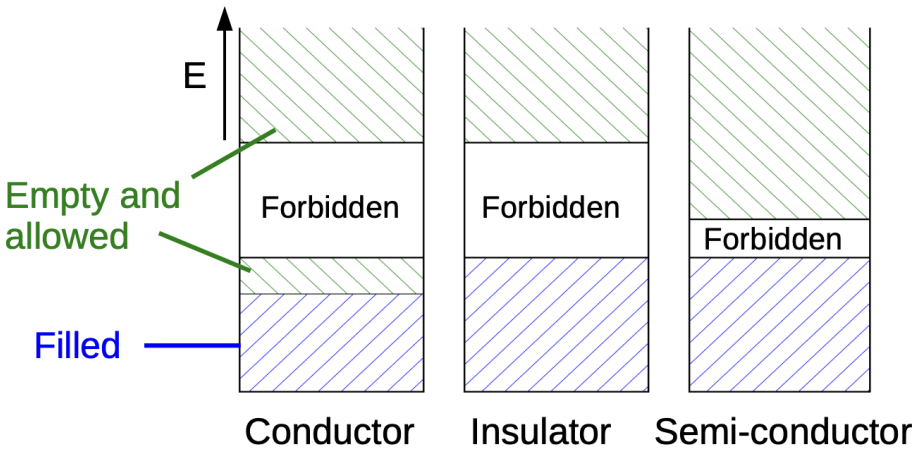


Figure 9.15: The different configurations of a conductor, insulator, and semiconductor[197].

have enough energy to push electrons into the upper energy level. Electrons can also be excited by thermal energy, producing intrinsic noise (called **dark current**), so CCDs are often cryogenically cooled to reduce this noise as much as possible.

This whole process is subdivided into individual elements called pixels. In each pixel, electrons are excited and held in place by a small electric field created by electrodes. Then, when the image is ready to be read out, the electrodes rapidly switch between positive and negative to transfer the electrons from pixel to pixel and collect them at the read-out location. This is first done along a single row, then all the rows are moved down and the next row is read out, rinse and repeat until all rows have been read out. See this all in Figure 9.16.

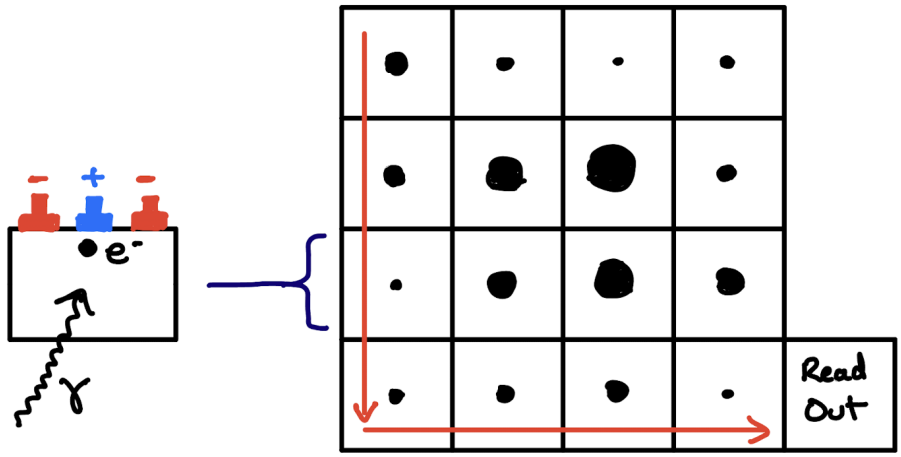


Figure 9.16: *Left*: A single CCD pixel where an electron can be excited by a photon and held in place by a positive electrode and surrounding negative electrodes. *Right*: A grid of pixels on a CCD with different numbers of electrons. During read out, the bottom row is transferred to the right. Then all of the rows are shifted down, and the new bottom row is shifted to the right again. This is all done by varying the electric fields in the electrodes[197].

There is one major difference between CCDs in the optical/UV/IR and in the X-ray. In the other wavebands, 1 photon excites 1 electron, so the number of electrons is equal to the number of photons.

However, in the X-ray 1 photon has enough energy to excite many electrons in the CCD, so the charge on the pixel is proportional to the *energy* of the photon. Even the brightest sources in the sky release much fewer photons in the X-ray than in the optical (about 0–1 per pixel every second), so we can literally count individual photons in the X-ray and record each of their energies. However, if the source is bright enough in the X-rays that we receive multiple photons in 1 pixel, a condition called **pile up**, we have an ambiguity between an event with *multiple, lower energy* photons vs. an event with *one, higher energy* photon.

V. Coronagraphs[207]

A coronagraph is a small attachment to a telescope that uses a disk to block out bright objects in the image—for example, stars or AGN. This allows us to resolve and study faint sources in the immediate surroundings of these objects without them being washed out by the bright PSF of the nearby point source. These are useful in studying the corona of stars (hence the name), directly imaging exoplanets around stars, and studying the host galaxies of AGN. See an example in Figure 9.17

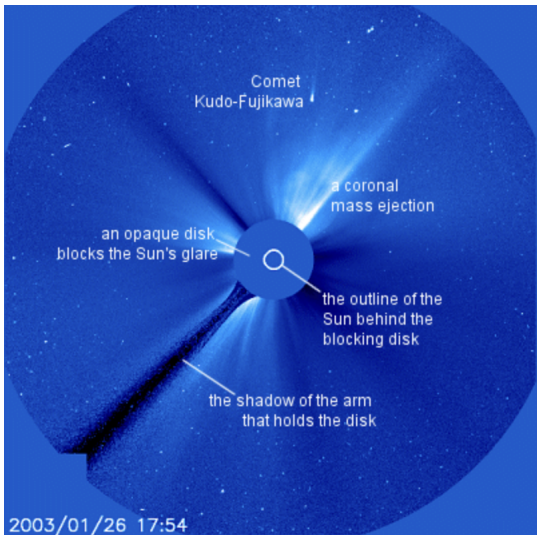


Figure 9.17: A schematic of a coronagraph being used to study the solar corona.

Examples: *HST*'s NICMOS, *JWST*'s NIRCam and MIRI

VI. Laser Interferometers[208]

A laser interferometer, as the name suggests, measures the interference pattern of two light beams arranged perpendicularly to each other. An incoming light beam is split evenly into each direction before being reflected back and collected at a detector. If the two paths the light traveled are exactly the same length, the output should be a coherent pattern. However, if the path lengths are slightly different, the light will interfere with itself. This is the setup that was used in the famous Michelson-Morley experiment that disproved the existence of the aether. See the setup in Figure 9.18. Today, interferometers are used to detect gravitational waves, since the stretching and compressing of spacetime will cause one of the paths to become shorter/longer than the other.

Examples: LIGO.

VII. Adaptive Optics[209]

Adaptive Optics (AO) refers to a technique where the effects of atmospheric turbulence, which blur and smear an image (causing stars to "twinkle", an effect known as **scintillation**), are corrected for

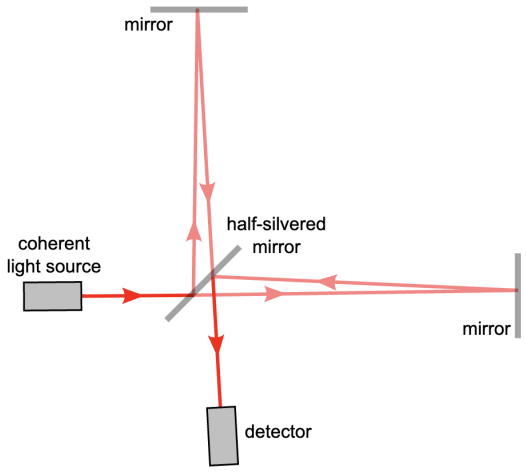


Figure 9.18: A simple laser interferometer.

in real time by deforming the telescope mirror in such a way that the turbulence effects cancel out. Typically, an artificial laser guide star is used as a point of reference for the AO system. An example of such a system is shown in Figure 9.19. This often allows us to obtain diffraction-limited seeing with ground-based telescopes.

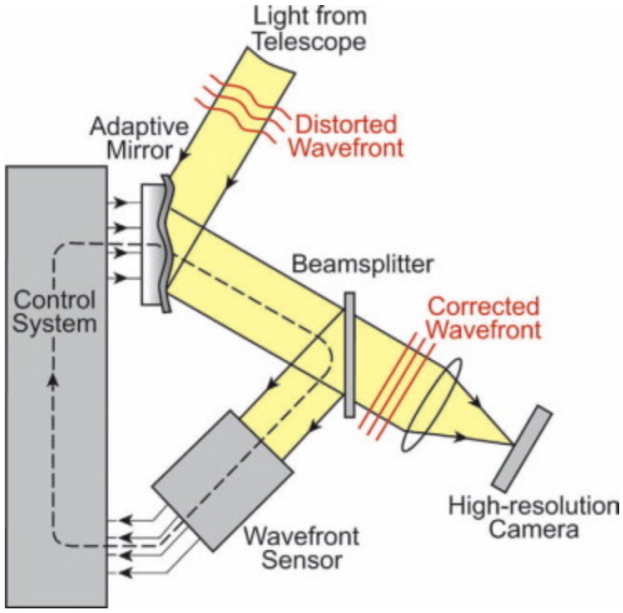


Figure 9.19: A typical AO system.

Examples: VLT, Keck

158. What limits the angular resolution of ground-based telescopes at 1 μm, 10 μm, or 100 m?

Recall the Rayleigh criterion for the diffraction-limited case (§9.Q156):

$$\theta = 1.22\frac{\lambda}{D} \tag{9.38}$$

- **1 μm** - Seeing limited

  Using a typical ground-based optical telescope size of $D = 10$ m, the diffraction limit here is $\theta = 0\farcs025$. The typical seeing limit due to atmospheric turbulence is much higher than the diffraction limit, $\theta \simeq 0\farcs5$. We can use AO to somewhat mitigate these effects (§9.Q157), but the seeing will still typically be larger than the diffraction limit.

- **10 μm** - Seeing limited

  Once again using $D = 10$ m, the diffraction limit in the MIR is $\theta = 0\farcs25$. This is still lower than the seeing limit from atmospheric turbulence, which is independent of wavelength, $\theta \simeq 0\farcs5$.

- **100 m** - Limited by the fact that we *literally cannot see* 100 m light from the ground

  See atmospheric transmission in §9.Q161. $\theta \to \infty$. I'm not sure why they even asked about this one. Maybe they meant 10 m?

- **10 m (BONUS)** - Diffraction limited

  I might as well cover this case too just in case. Using a typical radio telescope diameter of $D = 100$ m, the diffraction limit is $\theta = 7°$, much larger than the seeing limit from atmospheric turbulence. So in the radio the situation is reversed and we are diffraction limited. We can use radio interferometry to increase our effective telescope size and push down the angular resolution to sub-arcsecond (§9.Q153).

159. What are "apparent magnitude," "absolute magnitude" and "bolometric magnitude"? What are U, B and V colors?

Note: here I just write a quick review for each of these quantities. See the appendix section §12.11.16 for a full overview of magnitudes.

Magnitudes are a logarithmic measurement of the flux of an object relative to some zero point. They are defined such that a *smaller* magnitude corresponds to a *brighter* object. And yes, they can be negative.

Apparent Magnitude

The magnitude of the source as viewed from Earth. Since flux falls off with distance squared, the apparent magnitude also gets dimmer for more distant objects.

$$\boxed{m - m_0 = -2.5 \log_{10}\left(\frac{F_\nu}{F_{\nu,0}}\right)} \tag{9.39}$$

where m_0 and F_0 are the apparent magnitude/flux of the zero point. The most commonly used scales are Vega magnitudes, where the zero point is the magnitude of Vega, and AB magnitudes, where the zero point is 3631 Jy.

Absolute Magnitude

The magnitude of the source as it *would be* if it were viewed at a distance of 10 pc. This is kind of like the magnitude version of the luminosity, and is a more intrinsic quantity of the object than the flux. An apparent magnitude is converted to an absolute magnitude using the *distance modulus*

$$\boxed{m - M = 5 \log_{10}\left(\frac{d}{10\,\text{pc}}\right)} \tag{9.40}$$

Bolometric Mangitude

The magnitude of an object integrated over all frequencies/wavelengths. We typically convert magnitudes from one filter into bolometric magnitudes using a *bolometric correction* factor, i.e. for band V

$$\boxed{m_{\text{bol}} = m_V + BC} \tag{9.41}$$

Colors

U, B, and V are just the magnitudes in the U ("ultraviolet"), B ("blue"), and V ("visual") filters, which cover wavelengths:

- U: 3650 ± 680 Å
- B: 4400 ± 980 Å
- V: 5500 ± 890 Å

See §12.6 in the appendix for a full list of commonly used filters. The differences between two magnitudes, i.e. $B - V$, $U - B$, etc., are called *colors*.

160. Briefly describe the main objectives and capabilities of the following astronomical observatories *Voyager*, *HST*, Magellan telescopes, *Chandra*, *WMAP*, MWA, *SWIFT*, *Planck*, ROSAT, *Spitzer*, *TESS*, LIGO, *LISA*, NICER, CHIME, HERA, WINTER, SPT, *JWST*, *Roman Space Telescope*, Rubin Observatory

I. *Voyager* [210]

Two space probes, *Voyager 1* and *Voyager 2*, were launched in 1977 to study and image the atmospheres and ring systems of the giant planets, Jupiter, Saturn, Uranus, and Neptune. After the success of their initial mission, they continued to probe the outer parts of the Solar System. In 2012, *Voyager 1* passed the heliopause and officially reached interstellar space, followed by *Voyager 2* in 2018. They have made the first direct measurements of the density of the ISM, and are the furthest man-made objects in space, reaching a distance of ~ 160 AU today. They carried onboard golden records which contained sounds, images, and mathematical knowledge of human society on Earth, just in case they ever happened to be discovered by aliens.

II. *Hubble Space Telescope (HST or Hubble)* [211,212]

NASA's flagship optical telescope launched in 1990 in a low Earth orbit. Has a 2.4 m mirror and covers the UV/optical/NIR. It has made major contributions to just about every field of astrophysics. Instruments:

- UV/Optical/NIR Imagers: ACS, WFC3, STIS
- UV Spectrographs: COS, STIS (+optical)

III. Magellan Telescopes[213]

Twin 6.5 m telescopes located at the Las Campanas Observatory in Chile's Atacama Desert. Began observations in 2000/2002. Instruments:

- Optical Imagers: IMACS, Megacam, PISCO
- NIR Imagers: FourStar, MagAO
- Optical Spectrographs: IMACS, MagE, LDSS-3, MIKE, M2FS, PFS
- NIR Spectrographs: FIRE

IV. *Chandra X-ray Observatory (Chandra)*[214]

NASA's flagship X-ray space telescope launched in 1999 in a high-altitude orbit (must be above Earth's atmosphere which absorbs X-rays). Sensitive to ~ 0.08–2 keV. Still has the best angular resolution of any X-ray observatory today ($0\farcs5$). Instruments:

- Imagers: HRC
- Spectrographs: ACIS (PI: Mark Bautz, MKI), HETGS (PI: Claude Canizares, MKI), LETGS

V. *Wilkinson Microwave Anisotropy Probe (WMAP)*[175]

Space probe mission launched in 2001 to measure the CMB as a follow-up to COBE. Measured anisotropies down to angular resolutions $\sim 10'$. Placed strong constraints on cosmological parameters (H_0, Ω's). Ran until 2010.

VI. *Murchisonn Wide-field Array (MWA)*[215]

A low-frequency (70–300 MHz) radio interferometer with a wide FOV ($\sim 30°$). Good for probing the epoch of reionization with H I 21 cm science. Also has an all-sky radio survey of extragalactic sources (GLEAM). Jackie Hewitt at MKI is one of the founders.

VII. *Neil Gehrels Swift Observatory (Swift)*[216]

A gamma-ray/X-ray/UV/optical space telescope launched in 2004, originally designed to rapidly respond to GRBs (within $\lesssim 90$ s) to observe them and their afterglows. Now used for a wide variety of science interests. Instruments:

- Gamma-ray/X-ray Imagers: BAT (15–150 keV, wide FOV to find GRBs)
- X-ray Spectrographs: XRT (0.3–10 keV)
- UV/Optical Imagers: UVOT (1700–6500 Å)

VIII. *Planck*[177]

An ESA space probe launched in 2009 to study the CMB at higher angular resolutions than *WMAP* (down to $\sim 5'$). Currently our best measurements of the CMB power spectrum and early-universe cosmological parameters come from *Planck*. Ran until 2013.

IX. *Roentgen Satellite (ROSAT)*[217]

An X-ray survey satellite built through a collaboration between Germany, the US, and the UK. Launched in 1990 and ran until 1999. Began with a 6-month all-sky survey and used its remaining time on pointed observations. Instruments:

- Imagers: PSPC-B/PSPC-C (low-res, $\sim 25''$), HRI (high-res, $\sim 2''$)

X. *Spitzer Space Telescope (Spitzer)* [218]

NASA's previous-gen flagship IR telescope launched in 2003 and ran until 2020. Had a diameter of 0.85 m. It ran out of coolant in 2009, after which many of the filters became non-operational.

- IR Imagers: IRAC, MIPS
- IR Spectrographs: IRS

XI. *Transiting Exoplanet Survey Satellite (TESS)* [219]

An optical/NIR (one 0.6–1 μm filter) space telescope launched in 2018 with 4 wide-FOV cameras that stare at one portion of the sky for a few months before moving on to the next portion (called a sector). It observes with a cadence of 10 mins. Designed to study transiting exoplanets, but also useful for variables and transients of all kinds.

The pixel size is large, meaning the angular resolution is not the best (22″ since the FOV is so large). This means *TESS* is a good tool for *finding* transiting exoplanet candidates, but to be confirmed they often need follow-up observations with ground-based telescopes that have higher angular resolutions. This is coordinated via the *TESS* Follow-up Observer's Program (TFOP) working group. PI: George Ricker, MKI.

XII. Laser Interferometer Gravitational-wave Observatory (LIGO) [220]

A gravitational wave detector that uses a laser interferometer (see §9.Q157, §9.Q163). Has two locations: Hanover, WA and Livingston, LA, with 4 km arms. First run focused on improving the instrumentation ran from 2002–2010. The first real science run, dubbed "advanced LIGO" (aLIGO), started listening in 2015 and made its first detection within days: GW150914, a BBH merger.

XIII. *Laser Interferometer Space Antenna (LISA)* [221]

A proposed ESA space-based gravitational wave observatory that would allow for arm lengths of $\sim 10^6$ km. This would allow it to have sensitivity in the mHz range, opening up observations of SMBH mergers and EMRIs. Planned to be launched sometime in the 2030s.

XIV. Neutron Star Interior Composition Explorer (NICER) [222]

An X-ray instrument installed onboard the *International Space Station (ISS)* that was launched in 2017. Sensitive to 0.3–12 keV, does timing and spectroscopy. Its primary goal is to constrain the neutron star equation of state. It measures hot spots on neutron stars as they rotate in and out of the LOS, which can constrain the neutron star radius and magnetic field structure (i.e. non-dipole). Experienced a visible light leak in 2023 that caused increase noise at soft energies (< 0.8 keV) when the *ISS* is on the day side of its orbit.

XV. Canadian Hydrogen Intensity Mapping Experiment (CHIME) [223]

A radio telescope that is *not* a traditional dish. Instead it's a series of half-cylinder shaped detectors that has no moving parts, which gives it a wider FOV (200 deg^2). Operates between 400–800 MHz and is focused on measuring the H I 21 cm line (1420 MHz) to map neutral hydrogen in the local universe (sensitive to $z \sim 0.8$–2.5), see §8.Q126. Also good for finding FRBs. Kiyoshi Masui's group at MKI is heavily involved.

XVI. Hydrogen Epoch of Reionization Array (HERA) [224]

An array of radio dishes in South Africa sensitive to 120–200 MHz, corresponding to H I 21 cm (1420 MHz) at a $z \sim 6$–13. The successor of MWA with more collecting area and more sensitivity. See §8.Q126. PI: Jackie Hewitt, MKI.

XVII. Wide-field Infrared Transient Explorer (WINTER) [225]

A NIR (0.9–1.7 μm) time-domain instrument that is deployed on a dedicated 1 m robotic telescope at Palomar Observatory in San Diego, CA. Will perform a seeing-limited NIR time domain survey focusing on following up LIGO gravitational wave events to look for r-process material, kilonovae, TDEs, and more. Uses novel InGaAs photodetectors (as opposed to the traditional HgCdTe).

Note: Don't get it confused with the *Wide-field Infrared Survey Explorer (WISE)*

XVIII. South Pole Telescope (SPT) [179]

A 10 m microwave/millimeter/sub-millimeter ground-based telescope located at the NSF Amundsen-Scott South Pole Station. Began observations in 2007. Good for constraining the CMB power spectrum at angular scales $\lesssim 5'$ and finding galaxy clusters using the SZ effect. Instruments:

- Millimeter Imager: SPT-3G (90, 150, 220 GHz)

XIX. *Jelescope Welescope Space Telescope (JWST)* [226]

NASA's current-gen flagship IR telescope launched on Christmas day, 2021. Has a diameter of 6.5 m and is sensitive to 0.6–28.3 μm over its different instruments. Orbits at the L2 Lagrange point of the Earth-Sun system.

- NIR Imagers: NIRCam
- NIR Spectrographs: NIRSpec, NIRISS/FGS (slitless)
- MIR Imagers: MIRI (imaging modes)
- MIR Spectrographs: MIRI (spectroscopy modes)

XX. *Nancy Grace Roman Space Telescope* (the telescope formerly known as *WFIRST*) [227]

An infrared telescope with a 2.4 m diameter (same as *Hubble*) that is planned to be launched around 2027. Will have a much wider FOV (28 $\deg^2$) and has the science goals of direct imaging exoplanets with coronagraphy, and advancing understandings of dark energy and cosmology. Instruments:

- IR Imagers: WFI, CGI (coronagraphs)

XXI. Vera C. Rubin Observatory (the telescope formerly known as LSST) [228]

An 8.4 m ground-based telescope currently under construction in Chile. Observations are expected to begin in early 2025. Will perform an all-sky survey that is expected to find objects down to 25th magnitude in a single visit (1000s of transients expected per night). Results will be significantly affected by the satellite constellations contained hundreds of thousands of satellites planned to be launched by companies like SpaceX, Amazon, OneWeb, E-Space, etc.

161. What are the pros/cons of having a space-based versus ground-based observatory? At what wavelengths can you do both? At what wavelengths can you only observe from space?

	Pros	Cons
Ground	•Easy to maintain/repair •Cheaper •Can build larger	•Atmospheric turbulence •Not all wavelengths accessible •Light pollution
Space	•Diffraction-limited seeing •All wavelengths accessible •No light pollution	•Hard to maintain/repair •More expensive •Must build smaller

Wavelengths accessible from the ground:

- Optical
- NIR and some MIR
- Sub-mm/radio

Wavelengths only accessibile from space:

- Gamma-rays*
- X-rays
- MIR/FIR
- Long-wavelength radio

These are shown more precisely in Figure 9.20.

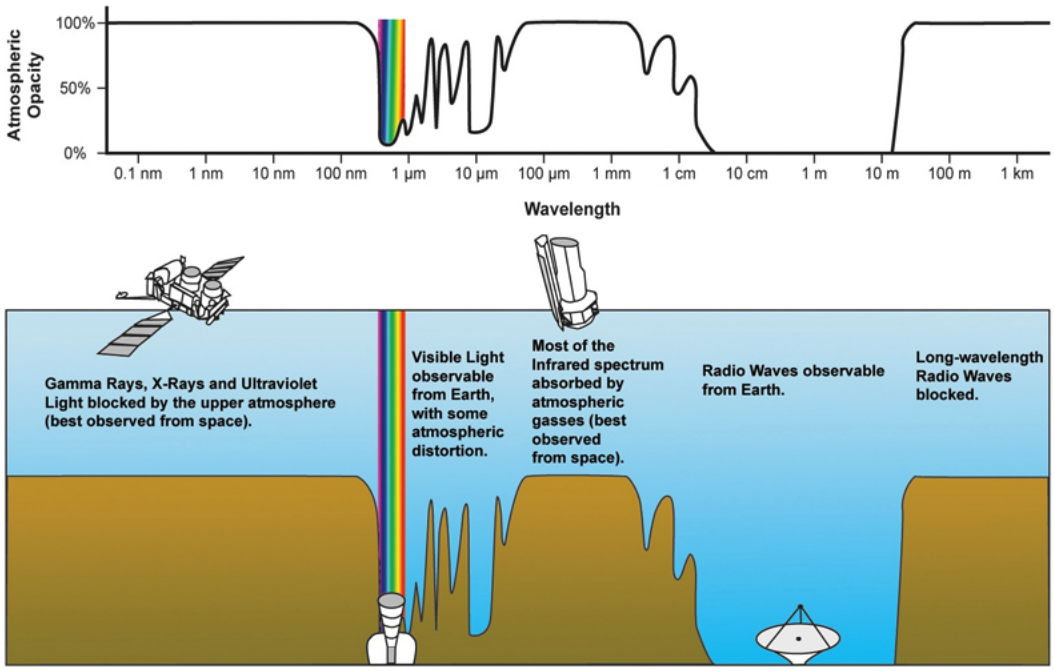


Figure 9.20: The opacity of Earth's atmosphere as a function of wavelength, showing the openings in the optical/NIR and sub-mm/radio[229].

*We can actually *indirectly* observe gamma rays from the ground using **Cherenkov Telescopes**. When gamma rays hit the upper atmosphere and interact with particles, they can create ultra-relativistic particles that travel faster than the speed of light in air (but still slower than the speed of light in a vacuum). This creates a flash of light that radiates in a cone (similar to a sonic boom when crossing the sound barrier). This is called **Cherenkov radiation**, and ground-based Cherenkov telescopes can detect this radiation and use its direction to pinpoint the astrophysical source of the gamma rays.

Examples: Cherenkov Telescope Array Observatory (CTAO), Major Atmospheric Gamma Imaging Cherenkov Telescope (MAGIC)

162. Derive the approximate (within a factor of a few) spectral resolution R required to detect a planet that perturbs its host star by a radial velocity amplitude of 10 cm/s. Describe some practical challenges with calibration and stability, and how they are addressed for such instruments. Why have we not yet achieved this precision in RV measurements of real stars?

The spectral resolving power is defined as

$$\boxed{R = \frac{\lambda}{\Delta\lambda}} \tag{9.42}$$

where $\Delta\lambda$ is the minimum wavelength difference for which two spectral line features are still resolvable by Rayleigh's criterion (i.e. the peak of one line must correspond to the first minimum of the neighboring line's peak)[9]. The Doppler shift, in the limit $\Delta v \ll c$, is approximately

$$z \approx \frac{\Delta v}{c} = \frac{\Delta\lambda}{\lambda} = \frac{1}{R} \tag{9.43}$$

So, we obtain

$$R \approx \frac{c}{\Delta v} \approx \frac{3 \times 10^{10}}{10} = 3 \times 10^{9} \tag{9.44}$$

However, this is the resolving power that would be required to resolve two close lines within 10 cm/s of each other. This is *NOT* the same thing as the resolving power that would be required to just measure the center of a line to 10 cm/s precision. The wavelengths are typically Nyquist sampled (i.e. $\geqslant 2$ pixels for each FWHM spectral resolution element $\Delta\lambda$), and in addition we can measure the center of a line by fitting a Gaussian model, which allows us to estimate the peak with sub-pixel precision depending on the errors of each data point we use (see Figure 9.21).

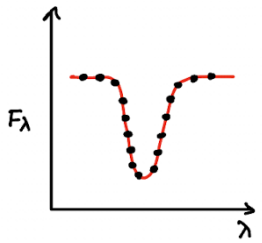


Figure 9.21: An absorption line modeled with a Gaussian profile, which resolves the peak location at a subpixel level of precision.

In general, the narrower the line itself is, the easier it is to measure the peak location (broad lines are so smeared out that the exact location of the peak becomes hard to distinguish), and having a higher SNR means the line is deeper, which also makes it easier. This gives an uncertainty σ_v

$$\sigma_v \sim \frac{\text{FWHM}}{\text{SNR}} \tag{9.45}$$

We can further lower this by measuring many spectral lines simultaneously. They all must have the

[9] See the end of this question for details

same spectral shift, so with N lines our measurement uncertainty goes down by a factor of $\sqrt{N}$:

$$\sigma_v \sim \frac{\text{FWHM}}{\sqrt{N}\text{SNR}} \tag{9.46}$$

For a non-rotating star, the FWHM of the lines will essentially just be the FWHM of the spectral resolution element as determined by the line-spread function (LSF), $\Delta v = c/R$. Then, a typical SNR ~ 100 and $N \sim 1000$, so

$$\boxed{\sigma_v \sim \frac{c}{R}\frac{1}{\sqrt{N}\text{SNR}}} \longrightarrow \boxed{R \sim 10^6 \text{ for } \sigma_v \sim 10 \text{ cm s}^{-1}} \tag{9.47}$$

For the highest resolution spectrographs today

$$R \sim 10^5 \longrightarrow \sigma_v \sim 1 \text{ m s}^{-1} \tag{9.48}$$

However, most stars *are* rotating, with typical speeds of at least ~ 2 km s^{-1} or more, in which case the uncertainty increases by an order unity factor ($c/R \to \sqrt{(c/R)^2 + v_{\text{rot}}^2}$). If $v_{\text{rot}} \gg c/R$, the scaling with R becomes slower, but not 0, since the SNR also increases as R increases (the lines get deeper as they become more resolved, SNR $\propto R^{0.5}$)[230,231,23210]

Stability and Calibration Challenges

At an $R \sim 10^5$, the spectral resolution element is $c/R \approx 3$ km s^{-1}, so when we measure differences of 1 m s^{-1}, we're measuring differences on the order of 1/1000th of a pixel. This means the instrument must be *extremely* stable against natural fluctuations:

- Pressure fluctuations $\longrightarrow$ use a vacuum chamber
- Temperature fluctuations on the mK level
- Must have no moving parts
- Must have undisturbed operations
- Do simultaneous calibrations
  - Use a gas cell or lamp with known emission lines to "anchor" the velocity shifts in the stellar spectrum to a known source. Downside: has some very bright lines that can saturate and bleed into the rest of the exposure.
  - Fabry-Pérot etalons create equally spaced and equally bright calibration lines. Downside: not a fundamental frequency tied to any atomic physics, just depends on the distance between two plates.

RV Measurements of Real Stars

Stellar variability is something we can't correct for with instrumentation, and is often much larger than the sub-m/s levels of precision we want to obtain. This encompasses the effects of stellar rotation (mentioned above), starspots rotating in and out of our FOV (which partially dim the side of the star rotating away from/towards us, creating a small blueshift/redshift), and granulation due to near-surface convection. There are many people working on analysis techniques to separate out the planet signals from the stellar signals. There are also *tellurics* (absorption lines from Earth's atmosphere) that further introduce uncertainty in our measurements by overlapping and blending with the lines

[10] Also see this helpful YouTube video

we actually care about. In the same vein, keep in mind that velocities are *relative*. This means that if we want to know the velocities of exoplanets around their host stars to within ∼10 cm/s of precision, then we also must know the velocity of Earth around the Sun and the velocity of the Sun around the galactic center to *at least* the same order of magnitude in precision.

Resolving Power for Gratings[199]

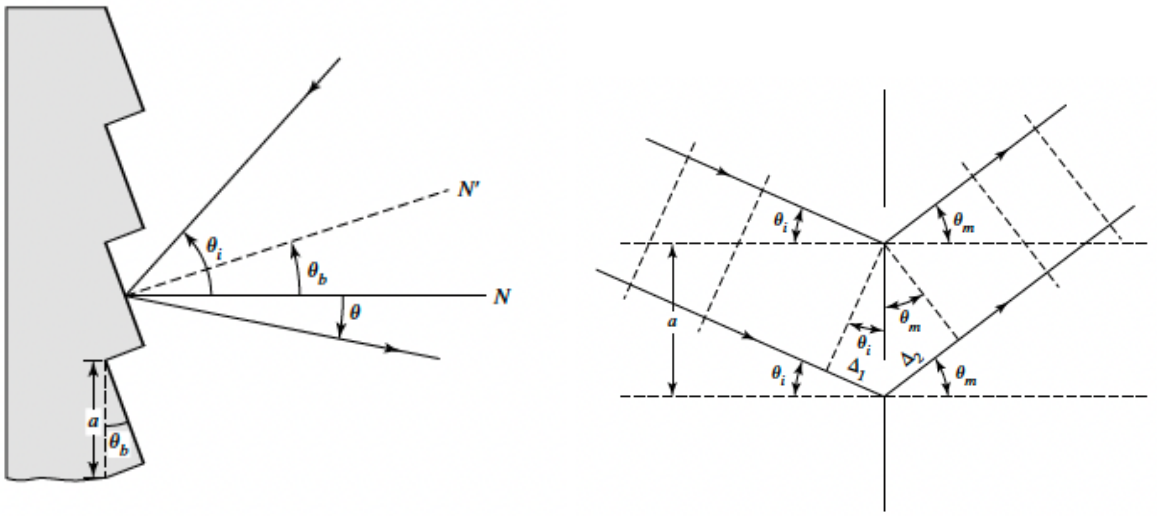


Figure 9.22: Light reflecting off of a diffraction grating[199].

The grating equation states that for a grating with a spacing a between each slit, the diffracted waves from two neighboring slits are in phase when

$$a(\sin\theta_i + \sin\theta_m) = m\lambda \tag{9.49}$$

See Figure 9.22 for an explanation of the symbols. If the waves are normally incident, $\theta_i = 0$, at a wavelength of $\lambda + \Delta\lambda$, then

$$a\sin\theta_m = m(\lambda + \Delta\lambda) \tag{9.50}$$

To satisfy Rayleigh's criterion, this must correspond to the minimum of the neighboring wavelength's peak in the same order, i.e.

$$a\sin\theta_m = \left(m + \frac{1}{N}\right)\lambda \tag{9.51}$$

where N is the number of grating lines. Therefore, we obtain

$$\boxed{R = \frac{\lambda}{\Delta\lambda} = mN} \tag{9.52}$$

This basically shows that at higher orders (m) and larger numbers of slits (N), wavelengths are better resolved. We can also quantify this with the angular dispersion, which gives the angular separation per unit wavelength

$$D = \frac{\mathrm{d}\theta_m}{\mathrm{d}\lambda} = \frac{m}{a\cos\theta_m} \tag{9.53}$$

With a focal length f, this becomes a linear dispersion across the detector of

$$\frac{\mathrm{d}y}{\mathrm{d}\lambda} = f\frac{\mathrm{d}\theta_m}{\mathrm{d}\lambda} = fD \tag{9.54}$$

163. Summarize the current state of experimental gravitational wave detection, and the prospects for improvement in the near future.

The current state-of-the-art gravitational wave observatory is LIGO, operated jointly by MIT/Caltech. It has two observatories in Hanover, WA and Livingston, LA. It has had a few observing runs:

- O1 (2015)

  GW150914: The first gravitational wave detection, a binary black hole merger, both with $M \sim 30\ M_\odot$.

  One other event.

- O2 (2017)

  15–25% improved sensitivity over O1.

  VIRGO comes online (Italy), which helps with localization.

  GW170817: The first binary neutron star merger that was also detected in EM observations (with Fermi).

  Total of 6 detections.

- O3 (2019–2020)

  KAGRA comes online (Japan).

  More than 50 binary BBH mergers and 1 BNS merger.

  GW190521: Binary black holes in the pair instability gap ($\sim 150\ M_\odot$).

- O4 (2023–2025)

  Still actively ongoing

LIGO, VIRGO, and KAGRA together form the LVK collaboration. To date, ∼100 gravitational wave detections have been recorded in total.

There are constant efforts to improve sensitivity and reduce noise. Sources of noise include:

- Thermal noise
- Quantum noise (shot noise and radiation pressure from the laser beams)
- Seismic activity

The 7.6 magnitude earthquake in Japan on January 1, 2024 damaged a lot of KAGRA's mirror suspension systems, so it is temporarily offline until they can be repaired before the end of O4.

For the future:

- Further improvements to the LVK observatories will push us to higher sensitivities.
- Planned LISA observatory, which will be a space-based interferometer with much longer arms and sensitivity in the mHz regime (more sensitive to SMBH binaries and EMRIs). See §9.Q160.
- Planned Cosmic Explorer observatory, which will be a ground-based interferometer with 40 km arms. Sensitivity $\propto$ length^2.

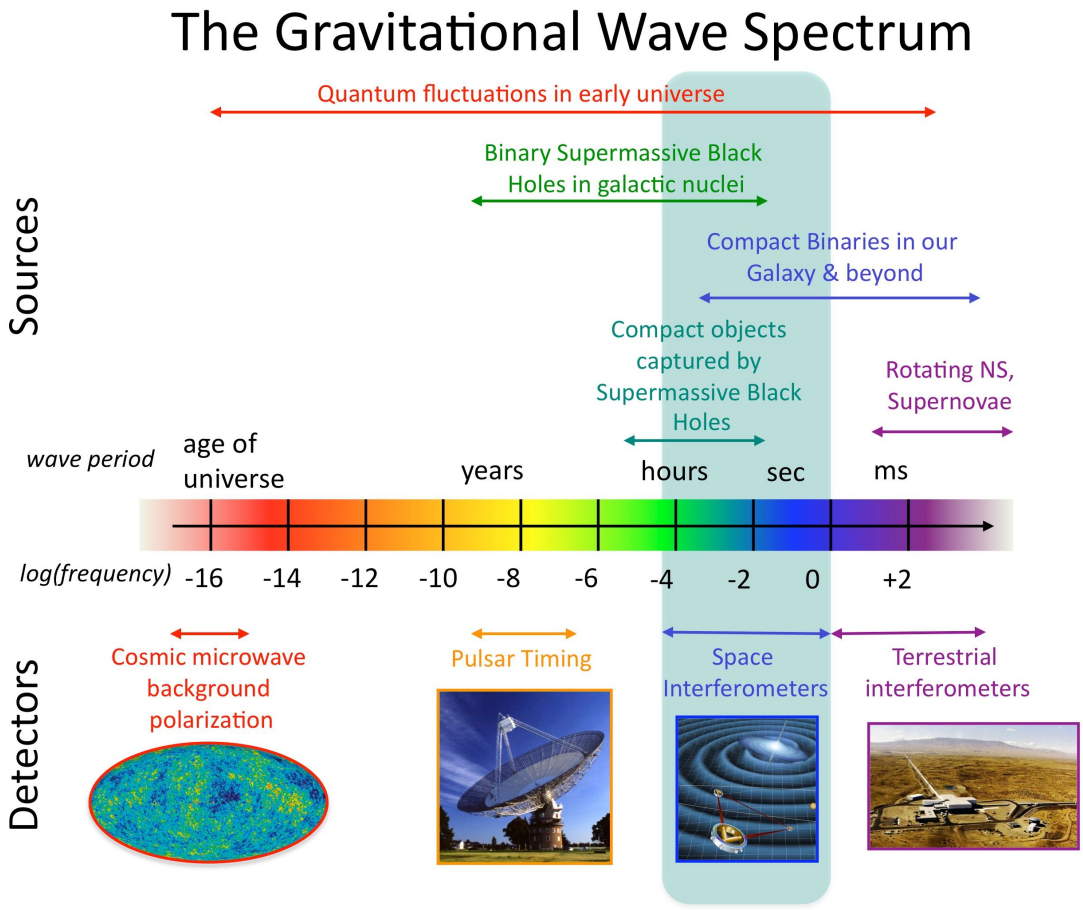


Figure 9.23: The gravitational wave spectrum.

10 Lighting Round / Miscellaneous

164. What kind of objects would you guess the following were? HD 128220, Abell 426, QSO 1013+277, α Crucis, Mk 509, 3C 234, NGC 2808, PSR 0950+08, BD +61° 1211, β Pictoris b

HD 128220

Regular star. HD stands for the Henry Draper catalogue, which contains stars.

Abell 426

AKA the Perseus Cluster. The Abell catalogue is a catalogue of galaxy clusters.

QSO 1013+277

A quasi-stellar object or quasar. The numbers here indicate truncated coordinates.

α Crucis

The brightest star in the Crux constellation, AKA the Southern Cross. Stars are labeled from brightest to dimmest following Greek letters: α Crucis, β Crucis, ... etc.

Mk 509

An AGN. Mk stands for the Markarian catalogue which contains galaxies with excess UV emission, indicative of AGN.

3C 234

A radio-loud AGN. 3C stands for the 3rd Cambridge Catalogue of radio sources.

NGC 2808

A globular cluster. NGC stands for the New General Catalogue, which contains all kinds of extended objects in the nearby universe, including galaxies, emission nebulae, and globular clusters.

PSR 0950+08

A pulsar. The numbers here are truncated coordinates, like with the QSO.

BD +61° 1211

A star. BD stands for the Bonner Durchmusterung catalogue, which contains stars. NOT A BROWN DWARF.

β Pictoris b

The first planet around the β Pictoris star. Planets are labeled with lowercase roman letters starting from b (β Pictoris a would be β Pictoris itself).

165. What are the strongest spectral lines seen in each of the following: Integrated light from a typical galaxy, a quasar, a 100 km/s interstellar shockwave, a giant molecular cloud?

Typical galaxy

Galaxy spectra are made up of a few different components[233]:

- **Stellar** (UV/optical/NIR): Contributions from many individual stars (i.e. blackbodies) at different temperatures, which is relatively flat apart from a break at 4000 Å caused by blanket absorption of high energy photons by metals, and breaks at the Balmer limit (3645 Å) and Lyman limit (911 Å). There are also features due to metals in cool stellar atmospheres.
- **Nebular** (UV–radio): Lyman continuum photons from hot stars ionize the surrounding gas, which heats up and then re-radiates the energy in a series of emission lines (free-bound to bound-bound) and continuum emission (free-free). The emission lines are prominent but the nebular continuum emission tends to be negligible in most cases and is usually ignored.
- **Dust Attenuation** (UV/optical/NIR): Dust grains in the ISM are efficient at absorbing short-wavelength radiation. This affects bluer wavelengths more than redder wavelengths and hence has the effect of "reddening" the spectrum. There is also a bump in the UV part of the attenuation curve.
- **Dust Emission** (MIR/FIR): The energy that dust absorbs in the UV/optical/NIR must go somewhere, so it is radiated away in the MIR/FIR domain. At shorter wavelengths, this is done by de-exciting PAH molecules, which produce broad emission features. Whereas longer wavelengths are dominated by larger, colder dust grains which emit like modified blackbodies.
- **Synchrotron Emission** (radio): Relativistic electrons from supernovae can produce non-thermal synchrotron radiation when interacting with the local magnetic field.
- **IGM Absorption** (UV): The intergalactic medium absorbs UV radiation past the Lyman limit, depending on the neutral Hydrogen content (i.e. depending on the redshift). See §8.Q142.

Now to answer the actual question, it depends on the type of galaxy[234].

- **Elliptical galaxies** have no young, hot, ionizing stars, so they don't produce any nebular emission. Thus, no emission lines. They will just have absorption lines from metals in stellar atmospheres. Some of the most prominent ones are:
  - Ca H and K (Ca II $\lambda\lambda$3968,3934)
  - The G band (Ca I λ4307.7 and Fe I λ4307.9)
  - Mg I λ5175
  - Na D (Na I $\lambda\lambda$5896,5889)
  - H balmer series (Hα 6563, Hβ 4861, Hγ 4340, ...)
- **Spiral galaxies** do have young, hot, ionizing stars, so they have emission lines in addition to the absorption lines. Most prominently:
  - [O II] $\lambda\lambda$3737,3739
  - [O III] $\lambda\lambda$4959,5007
  - [N II] $\lambda\lambda$6583,6548
  - H balmer series (Hα 6563, Hβ 4861, Hγ 4340, ...)

- [S IV] λ10.511 μm
- [Ne II] λ12.813 μm
- [Ne III] λ15.555 μm

Quasar

Like galaxy spectra, quasar spectra are complex and multifaceted. For a full breakdown of the SED like I did for the galaxy, see §4.Q71. The important thing to realize here is that the quasar still has a host galaxy around it, so we're still going to see the same prominent emission lines that we saw in the galaxy spectra. But thanks to the strong UV continuum emission from the AGN, we'll also see more prominent UV emission lines, and even higher-ionization "coronal" lines that wouldn't be possible with a normal galaxy (though these are typically not as bright).

- Lyα 1216
- C IV $\lambda\lambda$1548,1550

Coronal lines:

- [Ne V] $\lambda\lambda$3426,3346
- [Fe X] λ6374
- [Ne VI] λ7.652 μm
- [Mg VII] λ9.009 μm
- [Ne V] λ14.322 μm

100 km/s Interstellar shock wave[235]

Shocks are produced by supernovae. The temperature of a 100 km/s shock, using our results from §4.Q50, is

$$T \simeq 2 \times 10^7 \left(\frac{u_s}{1000\,\mathrm{km\,s^{-1}}} \right)^2 \mathrm{K} \quad \longrightarrow \quad T \simeq 2 \times 10^5 \mathrm{\ K} \simeq 18 \mathrm{\ eV} \tag{10.1}$$

Like with our argument in §8.Q138, even though this is below the ionization potential of He II (24.5 eV $\approx 3 \times 10^5$ K), since there are many more photons than baryons, we can still get significant ionization of He II lines, in addition to the Hydrogen lines

- He II λ4686
- H balmer series (Hα 6563, Hβ 4861, Hγ 4340, ...)

Giant molecular cloud

These clouds are mostly molecular H_2. If the cloud is cold ($\sim$ 10 K), the H_2 does not emit strongly, so we have to use the next most common molecule, CO:

- CO 1–0 2.6 mm
- CO 3–2 1.4 mm

Otherwise, if the cloud is warm ($\sim$ 100 K), we can use rovibrational lines in the IR:

- H_2 0–0 S(0) 28.221 μm
- H_2 1–0 S(0) 2.2235 μm

166. How far away are the following objects: The Sun, the Orion Nebula, M13, M31, the Virgo Cluster, the Coma Cluster, 3C 273? How large and how massive are they?

	Mass	Radius*	Distance
The Sun	$1\,M_\odot \approx 2\times10^{33}$ g	$1\,R_\odot \approx 7\times10^{10}$ cm	$1\,\mathrm{AU} \approx 1.5\times10^{13}$ cm
Orion Nebula	2000 $M_\odot$	4 pc	412 pc
M13 (Hercules Glob. Cluster)	$6\times10^5\ M_\odot$	26 pc	6.8 kpc
M31 (Andromeda)	$1.5\times10^{12}\ M_\odot$	23.3 kpc	765 kpc
Virgo Cluster	$1.2\times10^{15}\ M_\odot$	2.2 Mpc	16.5 Mpc
Coma Cluster	$7\times10^{14}\ M_\odot$	3 Mpc	100 Mpc ($z\sim0.02$)
3C 273†	$9\times10^8\ M_\odot$	3×10^{14} cm	749 Mpc ($z\sim0.16$)

*1 pc $= 3.086\times10^{18}$ cm

†Reported mass and radius are for the black hole (Schwarzschild radius), and reported distance is luminosity distance.

Refs. 236,237,238,239,240,241.

167. What are the brightest extra-solar-system sources at the following wavelengths: radio, mm, midIR, optical, UV, X-ray, gamma ray? Which, if any, are isotropic on the sky?

Wavelength	Individual examples (apparent)	Other bright sources
Radio[242]	Cas A, Cyg A, Cen A	Radio-loud AGN*, FRBs*
Millimeter	Perseus Molecular Cloud	The CMB*, cold molecular gas/dust, AGN*
MIR[243]	η Car, Omega Nebula, Orion Nebula	Cool stars, galactic dust, starburst galaxies*, AGN*
NIR[243]	Betelgeuse, R Doradus, Antares	Cool stars, galactic dust, starburst galaxies*, AGN*
Optical[243]	Sirius, Canopus, Arcturus	Stars, galaxies*, AGN*
UV	Adhara	Hot stars, hot white dwarfs, AGN*
X-ray[244]	SGR 1806-20, GRB 100621A, A 0620-00	XRBs, supernova remnants, galaxy clusters, AGN*
Gamma-ray[244]	GRB221009A	GRBs*, XRBs, AGN*

*Isotropic

168. Briefly describe the following famous astronomical objects: The Crab, M3, M31, M87, W51, Cyg X-1, 3C 273, Cyg A, SS 433, GW150914, GW170817 , LMC, Boötes Void, Virgo cluster, TRAPPIST-1, Prox Cen b, Sgr A*

The Crab Nebula[245]

AKA: M1, NGC 1952, Taurus A

Distance: 2 kpc | Radius: 1.7 pc | Mass: 5 $M_\odot$

A supernova remnant and pulsar wind nebula in the constellation Taurus. The supernova event was recorded in 1054 by Chinese astronomers. At the center is the Crab Pulsar, which powers the whole nebula. One of the brightest persistent X-ray and gamma-ray sources in the sky.

M3[246]

AKA: NGC 5272

Distance: 10.4 kpc | Radius: 27.6 pc | Mass: 4.5×10^6 $M_\odot$

A globular cluster in the constellation Canes Venatici. It is one of the largest and brightest globular clusters, with around 500,000 stars and an estimated age of 11.4 Gyr. Contains a lot of variable stars.

M31[238]

AKA: Andromeda, NGC 224

Distance: 765 kpc | Radius: 23.3 kpc | Mass: 1.5×10^{12} $M_\odot$

The closest galaxy to us that is not a dwarf/satellite galaxy of the Milky Way. About the same size/mass as the Milky Way. It is a member of the Local Group. Will collide with the Milky Way in about 5 Gyr.

M87[247]

AKA: Virgo A, NGC 4486

Distance: 16.4 Mpc | Radius: 150 kpc | Mass: 6×10^{12} $M_\odot$

The BCG of the Virgo Cluster—a massive supergiant elliptical galaxy (type cD). Has a relativistic jet extending 1.5 kpc from the nucleus, and is one of the brightest radio sources in the sky. Its SMBH, M87*, with a mass of 3.5×10^9 $M_\odot$, was imaged by the EHT in 2017.

W51[248]

Distance: 5 kpc | Radius: 100 pc | Mass: 10^6 $M_\odot$

A giant molecular cloud, one of the most active star-forming clouds in our Galaxy. Highly obscured by dust, so it's only visible in the IR.

Cyg X-1[249]

AKA: Cygnus X-1

Distance: 2.2 kpc | BH Radius: 63 km | BH Mass: 21.2 $M_\odot$

A stellar mass black hole in a high-mass X-ray binary system in the constellation Cygnus. It was discovered in 1965 as one of the brightest X-ray sources in the sky detectable from Earth, and it became the first discovered source to be widely accepted as a black hole. Its companion is a blue supergiant variable star.

3C 273[241]

AKA: PGC 41121, HIP 60936

Distance: 749 Mpc | BH Radius: 3×10^9 km | BH Mass: 9×10^8 $M_\odot$

A quasar at the center of a giant elliptical galaxy in the constellation Virgo (*not* the Virgo Cluster). It was the first quasar ever discovered and is the brightest quasar visible from Earth. Coronagraphic observations with *HST* later revealed the faint host galaxy.

Cyg A[250]

AKA: Cygnus A, 3C 405

Distance: 232 Mpc | Radius: 18 kpc | Mass: 10^{11} $M_\odot$

A radio galaxy with an obscured (type II) AGN that emits two strong radio jets that have inflated radio lobes (FR-II). It is the BCG of a cluster of galaxies. There are also X-ray cavities in the ICM coincident with the radio lobes.

SS 433[251,252]

Distance: 5.5 kpc | Orbital Radius: 8×10^7 km | System Mass: 36 $M_\odot$

An eclipsing X-ray binary system containing a stellar mass black hole and an A-type star. It was the first discovered "microquasar", with jets and an accretion disk producing very high energy photons.

GW150914[253]

Distance: 410 Mpc | BH Masses: 35 $M_\odot$, 30 $M_\odot$

The first gravitational wave detection. Came from a binary black hole merger and lasted about 0.2 seconds, with the frequency increasing from 35–150 Hz.

GW170817[254]

Distance: 40 Mpc | NS Masses: ~ 1.4 $M_\odot$, ~ 1.4 $M_\odot$

A gravitational wave signal from the merger of two neutron stars detected in 2017. Was the first GW observation to have an electromagnetic counterpart observation: *Fermi* detected a short GRB at the same time/place. This was then followed up across other wavebands in the optical/UV/IR/etc. to observe the afterglow, which localized the source.

LMC[255]

AKA: Large Magellanic Cloud

Distance: 50 kpc | Radius: 5 kpc | Mass: 10^{11} $M_\odot$

The largest satellite galaxy of the Milky Way. An irregular galaxy with lots of gas, dust, and star formation. Was likely a barred spiral before being tidally disrupted by the Milky Way. Named for Ferdinand Magellan, who first noticed the LMC/SMC during his voyages around the world.

Boötes Void[256]

AKA: "The Great Nothing"

Distance: 250 Mpc | Radius: 60 Mpc

A large spherical region in the constellation Boötes that contains a significant underdensity of galaxies (only about 60 galaxies where, for a region of this size, we should expect 2000). It is one of the largest known Voids.

Virgo Cluster[239]

Distance: 16.5 Mpc | Radius: 2.2 Mpc | Mass: $1.2 \times 10^{15}\ M_\odot$

A galaxy cluster in the constellation Virgo. It is a part of the Virgo supercluster, which the Local Group is also a member of. Its BCG is M87 (see above), and other prominent members include M49, M86, M60, all of which are elliptical galaxies. It has a total of about 1000 member galaxies.

TRAPPIST-1[257]

AKA: EPIC 246199087, K2-112

Distance: 12.5 pc | Radius: $0.1\ R_\odot$ | Mass: $0.09\ M_\odot$

A cool red dwarf star (spectral type M8V) with 7 known orbiting exoplanets (TRAPPIST-1b through TRAPPIST-1h), with 4 in the habitable zone. Discovered by the Transiting Planets and Planetesimals Small Telescope (TRAPPIST) in 2016. The planets are packed closely together and the whole system is much tighter than the Solar System. The planets all have orbital resonances of 8:5, 5:3, 3:2, 3:2, 4:3, and 3:2 between each pair, which can keep the system stable for billions of years.

Prox Cen b[258]

AKA: Proxima Centauri b

Distance: 1.3 pc | Radius: $1\ R_\oplus$ | Mass: $1\ M_\oplus$

The closest exoplanet to Earth, around the star Proxima Centauri (the closest star to the Sun). It orbits within the habitable zone, and is thought to be Earthlike in terms of its mass and radius. Discovered in 2016 with the RV method.

Sgr A*[259]

AKA: Sagittarius A*

Distance: 8.3 kpc | Radius: 1.3×10^7 km | Mass: $4.3 \times 10^6\ M_\odot$

The supermassive black hole at the center of the Milky Way galaxy, located in the Sagittarius constellation. Its mass has been determined by the orbits of stars around the galactic center, which confirmed it to be a SMBH (2020 Nobel Prize: Ghez & Genzel). Imaged by the EHT in 2017.

11 Prepared Question

Note: This section is obviously just for my own reference, as everyone's prepared question will be different. But perhaps looking at it might give others some ideas of the kinds of things to be looking for when coming up with a prepared question.

Q: How can one measure the mass of a galaxy's central supermassive black hole, assuming it is part of a binary black hole system, using only the galaxy's surface brightness profile? What kinds of systems can we actually use this technique on, and how do its results compare to other SMBH mass measurement techniques?

Everyone knows about the $M-\sigma$ relationship and reverberation mapping, but this method for estimating black hole masses is much less well known. There is a beautiful coalescence of many areas of physics in deriving this result, but it is held back by large uncertainties due to our rather limited sample size of galaxies with black hole masses that can be estimated this way.

To start off, we need to have a galaxy with a "cored" surface brightness profile—that is, the intensity I should level off to a near-constant value within some distance from the nucleus. Note that while many *dark matter* density profiles have cored shapes (i.e. the core-cusp problem, §7.Q125), we're talking here about a brightness profile, which strictly traces the *luminous* matter. Most galaxies do not have cored brightness profiles—they're traced well by Sérsic, de Vaucouleurs, or exponential profiles (i.e. §7.Q118), which all continue increasing as $r \to 0$.

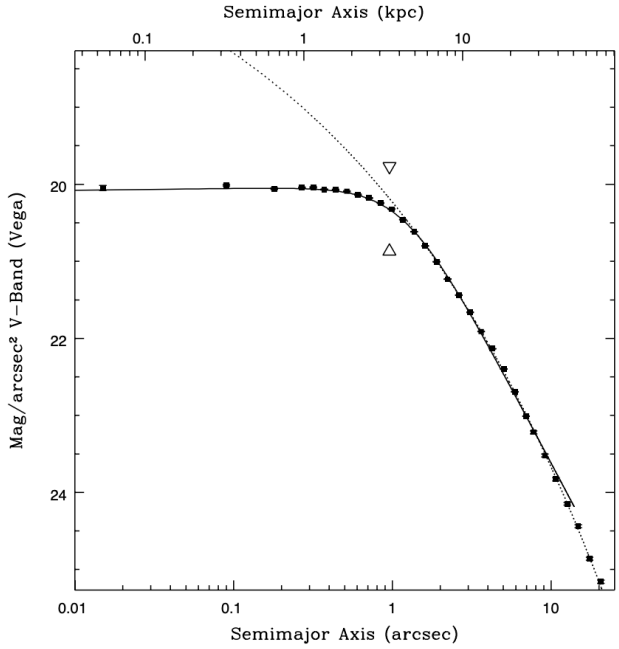


Figure 11.1: An example of a Nuker profile being used to fit the surface brightness profile of the BCG of Abell 2261[260]. The break radius is marked with triangles.

It is often conventient to parametrize this using the so-called **Nuker profile**

$$I(r) = 2^{(\beta-\gamma)/\alpha} I_b \left(\frac{r}{r_b}\right)^{-\gamma} \left[1+\left(\frac{r}{r_b}\right)^{\alpha}\right]^{(\gamma-\beta)/\alpha} \tag{11.1}$$

This is characterized by having an inner core slope γ that transitions to an outer envelope slope β, and α determines how quickly the transition occurs (i.e. the curvature). The transition between slopes happens at a characteristic "break radius" r_b, at which the intensity is $I_b = I(r_b)$. An example of this profile is use is shown in Figure 11.1.

Incredibly, we can measure the mass of the central black hole directly from the size of the break radius r_b. Why? The idea is that, if the central SMBH is in a binary system, dynamical interactions with the surrounding stars tends to give them a push, which boosts their energies and causes them to be ejected from the core, which flattens out the core shape on scales much larger than the gravitational influence of the binary itself. How would a binary black hole system form in the first place? The most likely culprit would be recent galaxy mergers where the central black holes of both galaxies have not coalesced yet.

So, to get our desired relationship (relating $M_{\rm BH}$ to r_b), we require 2 intermediate steps:

1. Relate the break radius r_b to the stellar mass that has been ejected from the core, $M_{\rm ej}$
2. Relate the ejected stellar mass $M_{\rm ej}$ to the black hole mass $M_{\rm BH}$

For step 1, we can estimate the amount of *light* that is missing from the core by extrapolating the outer envelope slope inwards and integrating the difference between this theoretical "unperturbed" profile and the true, cored profile. This is shown schematically in Figure 11.2.

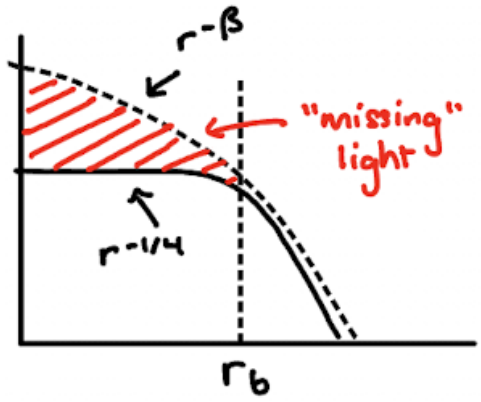


Figure 11.2: An example showing how the missing light would be calculated from the difference of two intensity profiles.

The light "deficit" can then be converted into the ejected mass by assuming some mass-to-light ratio $\Gamma = \langle M/L \rangle$, i.e.

$$\mathrm{d}M = 2\pi r I(r) \Gamma \mathrm{d}r \tag{11.2}$$

The Nuker profile is kind of cumbersome, so for simplicity, let's assume a profile that instantly transitions from a shallow core slope of $-1/4$ to a generic envelope slope $-\beta$ at r_b. Then our new profile is

$$I(r) = \begin{cases} I_1(r) = I_b (r/r_b)^{-1/4} & r < r_b \\ I_2(r) = I_b (r/r_b)^{-\beta} & r \geqslant r_b \end{cases} \tag{11.3}$$

We now want to integrate the difference betwen $I_1(r)$ and $I_2(r)$, extrapolated from 0 to r_b:

$$M_{\rm ej} = 2\pi\Gamma \int_0^{r_b} r[I_2(r) - I_1(r)]\mathrm{d}r \tag{11.4}$$

$$= 2\pi\Gamma I_b \int_0^{r_b} (r^{1-\beta} r_b^{\beta} - r^{3/4} r_b^{1/4})\mathrm{d}r \tag{11.5}$$

$$= 2\pi\Gamma I_b \left[r_b^{\beta} \frac{r^{2-\beta}}{2-\beta} - r_b^{1/4} \frac{r^{7/4}}{7/4} \right]_0^{r_b} \tag{11.6}$$

$$= 2\pi\Gamma I_b r_b^2 \left(\frac{1}{2-\beta} - \frac{4}{7} \right) \tag{11.7}$$

$$\boxed{M_{\rm ej} = 2\pi\Gamma \frac{4\beta - 1}{14 - 7\beta} I_b r_b^2} \tag{11.8}$$

It seems at first glance like $M_{\rm ej} \propto r_b^2$, but I_b and r_b are *not* independent parameters. These are intrinsic properties of the core that depend on the dynamics of the binary BH interactions. Intuitively, we would expect an inverse scaling, since as the black hole mass increases, the energies ejected into the scatterings increases, so the core size increases, while at the same time the light from the stars gets spread out further and decreases.

If we just require that the total integrated core light stays the same as I_b and r_b change, we recover the expected relationship. This makes sense—if the binary black hole changes in mass, then r_b and I_b will change, i.e. the core mass/light will be more or less spread out, but this can't magically add or remove mass/light from the core:

$$I_c = \int_0^{r_b} I_b \left(\frac{r}{r_b} \right)^{-1/4} \mathrm{d}r = \frac{4}{3} I_b r_b \quad \longrightarrow \quad I_b = \frac{3 I_c}{4 r_b} \propto \frac{1}{r_b} \tag{11.9}$$

Applying this to (11.8), we find that the ejected mass scales *linearly* with the break radius[261]:

$$\boxed{M_{\rm ej} = \frac{3}{2} \frac{4\beta - 1}{14 - 7\beta} \pi\Gamma I_c r_b} \tag{11.10}$$

And we're done with step 1! Now onto step 2. This one is a bit involved. To relate $M_{\rm ej}$ to $M_{\rm BH}$, we need to understand how the binary interacts with and loses energy to the surrounding stars. In general, these are complicated 3-body interactions, but we can simplify them by considering two extremes: longer range dynamical friction interactions where we can look at each BH individually, and shorter range ejection interactions where we can treat the binary as tightly bound.

We'll start with the first case. Assuming a wide binary orbit, as each BH travels along, it interacts with nearby stars, which in the rest frame of the BH are seen to be moving towards it. The gravitational force tends to pull the stars preferentially behind the BH's direction of motion, which leads to an overdensity and a drag force. Say a star with mass m is incident at an impact parameter b, see Figure 11.3. If we make the simplifying assumption that the perturbation in velocity is small, $\delta v_\perp \ll v$, such that the star travels along a nearly straight path, we can find $\delta v_\perp$ by integrating the perpendicular

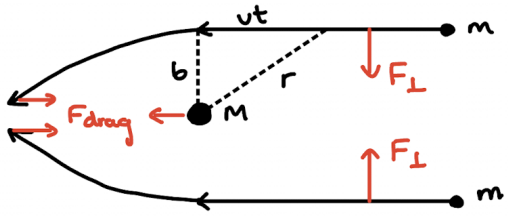


Figure 11.3: The dynamical friction caused by the pile-up of stars behind the direction of motion.

gravitational force $F_\perp$:

$$F_\perp = \frac{GMm}{r^2}\cos\theta = \frac{GMm}{b^2+x^2}\frac{b}{\sqrt{b^2+x^2}} = \frac{GMm}{b^2(1+v^2t^2/b^2)^{3/2}} \tag{11.11}$$

Integrating the force (per unit mass) over time yields

$$F_\perp \approx m\frac{\delta v_\perp}{\delta t} \quad\longrightarrow\quad \delta v_\perp \approx \frac{1}{m}\int_{-\infty}^{\infty} F_\perp \mathrm{d}t = \frac{GM}{b^2}\int_{-\infty}^{\infty}\frac{\mathrm{d}t}{(1+v^2t^2/b^2)^{3/2}} \tag{11.12}$$

Making the substitution $s = vt/b$

$$\delta v_\perp = \frac{GM}{bv}\int_{-\infty}^{\infty}\frac{\mathrm{d}s}{(1+s^2)^{3/2}} \tag{11.13}$$

And then making the further substitution $s = \tan u$, $\mathrm{d}s = \sec^2 u\,\mathrm{d}u$ and $1+\tan^2 u = \sec^2 u$, so

$$\delta v_\perp = \frac{GM}{bv}\int_{-\pi/2}^{\pi/2}\cos u\,\mathrm{d}u = \frac{2GM}{bv} \tag{11.14}$$

This gives the velocity change due to one interaction. Now, the total energy of the system must be conserved, so the star's *total* relative velocity before and after the interaction (at an equally large distance away) must be the same. Clearly, the relative parallel velocity must change to compensate:

$$v_i = v_f \quad\longrightarrow\quad v^2 = (v+\delta v_\parallel)^2 + \delta v_\perp^2 \tag{11.15}$$

$$0 = \delta v_\parallel^2 + 2v\delta v_\parallel + \delta v_\perp^2 \tag{11.16}$$

Using the quadratic formula and taking the positive root

$$\delta v_\parallel = -v + \sqrt{v^2 - \delta v_\perp^2} \tag{11.17}$$

$$= -v + v\sqrt{1-(\delta v_\perp/v)^2} \tag{11.18}$$

$$\approx -\frac{1}{2}v\left(\frac{\delta v_\perp}{v}\right)^2 \tag{11.19}$$

The drag force acting on the black hole, then, must equal the drag force that causes the change in the parallel velocity, i.e. in some small time window δt,

$$F_{\mathrm{drag}} = m\frac{\delta v_\parallel}{\delta t} = \frac{1}{2}mv\left(\frac{\delta v_\perp}{v}\right)^2\frac{1}{\delta t} \tag{11.20}$$

And we can estimate the amount of energy δE that is transferred from the BH to the star in this small time window (the perpendicular components cancel out but the parallel components add up)

$$\delta E = F_{\text{drag}} v \delta t = \frac{1}{2} m v^2 \left(\frac{\delta v_\perp}{v} \right)^2 = \frac{1}{2} m \delta v_\perp^2 = \frac{2G^2 M^2 m}{b^2 v^2} \tag{11.21}$$

Many stars will undergo these interactions, which we can quantify by considering a cylinder of constant impact parameter b around the BH (see Figure 11.4).

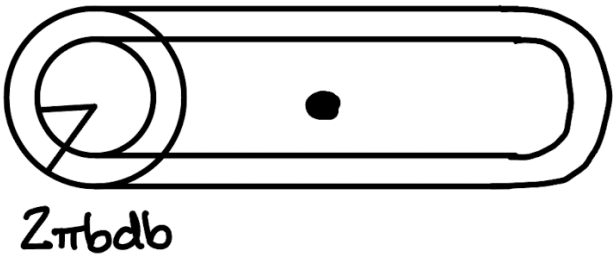


Figure 11.4: A number of stars at some range of impact parameters $b \pm \mathrm{d}b$ will pass by the binary and interact with it.

The number of stars passing through this cylinder will be proportional to its cross-sectional area, $2\pi b\,\mathrm{d}b$, and incident star flux $nv\Delta t$ where n is the number density of stars. Thus, $\mathrm{d}N = 2\pi b \mathrm{d}b n v \Delta t$. Each interaction causes the star to gain energy, while the BH loses energy. The total change in energy of the BH can then be found by integrating over b, assuming all incident stars have the same mass m

$$\Delta E = \sum_i \delta E_i = \int_{b_{\min}}^{b_{\max}} \frac{2G^2 M^2 m}{v^2} 2\pi b n v \Delta t \,\mathrm{d}b \tag{11.22}$$

Simplifying,

$$\Delta E = \frac{4\pi G^2 M^2 n m \Delta t}{v} \int_{b_{\min}}^{b_{\max}} \frac{\mathrm{d}b}{b} \tag{11.23}$$

And using the Coulomb logarithm $\ln\Lambda = \ln(b_{\max}/b_{\min})$

$$\frac{\Delta E}{\Delta t} = 4\pi \ln\Lambda \frac{G^2 M^2 n m}{v} \tag{11.24}$$

Now we can rewrite this using the mass density $\rho = nm$, and relabeling $v \to \langle v \rangle$. The constant we found, $4\pi \ln\Lambda$, in general will change depending on what assumptions are made about the distribution of stars in position and velocity space (i.e. phase space), so I'll just replace it with a generic C.

$$\dot{E} = C \frac{G^2 M^2}{\langle v \rangle} \rho \tag{11.25}$$

How does this loss of energy actually impact the binary? The orbit between the black holes must shrink, i.e. the semimajor axis a must get smaller. We know that a bound orbit has energy of the

form

$$E = -\frac{GM^2}{2a} \tag{11.26}$$

(assuming equal masses for the binary, $M_1 = M_2 = M$). So this corresponds to a change in a of

$$\dot{E} = -\frac{GM^2}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{a}\right) = C\frac{G^2M^2}{\langle v\rangle}\rho \quad \longrightarrow \quad \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{a}\right) = H\frac{G\rho}{\langle v\rangle} \tag{11.27}$$

where we defined the *hardening* coefficient $H = 2C$. "Hardening" in the binary sense means the orbit gets smaller.

In the regime where the star gets close to the black hole, i.e. $b \sim a$, the energy injection becomes significant, as $\delta v_\perp \sim v$:

$$\delta v_\perp \sim \frac{2GM}{bv} \quad \longrightarrow \quad v^2 \sim \frac{2GM}{a} \tag{11.28}$$

In this case, the assumption that δv is small breaks down, so our parallel velocity change becomes $\delta v_\parallel \sim \delta v_\perp \sim v$. Thus, the energy exchange for one strong encounter, using equation (11.21), is

$$\delta E = mv\delta v_\parallel \sim \frac{1}{2}mv^2 \sim \frac{GMm}{a} \tag{11.29}$$

Then, if we consider m as a small element of the total ejected mass, $m = \delta M_{\mathrm{ej}}$, then over a small time interval we get

$$\dot{E}_{\mathrm{ej}} = \frac{\delta E}{\delta t} \sim \frac{GM}{a}\frac{\delta M_{\mathrm{ej}}}{\delta t} \tag{11.30}$$

We can then equate the energy injected into the stars to the energy released from the binary.

$$\dot{E} \sim \frac{GM\dot{M}_{\mathrm{ej}}}{a} \sim C\frac{G^2M^2}{\langle v\rangle}\rho \quad \longrightarrow \quad \frac{1}{a}\frac{\mathrm{d}M_{\mathrm{ej}}}{\mathrm{d}t} \sim C\frac{GM}{\langle v\rangle}\rho \tag{11.31}$$

Using (11.27)

$$\mathrm{d}t = \frac{\langle v\rangle}{HG\rho}\mathrm{d}(1/a) \tag{11.32}$$

$$\mathrm{d}\ln(1/a) = \frac{\mathrm{d}(1/a)}{1/a} = a\mathrm{d}(1/a) \tag{11.33}$$

$$a\mathrm{d}t = \frac{\langle v\rangle}{HG\rho}\mathrm{d}\ln(1/a) \tag{11.34}$$

we can replace the time derivative with an a derivative. I will now relabel $M = M_{\mathrm{BH}}$ to be more clear[262]:

$$\frac{\mathrm{d}M_{\mathrm{ej}}}{\mathrm{d}\ln(1/a)} = JM_{\mathrm{BH}} \tag{11.35}$$

where the *mass ejection coefficient* J is roughly a constant of order unity, $J \sim C/H \sim 1/2$. We can

now integrate this equation to find $M_{\rm ej}$ as a function of $M_{\rm BH}$[263]:

$$M_{\rm ej} = J M_{\rm BH} \ln\left(\frac{a_{\rm ej}}{a}\right) \tag{11.36}$$

We see there is some upper limit, $a_{\rm ej}$, above which the binary mass ejection is not significant (and the inspiral is dominated by dynamical friction), and a lower limit $a \sim a_{\rm ej}/2$. The lower limit is caused by the fact that, as the black holes continue to scatter with stars and clear out the local neighborhood, eventually there are very few stars left that can actually be scattered and the binary hardening stalls. The exact ratio of $1/2$ is, of course, another approximation, but it does not affect the scaling, which we can see shows $M_{\rm ej}$ to be linearly proportional to $M_{\rm BH}$:

$$\boxed{M_{\rm ej} \sim J \ln 2\, M_{\rm BH}} \tag{11.37}$$

One may beg the question of how binary black holes can merge at all since the hardening stalls. From this analysis, we in fact would not expect them to merge within the age of the universe, stalling out at orbital separations on the order of parsecs. This is known as the **final parsec problem**, which I must say is a pretty sick name. At the very smallest orbital separations (i.e. 0.01–0.001 pc), gravitational radiation can kick in and become important in the final stages of the merger, but there is a critical gap between the parsec scales that we are left with from dynamical friction and the milliparsec scales needed for gravitational radiation. For more, see Ref. 264. But, this isn't what this question is about, so let's get back on track.

With this, we're done with step 2! Now all we have to do is combine them to see that the break radius is linearly porportional to the black hole mass!

$$\boxed{M_{\rm BH} \approx \frac{3}{2}\frac{4\beta-1}{14-7\beta}\frac{\pi\Gamma I_c}{J\ln 2} r_b} \quad \longrightarrow \quad \boxed{M_{\rm BH} \propto r_b} \tag{11.38}$$

This is backed up by observed scaling relations which find a near-linear dependence[261]

$$\boxed{\log_{10}\left(\frac{r_\gamma}{\rm pc}\right) = (0.96 \pm 0.09)\log_{10}\left(\frac{M_{\rm BH}}{10^9 M_\odot}\right) + (2.0 \pm 0.1)} \tag{11.39}$$

Here, r_γ is related to r_b by a constant and is used more frequently in the literature as it has been shown to have a tighter correlation with black hole parameters.

Now, the question becomes, if this scaling relation works, why is it used so much less frequently than the more well known scaling relations like $M - \sigma$? The simple answer is that there aren't a lot of systems where we can actually *use* this scaling relation, so the precise values in the above relation are not as constrained as the uncertainties would lead you to believe. The inner intensity cores for black holes of $10^7\ M_\odot$ are only on the order of parsecs, which we cannot resolve for most galaxies with our current best spatial resolutions. You need to look at the highest mass/luminosity galaxies with the highest mass black holes, such as brightest cluster galaxies (BCGs), to find ones where we can resolve the cores of the intensity profiles. Circa 2007, there were only 11 core galaxies which had independent measurements of black hole mass for which the relation can be calibrated. Depending on the method one uses to fit it, you can recover power law indices ranging from 0.6–1.5[261] (see Figure 11.5).

However, it is still useful to constrain the scaling at the high mass/luminosity end of galaxies/BCGs,

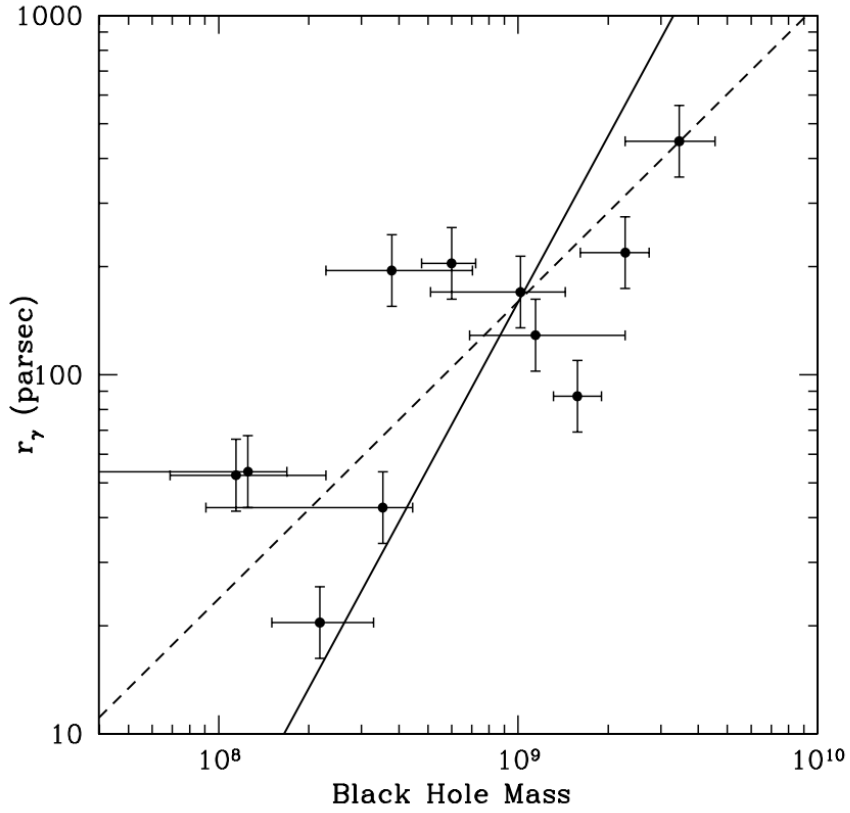


Figure 11.5: A fitted power law relationship between black hole mass and core radius from 261.

as the scaling deviates significantly from the canonical $M_{\rm BH} \propto \sigma^4$. In fact, using the scaling with r_γ reveals a much steeper dependence,

$$r_\gamma \propto \sigma^7 \quad \longrightarrow \quad M_{\rm BH} \propto \sigma^7 \tag{11.40}$$

for BCGs and cored galaxies[261]. In other words, there is curvature in the gravitational scaling relation. This hints that the relative importance of different physical processes changes as a function of galaxy size. There is still a lot unknown about this, but modern research suggests that "dissipationless" mergers become more important in the most massive elliptical galaxies. In these mergers, there is less gas, so there is less dissipation of energy through collisions (the stars are collisionless). The lack of gas could be explained as a result of ram pressure stripping in galaxy clusters, which coincides with the most massive galaxies that this turnover is observed in[265].

12 Appendices

12.1 Physical Constants and Units

Mathematical constants

pi	π	3.141592653589793238462...	
e (Euler's number)	e	2.718281828459045235360...	

Physical Constants

Speed of light in a vacuum	c	$2.99792458 \times 10^{10}\,\mathrm{cm\,s^{-1}}$	$299792458\,\mathrm{m\,s^{-1}}$
Gravitational constant	G	$6.674 \times 10^{-8}\,\mathrm{dyne\,cm^2\,g^{-2}}$	
Boltzmann constant	k	$1.381 \times 10^{-16}\,\mathrm{erg\,K^{-1}}$	$8.617 \times 10^{-5}\,\mathrm{eV\,K^{-1}}$
Planck constant	h	$6.626 \times 10^{-27}\,\mathrm{erg\,s}$	$4.136 \times 10^{-15}\,\mathrm{eV\,s}$
Reduced Planck constant	$\hbar \equiv h/2\pi$	$1.055 \times 10^{-27}\,\mathrm{erg\,s}$	$6.582 \times 10^{-16}\,\mathrm{eV\,s}$
Proton charge	e	$4.803 \times 10^{-10}\,\mathrm{statC}$	$1.602 \times 10^{-19}\,\mathrm{C}$
Proton mass	m_p	$1.673 \times 10^{-24}\,\mathrm{g}$	$938\,\mathrm{MeV}/c^2$
Electron mass	m_e	$9.109 \times 10^{-28}\,\mathrm{g}$	$511\,\mathrm{keV}/c^2$
Stefan-Boltzmann const.	$\sigma_{\rm SB} \equiv \pi^2 k^4/60\hbar^3 c^2$	$5.6705 \times 10^{-5}\,\mathrm{erg\,s^{-1}\,cm^{-2}\,K^{-4}}$	
Radiation constant	$a \equiv 4\sigma_{\rm SB}/c$	$7.5659 \times 10^{-15}\,\mathrm{erg\,cm^{-3}\,K^{-4}}$	
Classical electron radius	$r_e \equiv e^2/m_e c^2$	$2.8179 \times 10^{-13}\,\mathrm{cm}$	
Thomson cross section	$\sigma_{\rm T} \equiv 8\pi r_e^2/3$	$6.6525 \times 10^{-25}\,\mathrm{cm^2}$	
Fine structure constant	$\alpha \equiv e^2/\hbar c$	$1/137$	

Non-standard (but commonly used) units and conversions

ångström	Å	$10^{-8}\,\mathrm{cm}$	$10^{-10}\,\mathrm{m}$
electron-volt	eV	$1.602 \times 10^{-12}\,\mathrm{erg}$	$1.602 \times 10^{-19}\,\mathrm{J}$
Rydberg	Ryd	13.6 eV	
Jansky	Jy	$10^{-23}\,\mathrm{erg\,s^{-1}\,cm^{-2}\,Hz^{-1}}$	

Astronomical units and conversions

solar day	$\mathrm{day_{sol}}$	86,400 s	24 hr
sidereal day	$\mathrm{day_{sid}}$	86,164 s	23 hr 56 min 4 s
solar year	$\mathrm{yr_{sol}}$	$3.1557 \times 10^7\,\mathrm{s}$	$365.242\,\mathrm{day_{sol}}$
sidereal year	$\mathrm{yr_{sid}}$	$3.1558 \times 10^7\,\mathrm{s}$	$365.256\,\mathrm{day_{sol}}$
Lunar mass	M_M	$7.349 \times 10^{25}\,\mathrm{g}$	
Earth mass	$M_\oplus$	$5.972 \times 10^{27}\,\mathrm{g}$	$81\,M_M$
Jupiter mass	M_J	$1.898 \times 10^{30}\,\mathrm{g}$	$318\,M_\oplus$
Solar mass	$M_\odot$	$1.989 \times 10^{33}\,\mathrm{g}$	$1047\,M_J$
Lunar radius	R_M	$1.738 \times 10^8\,\mathrm{cm}$	
Earth radius	$R_\oplus$	$6.371 \times 10^8\,\mathrm{cm}$	$3.66\,R_M$
Jupiter radius	R_J	$6.991 \times 10^9\,\mathrm{cm}$	$10.97\,R_\oplus$
Solar radius	$R_\odot$	$6.958 \times 10^{10}\,\mathrm{cm}$	$9.95\,R_J$
Astronomical Unit	AU	$1.496 \times 10^{13}\,\mathrm{cm}$	
lightyear	ly	$9.461 \times 10^{17}\,\mathrm{cm}$	
parsec	pc	$3.086 \times 10^{18}\,\mathrm{cm}$	206,265 AU
Solar luminosity	$L_\odot$	$3.839 \times 10^{33}\,\mathrm{erg\,s^{-1}}$	

12.2 The Electromagnetic Spectrum

Gamma-rays	$\lambda \lesssim 0.1\,\text{Å}$	$124\,\text{keV} \lesssim h\nu$
Hard X-rays	$0.1\,\text{Å} \lesssim \lambda \lesssim 6\,\text{Å}$	$2\,\text{keV} \lesssim h\nu \lesssim 124\,\text{keV}$
Soft X-rays	$6\,\text{Å} \lesssim \lambda \lesssim 60\,\text{Å}$	$0.2\,\text{keV} \lesssim h\nu \lesssim 2\,\text{keV}$
Extreme Ultraviolet (EUV)	$60\,\text{Å} \lesssim \lambda \lesssim 900\,\text{Å}$	$13.6\,\text{eV} \lesssim h\nu \lesssim 0.2\,\text{keV}$
Far Ultraviolet (FUV)	$900\,\text{Å} \lesssim \lambda \lesssim 1200\,\text{Å}$	$10\,\text{eV} \lesssim h\nu \lesssim 13.6\,\text{eV}$
Near Ultraviolet (NUV)	$1200\,\text{Å} \lesssim \lambda \lesssim 3100\,\text{Å}$	$4\,\text{eV} \lesssim h\nu \lesssim 10\,\text{eV}$
Optical	$3100\,\text{Å} \lesssim \lambda \lesssim 7000\,\text{Å}$	$2\,\text{eV} \lesssim h\nu \lesssim 4\,\text{eV}$
Near Infrared (NIR)	$7000\,\text{Å} \lesssim \lambda \lesssim 3\,\mu\text{m}$	$100\,\text{THz} \lesssim \nu \lesssim 430\,\text{THz}$
Mid Infrared (MIR)	$3\,\mu\text{m} \lesssim \lambda \lesssim 25\,\mu\text{m}$	$10\,\text{THz} \lesssim \nu \lesssim 100\,\text{THz}$
Far Infrared (FIR)	$25\,\mu\text{m} \lesssim \lambda \lesssim 1\,\text{mm}$	$300\,\text{GHz} \lesssim \nu \lesssim 10\,\text{THz}$
Microwave	$1\,\text{mm} \lesssim \lambda \lesssim 1\,\text{cm}$	$30\,\text{GHz} \lesssim \nu \lesssim 300\,\text{GHz}$
Radio	$1\,\text{cm} \lesssim \lambda$	$\nu \lesssim 30\,\text{GHz}$

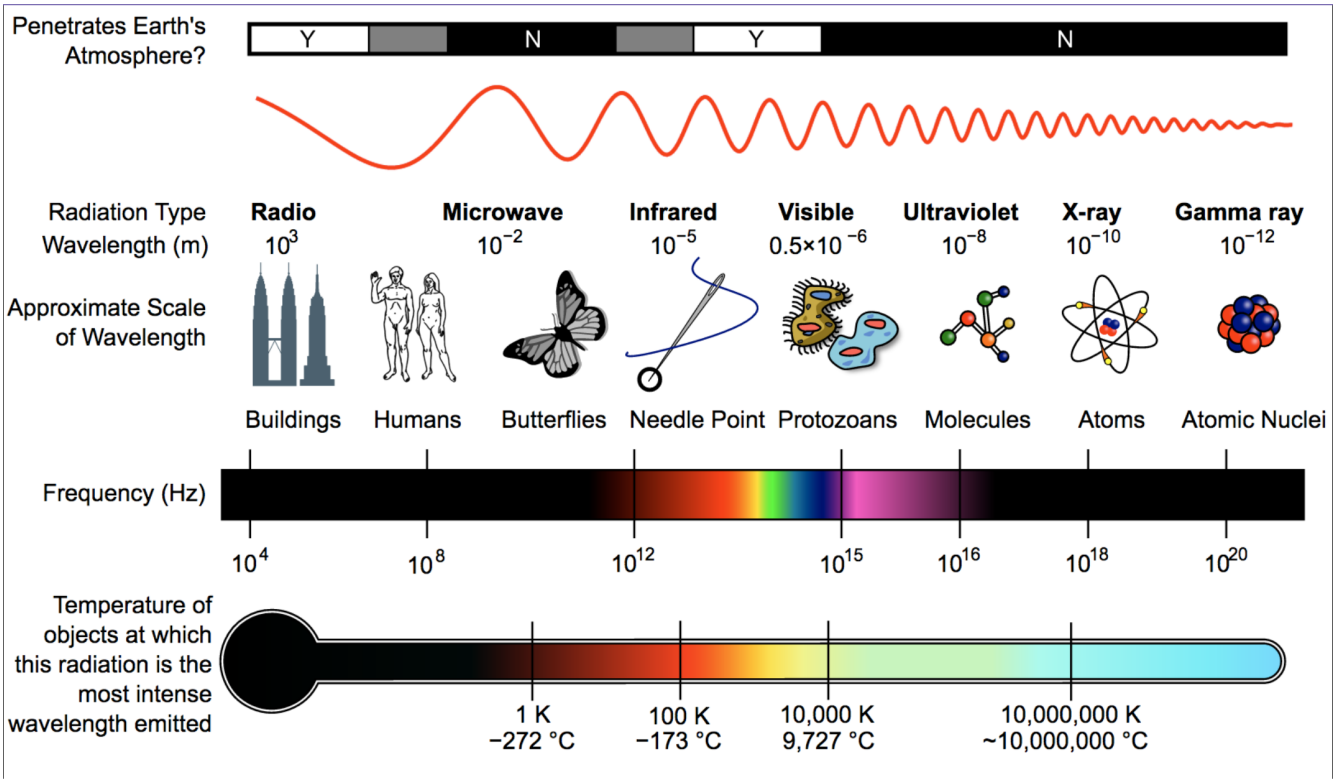


Figure 12.1: A diagram of the electromagnetic spectrum showing various properties of each wavelength band.

12.3 Thermal History of the Universe

- $t = 0$, $T = \infty$, $z = \infty$
  - **The Big Bang**
- $t \sim 10^{-43}$ s, $T \sim 10^{19}$ GeV, $z \sim 10^{40}$
  - **Planck era** (quantum gravity regime; read: we have no f*cking clue what happens here)
- $t \sim 10^{-38}$ s, $T \sim 10^{16}$ GeV, $z \sim 10^{37}$
  - **GUT phase transition** (Strong and electroweak interactions become distinguishable)
- $t \sim 10^{-36}$–10^{-32} s, $T \sim 10^{15}$ GeV, $z \sim 10^{35}$
  - **Inflation**
- $t \sim 10^{-30}$ s, $T \sim 10^{12}$ GeV, $z \sim 10^{31}$
  - **Peccei-Quinn phase transition** (if PQ is responsible for the charge/partiy problem)
- $t \sim 10^{-11}$ s, $T \sim 100$ GeV, $z \sim 10^{19}$
  - **Electroweak phase transition** (weak and electromagnetic interactions separate)
- $t \sim 10^{-8}$ s, $T \sim$ 10s of GeV, $z \sim 10^{17}$
  - **WIMPs freeze out** (if WIMPs are the explanation for dark matter)
- $t \sim 10^{-5}$ s, $T \sim$ 100–300 MeV, $z \sim 10^{15}$
  - **Quark-Hadron phase transition** (Quarks/gluons become bound into hadrons)
- $t \sim 0.1$ s, $T \sim 200$ MeV, $z \sim 10^{12}$
  - **Neutrino decoupling** (neutrinos stop interacting with other matter)
- $t \sim$ seconds–minutes, $T \sim$ 0.1–10 MeV, $z \sim 10^{10-11}$
  - **Big Bang Nucleosynthesis** (Neutrons and protons fuse into H, He, D, Li)
  - **Electron-positron Annihilation** (ambient temperature no longer high enough to create e^+e^- pairs)
- $t \sim$ days, $T \sim$ keV, $z \sim 10^8$
  - **Photon equilibrium** (interactions that can change the photon number freeze out)
- $t \sim 10^4$ yr, $T \sim 3$ eV, $z \sim 3000$
  - **Matter-Radiation Equality** (previously were radiation dominated)
- $t \sim 400$ kyr, $T \sim$ eV, $z \sim 1100$
  - **Recombination** (electrons and protons combine to form atoms)
  - CMB photons begin free-streaming
- $t \sim 400$ kyr–1 Gyr, $T \sim 10^{-3}$–1 eV, $z \sim$ 10–1100
  - **Cosmic Dark Ages** (before stars and galaxies start forming)

- $t \sim$ Gyr, $T \sim 10^{-3}$ eV, $z \sim 10$
  - **First stars and (small) galaxies form**
- $t \sim$ Gyr, $T \sim 10^{-3}$ eV, $z \sim 6$
  - **Reionization** (radiation from stars and quasars reionizes the universe)
- $t \sim 3$ Gyr, $T \sim 10^{-4}$ eV, $z \sim 2.5$
  - **Cosmic Noon** (the peak of star formation and AGN activity)
- $t \sim 10$ Gyr, $T \sim 10^{-4}$ eV, $z \sim 0.3$
  - **Matter-Λ Equality** (previously were matter dominated)
- $t \sim 13.7$ Gyr, $T \sim 10^{-4}$ eV, $z \sim 0$
  - **Today**

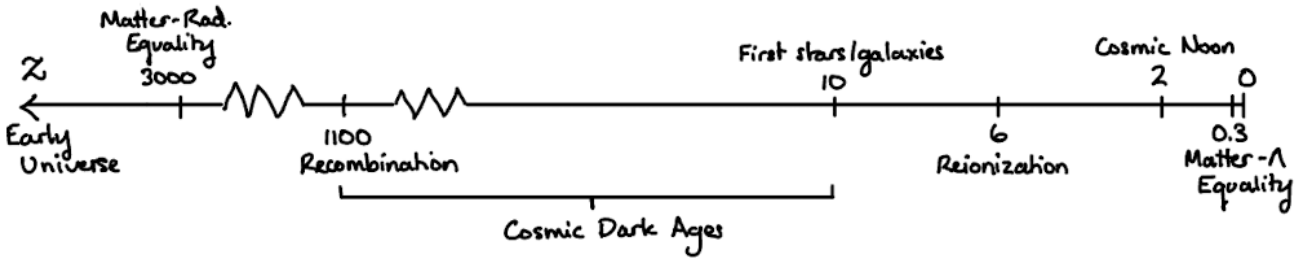


Figure 12.2: Timeline of the late univererse.

See Ref. 266.

12.4 Acronyms and Abbreviations

General terms:

- AGB: Asymptotic Giant Branch
- AGN: Active Galactic Nucleus
- AO: Adaptive Optics
- AXP: Anomalous X-ray Pulsar
- BAO: Baryon Acoustic Oscillations
- BBN: Big Bang Nucleosynthesis
- BCG: Brightest Cluster Galaxy
- BH: Black Hole
- BLR: Broad Line Region
- CCD: Charge-Coupled Device
- CDM: Cold Dark Matter
- CGRH: Cosmic Gamma-Ray Horizon
- CIE: Collisional Ionization Equilibrium
- CMB: Cosmic Microwave Background
- CGM: Circumgalactic Medium
- CNM: Cold Neutral Medium
- DM: Dark Matter
- DIB: Diffuse Interstellar Band
- EBL: Extragalactic Background Light
- EM: Electromagnetic
- EMRI: Extreme Mass Ratio Inspiral
- FLRW: Friedmann-Lemaître-Robertson-Walker
- FOV: Field of View
- FRB: Fast Radio Burst
- FWHM: Full-Width at Half-Maximum
- GR: General Relativity
- GRB: Gamma-Ray Burst
- GUT: Grand Unified Theory
- GW: Gravitational Waves
- HH: Herbig-Haro (Object)
- HIM: Hot Ionized Medium
- HMXB: High-Mass X-ray Binary
- HR: Hertzsprung-Russell (Diagram)
- ICM: Intracluster Medium
- ICL: Intracluster Light
- IFU: Integral Field Unit
- IGM: Intergalactic Medium
- ISCO: Innermost Stable Circular Orbit
- ISM: Interstellar Medium
- ISW: Integrated Sachs-Wolfe (Effect)
- KH: Kelvin-Helmholtz
- LMXB: Low-Mass X-ray Binary
- LSF: Line-Spread Function
- MACHO: Massive Compact Halo Object
- Maser: Microwave Amplification by Stimulated Emission of Radiation
- MHD: Magnetohydrodynamics
- MK: Morgan-Keenan (Luminosity class)
- MOND: Modified Newtonian Dynamics
- NLR: Narrow Line Region
- NS: Neutron Star
- PAH: Polycyclic Aromatic Hydrocarbon
- PSF: Point-Spread Function
- PSR: Pulsar / Pulsating Radio Source
- RGB: Red Giant Branch
- RTE: Radiative Transfer Equation
- RV: Radial Velocity
- SED: Spectral Energy Distribution
- SFR: Star Formation Rate
- SGB: Subgiant Branch
- SGR: Soft Gamma Repeater
- SIDM: Self-Interacting Dark Matter
- SMBH: Supermassive Black Hole
- SN(e): Supernova(e)
- SR: Special Relativity
- SSP: Simple Stellar Population
- SW: Sachs-Wolfe (Effect)
- SZ: Sunyaev-Zel'dovich (Effect)
- TAMS: Terminal-Age Main Sequence
- TRGB: Tip of the Red Giant Branch
- TTV: Transit Timing Variation
- WD: White Dwarf
- WDM: Warm Dark Matter
- WHIM: Warm-Hot Intergalactic Medium
- WIM: Warm Ionized Medium
- WIMP: Weakly Interacting Massive Particle
- WNM: Warm Neutral Medium
- XRB: X-Ray Binary
- ZAMS: Zero-Age Main Sequence

AGN classifications:

- BL Lac: BL Lacertae Objects
- Blazar: "BL Lac Quasar" (or what I like to call them, "Blasi-Ztellar Radio Source")
- BLRG: Broad Line Radio Galaxy
- FR-I/II: Faranroff-Riley class I/II
- NLRG: Narrow Line Radio Galaxy
- OVV: Optically Violently Variable Quasar
- QSO: Quasi-Stellar Object
- Quasar: Quasi-Stellar Radio Source
- Sy I/II: Seyfert I/II

Galaxy morphologies:

- BCD: Blue Compact Dwarf galaxy
- cD: Supergiant elliptical galaxy ("central dominant" backronym; "D" = diffuse)
- dE: Dwarf Elliptical galaxy
- dSph: Dwarf Spheroidal galaxy
- E0–E7: Elliptical galaxy
- Im–Ir: Irregular galaxy
- $S0_1$–$S0_3$: Lenticular galaxy
- $SB0_1$–$SB0_3$: Barred lenticular galaxy
- Sa–Sm: Spiral galaxy
- SBa–SBm: Barred spiral galaxy

12.5 Telescopes and Surveys

Some widely used telescopes in each wavelength band. Telescopes with MIT involvement are annotated: either MIT was directly involved in its development and/or funding (*), or there are MIT faculty who served some leadership role for it (†).

Gamma-rays:

- *Fermi*: *Fermi Gamma-ray Space Telescope*
- CTAO: Cherenkov Telescope Array Observatory (Cherenkov)
- MAGIC: Major Atmospheric Gamma Imaging Cherenkov Telescope (Cherenkov)

X-rays:

- *Arcus** (PLANNED): *Arcus*
- *Athena* (PLANNED): Advanced Telescope for High Energy Astrophysics
- *AXIS* (PLANNED): Advanced X-ray Imaging Satellite
- *Chandra**: *Chandra X-ray Observatory*
- *HETE-2** (DEFUNCT): *High-Energy Transient Explorer 2*
- *Hitomi* (DEFUNCT): *Hitomi*
- *IXPE*: Imaging X-ray Polarimetry Explorer
- *NICER**: *Neutron Star Interior Composition Explorer*
- *NuSTAR*: *Nuclear Spectroscopic Telescope Array*
- *REDSOX** (PLANNED): *sounding Rocket Experiment Demonstration of a Soft X-ray polarimeter*
- *ROSAT*: *Roentgen Satellite*
- *STAR-X** (PLANNED): *Survey and Time-domain Astrophysical Research eXplorer*
- *Suzaku** (DEFUNCT): *Suzaku*
- *Swift*: *Neil Gehrels Swift Observatory* (X-ray/FUV/NUV/optical)
- *XMM-Newton*: *X-ray Multi-Mirror Newton*
- *XRISM*†: *X-Ray Imaging and Spectroscopy Mission*

Ultraviolet:

- *FUSE* (DEFUNCT): Far Ultraviolet Spectroscopic Explorer (FUV)
- *GALEX* (DEFUNCT): Galaxy Evolution Explorer (FUV/NUV)
- *HWO* (PLANNED): *Habitable Worlds Observatory* (UV/Optical/IR)

Optical/NIR:

- ELT (PLANNED): Extremely Large Telescope (optical/NIR)
- *Gaia*: *Gaia* Astrometry Mission (NUV/optical/NIR)
- Gemini: Gemini North and Gemini South (optical/NIR)
- GMT (PLANNED): Giant Magellan Telescope (optical/NIR)
- *HST*: *Hubble Space Telescope* (FUV/NUV/optical/NIR)
- Keck: Keck Telescopes (optical/NIR)
- *Kepler* (DEFUNCT): *Kepler Space Telescope* (optical)
- LBT: Large Binocular Telescope (optical/NIR)
- Magellan*: Magellan Telescopes (optical/NIR) (MIT-built instruments: LLAMAS, FIRE)

- *Roman* (PLANNED): *Nancy Grace Roman Space Telescope* (NIR), formerly *WFIRST*
- Rubin (PLANNED): Vera C. Rubin Observatory (optical/NIR), formerly *LSST*
- Subaru: Subaru Telescope (optical/NIR/MIR)
- *TESS*[*]: *Transiting Exoplanet Survey Satellite* (red-optical/NIR)
- TMT (PLANNED): Thirty Meter Telescope (optical/NIR)
- VLT: Very Large Telescope (optical/NIR)
- Wallace[*]: MIT Wallace Observatory (mainly an educational facility) (optical/NIR)
- WINTER[*]: Wide-field Infrared Transient Explorer (NIR)
- WIYN[*]: WIYN Consortium telescope [University of Wisconsin-Madison (W), Indiana University (I), Yale University (Y), and National Optical Astronomy Observatories (N)] (optical/NIR) (MIT-built instruments: WISDOM)
- ZTF: Zwicky Transient Facility (optical/NIR)

MIR/FIR:

- *Akari* (DEFUNCT): *Akari*, the Japanese word for "light" (NIR/MIR/FIR)
- *Herschel* (DEFUNCT): *Herschel Space Observatory* (FIR)
- IRTF: NASA Infrared Telescope Facility (NIR/MIR)
- *JWST*: *James Webb Space Telescope* (NIR/MIR)
- SOFIA (DEFUNCT): Stratospheric Observatory for Infrared Astronomy (NIR/MIR/FIR)
- *Spitzer* (DEFUNCT): *Spitzer Space Telescope* (NIR/MIR/FIR)
- *WISE*: *Wide-field Infrared Survey Explorer* (NIR/MIR)

Microwave:

- ACT: Atacama Cosmology Telescope
- *COBE* (DEFUNCT): *Cosmic Background Explorer*
- *Planck* (DEFUNCT): *Planck* Mission
- SPT[†]: South Pole Telescope (microwave/radio)
- *WMAP* (DEFUNCT): *Wilkinson Microwave Anisotropy Probe*

Radio:

- ALMA: Atacama Large Millimeter/Submillimeter Array
- Arecibo (DEFUNCT): Arecibo Observatory
- CHIME[†]: Canadian Hydrogen Intensity Mapping Experiment
- EHT: Event Horizon Telescope
- GBT: Green Bank Telescope
- Haystack[*]: MIT Haystack Observatory
- HERA[†]: Hydrogen Epoch of Reionization Array
- LOFAR: Low Frequency Array
- MWA[†]: Murchison Wide-field Array
- PAPER: Precision Array to Probe the Epoch of Reionization
- SKA: Square Kilometer Array
- VLA: Karl G. Jansky Very Large Array

Gravitational waves:

- Cosmic Explorer (PLANNED)
- KAGRA: Kamioka Gravitational Wave Detector
- LIGO*: Laser Interferometer Gravitational-wave Observatory
- LISA (PLANNED): Laser Interferometer Space Antenna
- Virgo: The Virgo Interferometer

Some of the most important catalogues and surveys of astronomical objects:

- 2MASS: 2 Micron All-Sky Survey (NIR all-sky survey)
- 3C: Third Cambridge Catalogue of Radio Sources (radio sources)
- Abell: Abell Catalogue of Rich Clusters of Galaxies (rich galaxy clusters, $z \lesssim 0.2$)
- BD: Bonner Durchmusterung catalogue (stars)
- BOSS/eBOSS: (extended) Baryon Oscillation Spectroscopic Survey (galaxies)
- C: Caldwell catalogue, a complement to Messier (109 bright star clusters, nebulae, and galaxies)
- DSS: Digitized Sky Survey (optical/NIR all-sky survey)
- FK5: The 5th Fundamental Catalogue (highly precise star locations used to define a reference frame)
  - Superseded by the ICRS: International Celestial Reference System, which uses quasars
- *Gaia* DR3: *Gaia* Data Release 3 catalogue (stars)
- HD: Henry Draper catalogue (stars)
- HIP: *Hipparcos* catalogue (stars)
- IC: Index Catalogue, supplement to the NGC (star clusters, nebulae, galaxies)
- KIC: *Kepler* Input Catalogue (stars/variables/transients)
- M: Messier catalogue (110 star clusters, nebulae, and galaxies)
- MACS: Massive Cluster Survey (X-ray luminous galaxy clusters selected from ROSAT)
- Mrk/Mk: Markarian Catalogue (galaxies with a UV-bright continuum, i.e. AGN)
- NGC: New General Catalogue of Nebulae and Clusters of Stars (star clusters, nebulae, galaxies, $z \lesssim 0.1$)
- PGC: Principal Galaxies Catalogue (galaxies)
- SDSS: Sloan Digital Sky Survey (optical/NIR all-sky survey)
- TIC: *TESS* Input Catalogue (stars/variables/transients)
- TOI: *TESS* Objects of Interest (exoplanet candidates)
- W: Westerhout catalogue (local radio sources—star-forming regions, molecular clouds)
- Zw: Zwicky Catalogue (galaxies and galaxy clusters)

12.6 Common Photometric Systems and Filters

Filter name	Central wavelength	FWHM
Johnson-Cousins	Å	Å
U	3650	680
B	4400	980
V	5500	890
R_C	6470	1515
I_C	7865	1090
Strömgren	Å	Å
u	3500	340
v	4100	200
b	4670	160
y	5470	240
SDSS	Å	Å
u'	3643	58
g'	4770	1263
r'	6231	1150
i'	7625	1239
z'	9134	994
JHKLM	μm	μm
J	1.25	0.38
H	1.65	0.48
K	2.2	0.70
L	3.5	1.20
L'	3.8	0.60
M	4.8	5.70
WISE	μm	μm
WISE 1	3.35	0.66
WISE 2	4.60	1.04
WISE 3	11.56	5.50
WISE 4	22.09	4.10

...and many many more![267,268,269]

12.7 Chemical Notations

Often in astrophysics we encounter chemical formulas and borrow the notations used in nuclear physics. This section is dedicated to explaining all of these notations since I often forget what all of the different numbers mean.

12.7.1 Atoms and Isotopes

For a lone atom, we'll often see it written as, e.g.,

$$^{12}_{6}\mathrm{C}$$

Here, the upper left number indicates the **atomic mass number**, which is the total number of protons **and** neutrons in the atomic nucleus. The lower left number indicates the **atomic number**, which is just the number of protons in the atomic nucleus. Often the atomic number is left off since it can be implied by the element (C always has 6 protons). But different isotopes will have different numbers of neutrons, so the atomic mass number can change, e.g.

$$^{13}_{6}\mathrm{C}$$

indicates an isotope of Carbon with 7 neutrons.

12.7.2 Molecules

In chemical formulas for molecules, we'll often see things like

$$\mathrm{H}_2$$

Here, the lower right number on the H indicates the **number of atoms** of H that are contained in the molecule.

12.7.3 Ions

In standard notation, an atom that has gained or lost electrons (called an ion) is typically written like

$$\mathrm{He}^{2+}$$

Here, the number on the upper right indicates the number of electrons that the atom has gained or lost. If there is a + sign, this indicates a positive charge, so the atom has **lost** this many electrons, and if there is a − sign, this indicates a negative charge, so the atom has **gained** this many electrons.

However, astrophysics actually has its own more commonly used notation for ions. In this notation, the above example would be written

$$\mathrm{He\,III}$$

Where the roman numerals designate the ionization state. I indicates a neutron atom, so He I is just He. II would indicate singly ionized, so the atom has lost one electron, i.e. He II is He^{+}. And so on. The number on the roman numerals is just 1 more than the number of electrons that have been lost.

Note that in this system, there is no way to indicate an ion that has **gained** electrons, so the standard notation is used in these cases (i.e. H^- is still H^-). Also notice that the roman numerals are written in "small caps" (H I, not H I).

12.7.4 Lines

Emission and absorption lines produced by an atom or ion are often notated by that atom/ion's symbol followed by the wavelength that the transition occurs at. For example,

$$\text{He II } \lambda 4686\,\text{Å}$$

If the line occurs in a **doublet**, both wavelengths are annotated, e.g.

$$\text{O VI } \lambda\lambda 1032, 1038\,\text{Å}$$

And if the transition is **forbidden**, the atom/ion is put in brackets, e.g.

$$[\text{Ne VI}]\ \lambda 7.652\,\mu\text{m}$$

There are also **semi-forbidden** transitions, which get an awkward half-bracket

$$\text{Si III}]\ \lambda 1892\,\text{Å}$$

For particularly common lines, they are often given shorthand notations. For example, the Balmer series of Hydrogen lines ($n \to 2$) is written

$$\begin{aligned}
\text{H}\alpha &= \text{H I } \lambda 6563\,\text{Å} = \text{H I } 3-2 \\
\text{H}\beta &= \text{H I } \lambda 4861\,\text{Å} = \text{H I } 4-2 \\
\text{H}\gamma &= \text{H I } \lambda 4340\,\text{Å} = \text{H I } 5-2 \\
\text{H}\delta &= \text{H I } \lambda 4102\,\text{Å} = \text{H I } 6-2 \\
\text{H}\epsilon &= \text{H I } \lambda 3970\,\text{Å} = \text{H I } 7-2 \\
\text{H}\zeta &= \text{H I } \lambda 3889\,\text{Å} = \text{H I } 8-2 \\
\text{H}\eta &= \text{H I } \lambda 3835\,\text{Å} = \text{H I } 9-2 \\
\text{H10} &= \text{H I } \lambda 3798\,\text{Å} = \text{H I } 10-2
\end{aligned}$$

On the far right column, the names are annotated by the different energy levels that produce the line. The first number is the upper energy level and the second number is the lower energy level. For example, H I $3-2$ is the line produced by an electron in the third orbital dropping to the second orbital. After Hη (or sometimes even before) people usually get lazy and just start writing the upper level of the transition instead of the next greek letter. Similarly, the other Hydrogen line series are named after the various people who discovered them and are written

$$\begin{aligned}
\text{Lyman Series } (n \to 1): &\quad \text{Ly}\alpha,\ \text{Ly}\beta,\ \text{Ly}\gamma,\ \text{etc.} \\
\text{Paschen Series } (n \to 3): &\quad \text{Pa}\alpha,\ \text{Pa}\beta,\ \text{Pa}\gamma,\ \text{etc.} \\
\text{Brackett Series } (n \to 4): &\quad \text{Br}\alpha,\ \text{Br}\beta,\ \text{Br}\gamma,\ \text{etc.}
\end{aligned}$$

Often for series above this, even for the two that have names, the named notation is dropped and they are just written in terms of the transition

$$\begin{aligned}\text{Pfund Series } (n \to 5): &\quad \text{H I } 6-5,\ \text{H I } 7-5,\ \text{H I } 8-5,\ \text{etc.}\\ \text{Humphreys Series } (n \to 6): &\quad \text{H I } 7-6,\ \text{H I } 8-6,\ \text{H I } 9-6,\ \text{etc.}\end{aligned}$$

Lines from **molecular transitions** get a different notation that is also based on the different energy levels. These molecular transitions happen due to rotations and vibrations in the molecule.

$$\text{H}_2\ 1-0\ S(3)$$

The first numbers $1-0$ indicate the transition in vibrational energy levels, while the $S(3)$ indicates the transition in rotational energy levels—the 3 is the lower energy level, and the S indicates that $\Delta J = 2$. Therefore, $S(3)$ indicates $J = 5 \to 3$.

There is also a subset of lines called the **Fraunhofer lines** that are given special labels. They were first observed in absorption in the solar spectrum by Joseph von Fraunhofer in 1814, before we knew that they were caused by atomic transitions, so they were originally just deesignated by capital letters. Today we know what elements cause these lines, but we still often use Fraunhofer's letters to identify them as a shorthand instead of using their wavelengths (except for the ones that correspond to Balmer lines and tellurics). They are written[270]

$$\begin{aligned}\text{Na I D} &= \text{Na I } \lambda\lambda 5896, 5890\,\text{Å}\\ \text{Fe I E} &= \text{Fe I } \lambda 5270\,\text{Å}\\ \text{G band} &= \text{Ca I } \lambda 4307.7\,\text{Å},\ \text{Fe I } \lambda 4307.9\,\text{Å},\ \text{CH},\ \text{and others}\\ \text{Ca II H and K} &= \text{Ca II } \lambda\lambda 3968, 3934\,\text{Å} \longrightarrow \text{start of 4000 Å break}\\ \text{Fe I L} &= \text{Fe I } \lambda 3820\,\text{Å}\\ \text{Fe I N} &= \text{Fe I } \lambda 3581\,\text{Å}\end{aligned}$$

Finally, we'll talk about **characteristic X-ray** line notation. This grew out of a way of labeling electron orbital shells called X-ray notation. Each shell gets a letter. This is sort of like how s, p, d, etc. label the orbital quantum number ℓ, but in this case we're labeling the principle quantum number n.

	K	L	M	N
$n =$	1	2	3	4

If X-ray radiation is incident on an atom, it can kick out electrons in the inner orbitals without removing the electrons in the outer orbitals, leaving a "hole" that can be filled by one of the outer orbital electrons. The X-ray line notation then labels the lines this produces by the lower orbital, e.g.

$$\begin{aligned}\text{Fe K}\alpha &= \text{Fe L} \to \text{K } (n = 2 \to 1, 6.40\,\text{keV})\\ \text{Fe K}\beta &= \text{Fe M} \to \text{K } (n = 3 \to 1, 7.06\,\text{keV})\\ \text{Fe L}\alpha &= \text{Fe M} \to \text{L } (n = 3 \to 2)\\ \text{Fe L}\beta &= \text{Fe N} \to \text{L } (n = 4 \to 2)\end{aligned}$$

Since the hole opens up regardless of the state of the outer shell electrons, these lines can be produced by multiple different ionization states of the same species, which is why the ionization state is not

annotated. This is *fundamentally different* than other lines, which are only produced by electrons transitioning in the outer shells of an atom/ion (the inner shells have no freedom to move between states). The iron lines above could be produced by Fe I all the way up to Fe XXVI (hydrogen-like ion with only 1 electron). Even if we have fewer than 3 electrons (so there are none in the L shell), we can still produce the lines through recombination with free electrons. This has the effect of broadening the lines since the different ionization states will produce *slightly* different transition energies[46]

12.7.5 Electron Configurations

Electron configurations in atomic physics indicate the distribution of electrons in an atom's or molecule's orbitals. The orbitals are filled up in the following order:

1s, 2s, 2p, 3s, 3p, 4s, 3d, 4p, 5s, 4d, 5p, 6s, 4f, 5d, 6p, 7s, 5f, 6d, 7p, 8s, 5g, 6f, 7d, 8p, 9s

In each of these, the first number indicates the principle quantum number n, where the nth shell can accommodate $2n^2$ electrons. Subshells are split up by their orbital angular momentum quantum number ℓ, each of which can accommodate $2(2\ell+1)$. The letter in each orbital indicates the value of ℓ. So $s:\ell=0$, $p:\ell=1$, $d:\ell=2$, $f:\ell=3$, etc. The number of electrons in each shell is indicated as a superscript. For example, for phosphorus (atomic number 15) we would have

$$1s^2 2s^2 2p^6 3s^2 3p^3$$

12.7.6 Term Symbols

Term symbols in atomic physics indicate an abbreviation for the total spin and orbital angular momentum quantum numbers of the electrons in an atom. The total orbital quantum number is L, the total spin quantum number is S, and the total angular momentum quantum number is J. We also have a parity of P which can be either $+1$ or -1. The term symbol then has the form

$$^{2S+1}L_J^P$$

S and J are written as actual numbers, but L is written as letters which each correspond to a number[271].

	S	P	D	F	G	H	I	K	L	M	N	O	Q	R	T	U	V
$L=$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

After V, this continues alphabetically. The parity P is also not written as a number—if the parity is even, nothing is written, but if it's odd, then a superscript letter "o" is written[271].

Term symbols are often written in combination with electron configurations. So for example, the ground state of a neutral carbon atom would be written

$$1s^2 2s^2 2p^2\,{}^3P_0$$

This tells us that the quantum numbers are $S=1$ (integer, boson-like), $L=1$, and $J=0$.

Now let's do an example for a transition that produces an emission line[272]:

$$[\mathrm{O\,III}]\ \lambda 5007\,\text{Å} = 1s^2 2s^2 2p^2\,{}^3P_2 - 1s^2 2s^2 2p^2\,{}^1D_2$$

For **molecular term symbols**, there are many similarities, but in particular the L number is written with greek instead of latin letters, and there are a number of additional notations to indicate symmetries:

$$^{2S+1}\Lambda^{(+/-)}_{\Omega,(g/u)}$$

Here, S as before is the total spin quantum number, Λ is the *projection* of L along the internuclear axis, and Ω is the *projection* of J along the internuclear axis. The g/u indicates symmetry/parity with respect to inversion, and the $+/-$ indicates symmetry along an arbitrary plane containing the internuclear axis[?] .

Electronic states are often labeled with a single letter before the term, with X being the ground state, and excited states with same multiplicity given capital letters in alphabetical order: A, B, C, etc. Excited states with different multiplicities are given lowercase letter in the same way: a, b, c, etc. Polyatomic molecules are given a tilde (i.e. $\tilde{X}$, $\tilde{a}$). For example, the ground state of H_2 would be written

$$X^1\Sigma^+_g$$

And the first two binding electronic states with electric dipole transitions from the ground state are

$$B^1\Sigma^+_u \text{ and } C^1\Pi_u \tag{12.1}$$

Transitions from the ground state to these two states are called the **Lyman and Werner bands** respectively[65].

12.8 Nuclear Processes

12.8.1 Triple-alpha Process

As the name suggests, it involves the fusion of three alpha particles (i.e. ^{4}He nuclei) into a ^{12}C. This can only happen in very massive stars with hot core temperatures (or post-main sequence stars), since you need to have 3 alpha particles in close proximity to fuse at nearly the same time. This is necessary because the intermediate product, ^{8}Be, has a very short half-life and decays almost immediately:

$$^4\text{He} + {}^4\text{He} \longrightarrow {}^8\text{Be} + \gamma \tag{12.2}$$

$$^8\text{Be} + {}^4\text{He} \longrightarrow {}^{12}\text{C} + \gamma \tag{12.3}$$

In the second step, a ^{12}C is created in an excited state. Most of the time, it will just decay back into 3 ^{4}He, but there is a non-negligible fraction that decays into the ground state of ^{12}C and therefore produces some Carbon in the star. There exists a coincidental *resonance* in this reaction, meaning that the energy of the beryllium and helium is almost exactly the same as that of the excited state of Carbon. This increases the cross section of the reaction and makes it much more likely than it would normally be. This is called the **Hoyle resonance**, named after Fred Hoyle, who predicted its existence based on the fact that it must exist in order to explain the observed amount of Carbon in stars. It was confirmed not long after with the observation of an emission line at ∼7.65 MeV corresponding to $^{12}\text{C}^* \longrightarrow {}^{12}\text{C} + \gamma$.

The energy generation of the triple alpha process as a function of temperature is even more steep than the CNO cycle, having a general scaling of

$$\boxed{\varepsilon_m(3\alpha) \propto \rho^2 Y^3 T^{41}} \tag{12.4}$$

where Y is the helium fraction[11]. This is compared to the pp-chain and CNO cycle in Figure 12.3.

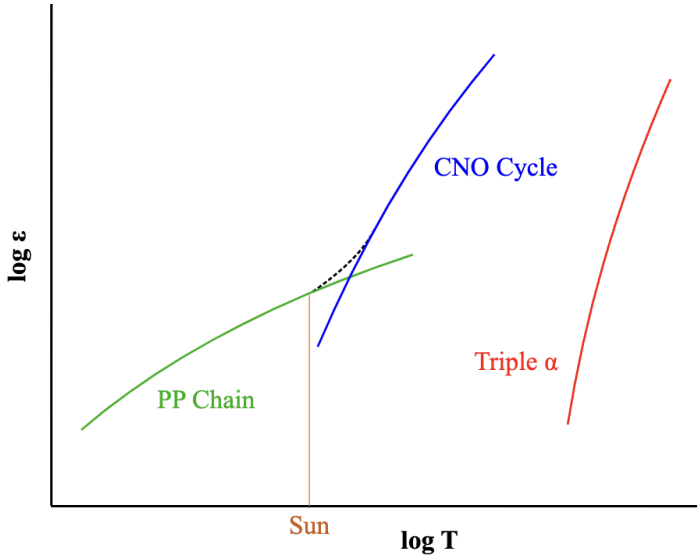


Figure 12.3: The energy generation rates of the pp-chain, CNO cycle, and triple-alpha process as a function of temperature.

12.8.2 s-process

Meaning "slow" process. This, along with the r-process ("rapid" process, see next section), explain how we get elements heavier than Fe occurring naturally in our universe.

In the s-process, we take an Fe nucleus and keep adding neutrons to it slowly, one at a time, allowing it to undergo a β decay between each addition and move up to the next element. This mostly happens in 2nd generation stars that already had some Fe in them to start with. By the end of the chain, we reach a stable point at lead:

$$\text{Hg} \longrightarrow \text{Tl} \longrightarrow \text{Pb} \longrightarrow \text{Bi} \longrightarrow \text{Po} \longrightarrow \text{alpha decay back to Pb} \tag{12.5}$$

12.8.3 r-process

The counterpart to the s-process. In the r-process, we add many neutrons quickly, not giving time for anything to β decay, and just run up to the neutron drip line. Then we wait for everything to β decay into stable states. For this to occur, we need a free neutron density of $\sim 1\ \text{g}\,\text{cm}^{-3}$ (giving ~100 neutron captures/second). This happens in **neutron star–neutron star mergers** and the **ejecta of core-collapse supernovae**.

The sources of all elements is shown in Figure 12.4.

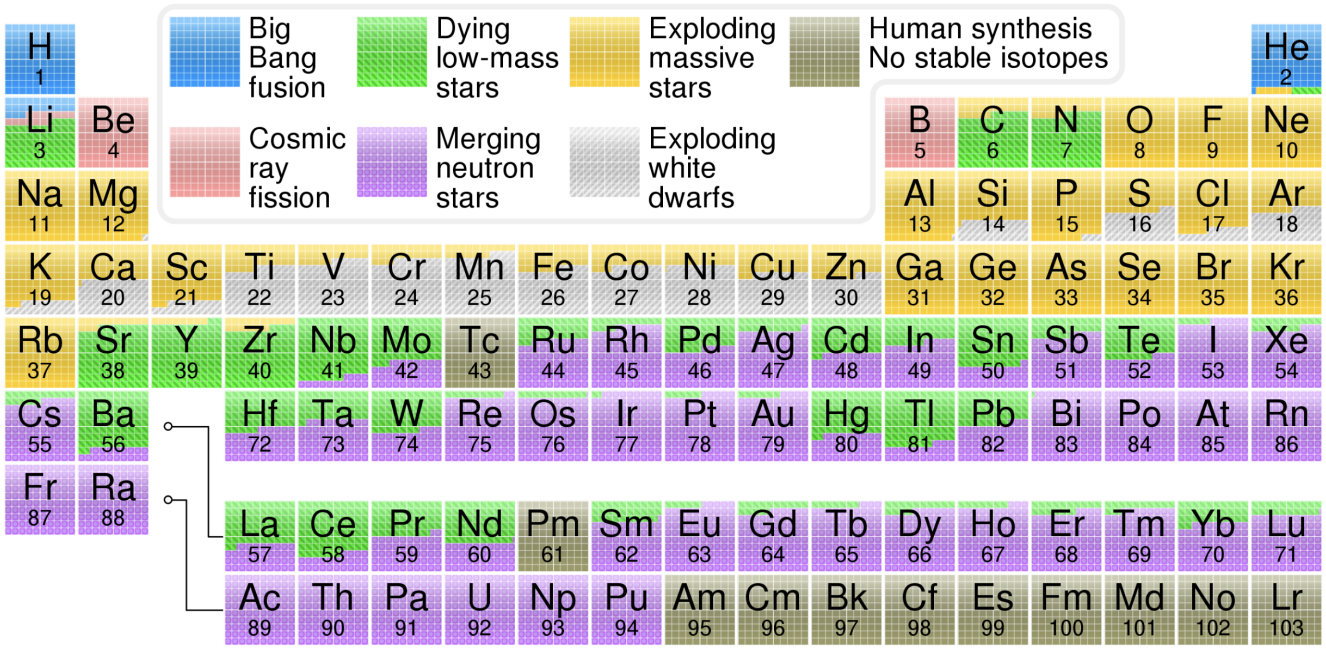


Figure 12.4: The periodic table of the elements, where each element is colored according to how much of the current abundance in the universe is attributed to different nuclear processes, including Big Bang Nucleosynthesis, stellar nucleosynthesis, s-process and r-process, and purely human synthesized elements. Notice the outlier, Technetium (Tc). This, along with the heavy elements above Plutonium (Pu), *may* be produced by some of the nucleosynthesis processes, but they have no stable isotopes, so they quickly decay.

12.9 Orbital Dynamics

12.9.1 Kepler's Laws

In this section, we'll derive Kepler's 3 laws of orbital motion from first principles.

Kepler's First Law

Kepler's 1st law says that planets orbit in ellipses. To prove this, we'll work in the reduced 1-body problem where M is the total mass, μ is the reduced mass, $\mathbf{r}$ is the relative position of the two bodies, and $\mathbf{R}$ is the center of mass position, which we set to the origin. See §3.Q20 for definitions of these quantities. We only have a conservative force (gravity), so this is the perfect situation to use Lagrangian mechanics. Our lagrangian is

$$\mathscr{L} = T - U = \frac{1}{2}\mu v^2 + \frac{GM\mu}{r} \tag{12.6}$$

Let's work in plane-polar coordinates (r, ϕ):

$$\mathbf{r} = r\hat{\mathbf{r}} \ , \ \dot{\hat{\mathbf{r}}} = \dot{\phi}\hat{\boldsymbol{\phi}} \ , \ \dot{\hat{\boldsymbol{\phi}}} = -\dot{\phi}\hat{\mathbf{r}} \tag{12.7}$$

$$\dot{\mathbf{r}} = \dot{r}\hat{\mathbf{r}} + r\dot{\hat{\mathbf{r}}} = \dot{r}\hat{\mathbf{r}} + r\dot{\phi}\hat{\boldsymbol{\phi}} \quad \longrightarrow \quad v^2 = \dot{r}^2 + r^2\dot{\phi}^2 \tag{12.8}$$

$$\therefore \quad \mathscr{L} = \frac{1}{2}\mu(\dot{r}^2 + r^2\dot{\phi}^2) + \frac{GM\mu}{r} \tag{12.9}$$

Now, using the Euler-Lagrange equations:

$$\frac{\partial \mathscr{L}}{\partial q} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathscr{L}}{\partial \dot{q}}\right) \tag{12.10}$$

We get two differential equations for r and ϕ:

$$\mu\ddot{r} = \mu r\dot{\phi}^2 - \frac{GM\mu}{r^2} \tag{12.11}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mu r^2\dot{\phi}) = 0 \tag{12.12}$$

The second equation gives us the conservation of angular momentum

$$\ell \equiv \mu r^2\dot{\phi} = \text{constant} \tag{12.13}$$

Then we can rewrite the first equation as

$$\mu\ddot{r} - \frac{\ell^2}{\mu r^3} + \frac{GM\mu}{r^2} = 0 \tag{12.14}$$

Now let's change from time derivatives to angle derivatives

$$\dot{r} = \frac{\mathrm{d}r}{\mathrm{d}t} = \frac{\mathrm{d}r}{\mathrm{d}\phi}\frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{\mathrm{d}r}{\mathrm{d}\phi}\dot{\phi} = \frac{\ell}{\mu r^2}\frac{\mathrm{d}r}{\mathrm{d}\phi} \tag{12.15}$$

$$\ddot{r} = \frac{\mathrm{d}^2r}{\mathrm{d}t^2} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\ell}{r^2}\frac{\mathrm{d}r}{\mathrm{d}\phi}\right) = \frac{\mathrm{d}}{\mathrm{d}\phi}\left(\frac{\ell}{\mu r^2}\frac{\mathrm{d}r}{\mathrm{d}\phi}\right)\dot{\phi} = \frac{\ell^2}{\mu^2 r^2}\frac{\mathrm{d}}{\mathrm{d}\phi}\left(\frac{1}{r^2}\frac{\mathrm{d}r}{\mathrm{d}\phi}\right) \tag{12.16}$$

And now we'll recast the problem in terms of the variable $u = 1/r$:

$$\frac{\mathrm{d}u}{\mathrm{d}\phi} = -\frac{1}{r^2}\frac{\mathrm{d}r}{\mathrm{d}\phi} \quad \longrightarrow \quad \ddot{r} = -\frac{\ell^2}{\mu^2}u^2\frac{\mathrm{d}^2u}{\mathrm{d}\phi^2} \tag{12.17}$$

Plugging this into our differntial equation and rewriting everything in terms of u, we have

$$\frac{\mathrm{d}^2u}{\mathrm{d}t^2} + u - \frac{GM\mu^2}{\ell^2} = 0 \tag{12.18}$$

Now we'll make one more substitution, $w = u - GM\mu^2/\ell^2$, such that

$$\frac{\mathrm{d}^2w}{\mathrm{d}t^2} = -w \tag{12.19}$$

This is just a simple harmonic oscillator! We have solutions of the form

$$w(\phi) = A\cos(\phi - \phi_0) \tag{12.20}$$

Therefore, converting this back to $r = 1/(w + GM\mu^2/\ell^2)$, we get

$$r(\phi) = \frac{\ell^2/GM\mu^2}{1 + (\ell^2 A/GM\mu^2)\cos(\phi - \phi_0)} \tag{12.21}$$

This is just the equation of an ellipse with an eccentricity $e = \ell^2 A/GM\mu^2$ and a semimajor axis $a = \ell^2/GM\mu^2(1-e^2)$.

$$\boxed{r(\phi) = \frac{a(1-e^2)}{1 + e\cos\phi}} \tag{12.22}$$

This gives the relation

$$\boxed{\ell^2 = GM\mu^2 a(1-e^2)} \tag{12.23}$$

Kepler's Second Law

Kepler's 2nd law states that planets sweep out equal areas in equal times during their orbits. This is essentially a restatement of conservation of angular momentum. Consider a small time interval $\mathrm{d}t$ during with the planet, at radius r, covers an angle $\mathrm{d}\phi$. Assuming $r(\phi + \mathrm{d}\phi) \approx r(\phi)$ (the change in radius is negligible), the area swept out during this time interval is just a triangle with a base of r and a height of $r\mathrm{d}\phi$, so

$$\mathrm{d}A = \frac{1}{2}r^2\mathrm{d}\phi \quad \longrightarrow \quad \boxed{\frac{\mathrm{d}A}{\mathrm{d}t} = \frac{1}{2}r^2\frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{\ell}{2\mu} = \text{constant}} \tag{12.24}$$

Since ℓ and μ are both constants, $\mathrm{d}A/\mathrm{d}t$ must also be constant! QED.

Kepler's Third Law

Kepler's 3rd law says that the orbital period squared, P^2, is proportional to the semimajor axis cubed, a^3. We can find this by recognizing that the orbital period is just the total area of the orbit A over the

area rate of change $\mathrm{d}A/\mathrm{d}t$, which we proved in the 2nd law is constant:

$$P = \frac{A}{\mathrm{d}A/\mathrm{d}t} = \frac{\pi a b}{\ell/2\mu} = \frac{2\pi a b \mu}{\ell} \tag{12.25}$$

The semiminor axis is related to the semimajor axis by

$$b = a\sqrt{1-e^2} \tag{12.26}$$

And we can use equation (12.23) for ℓ

$$P = 2\pi \frac{a^2\sqrt{1-e^2}}{\sqrt{GMa(1-e^2)}} \longrightarrow \boxed{P^2 = \frac{4\pi^2}{GM}a^3} \tag{12.27}$$

12.9.2 The Virial Theorem

This section will be dedicated to deriving the virial theorem, which I use often in these notes, but no question directly asks for a proof. Let's start by assuming we have a population of stars orbiting in a galaxy with masses m_i, positions $\mathbf{r}_i$, and momenta $\mathbf{p}_i = m_i\mathbf{v}_i$. We define a quantity called the "virial" as

$$G \equiv \sum_i \mathbf{p}_i \cdot \mathbf{r}_i = \sum_i m_i \dot{\mathbf{r}}_i \cdot \mathbf{r}_i \tag{12.28}$$

We can rewrite this by using the derivative of r_i^2, since

$$\frac{\mathrm{d}}{\mathrm{d}t}(r_i^2) = \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{r}_i \cdot \mathbf{r}_i) = \dot{\mathbf{r}}_i \cdot \mathbf{r}_i + \mathbf{r}_i \cdot \dot{\mathbf{r}}_i = 2\dot{\mathbf{r}}_i \cdot \mathbf{r}_i \tag{12.29}$$

In which case we obtain

$$G = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\sum_i m_i r_i^2 \tag{12.30}$$

Now we'll recall that the moment of inertia about the origin is $I = \sum_i m_i r_i^2$, meaning we can write

$$G = \frac{1}{2}\frac{\mathrm{d}I}{\mathrm{d}t} \tag{12.31}$$

Now let's look at the time derivative of the virial

$$\frac{\mathrm{d}G}{\mathrm{d}t} = \frac{1}{2}\frac{\mathrm{d}^2 I}{\mathrm{d}t^2} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_i \mathbf{p}_i \cdot \mathbf{r}_i\right) \tag{12.32}$$

$$= \sum_i \dot{\mathbf{p}}_i \cdot \mathbf{r}_i + \sum_i \mathbf{p}_i \cdot \dot{\mathbf{r}}_i \tag{12.33}$$

$$= \sum_i \mathbf{F}_i \cdot \mathbf{r}_i + \sum_i m_i v_i^2 \tag{12.34}$$

The first term $\mathbf{F}_i \cdot \mathbf{r}_i$ is just the work required to place a particle at $\mathbf{r}_i$ from ∞, and for a conservative force like gravity, this is just the potential energy U. The second term, by inspection, is just twice the

kinetic energy $2T$. Therefore

$$\frac{\mathrm{d}^2 I}{\mathrm{d}t^2} = U + 2T \tag{12.35}$$

If a system is in a steady-state, it is said to be **virialized**. In this case, the moment of inertia should not have a second time derivative $\mathrm{d}^2 I \big/ \mathrm{d}t^2 = 0$, so we get the **virial theorem**

$$\boxed{2T + U = 0} \tag{12.36}$$

It's important to remember the assumptions we made in deriving this equation:

1. The only forces **F** acting on the system are conservative
2. The system is virialized, i.e. $\mathrm{d}^2 I \big/ \mathrm{d}t^2 = 0$.

12.10 Equations of State

12.10.1 Equation of State Regimes

There are a few different equations of state of interest:

$$\text{Classical ideal gas:} \quad P = \frac{\rho}{\mu m_p} kT \tag{12.37}$$

$$\text{Non-relativistic degenerate:} \quad P = K_{\text{NR}} \left(\frac{\rho}{\mu} \right)^{5/3} \tag{12.38}$$

$$\text{Ultra-relativistic degenerate:} \quad P = K_{\text{UR}} \left(\frac{\rho}{\mu} \right)^{4/3} \tag{12.39}$$

$$\text{Radiation pressure:} \quad P = \frac{1}{3} a T^4 \tag{12.40}$$

These become important in different regimes of temperature and density, as shown in Figure 12.5.

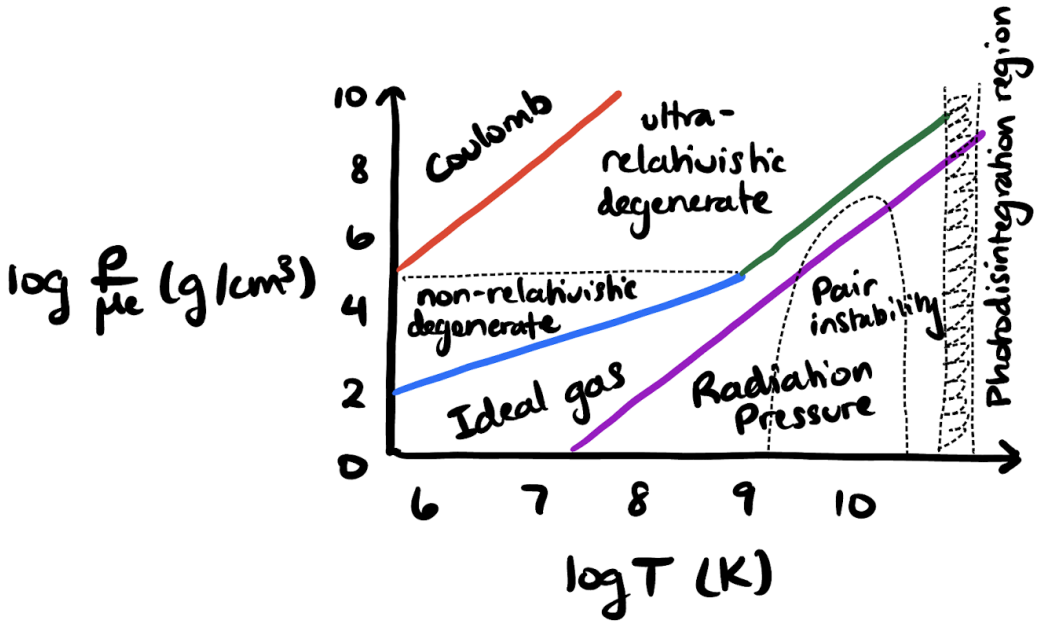


Figure 12.5: The dominant equation of state across different regimes of temperature and density. Note also the regions of pair instability and photodisintegration.

12.10.2 Polytropes

We often model the interiors of stars and stellar remnants using **polytropic** equations of state, with the form

$$P = K\rho^{\gamma} = K\rho^{1+1/n} \tag{12.41}$$

where γ is the adiabatic index and $n = 1/(\gamma - 1)$ is the polytropic index. When we have an equation of state of this form, we can formulate a framework for modeling interior stellar density, temperature,

pressure, etc. profiles. First, using hydrostatic equilibrium:

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{GM\rho}{r^2} \quad \longrightarrow \quad \frac{r^2}{\rho}\frac{\mathrm{d}P}{\mathrm{d}r} = -GM \tag{12.42}$$

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{r^2}{\rho}\frac{\mathrm{d}P}{\mathrm{d}r}\right) = -G\frac{\mathrm{d}M}{\mathrm{d}r} \tag{12.43}$$

$$\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{r^2}{\rho}\frac{\mathrm{d}P}{\mathrm{d}r}\right) = -4\pi G\rho \tag{12.44}$$

Notice that this is just a Poisson's equation. Now, we can define the density profile $\rho(r) = \rho_c\Theta^n(r)$, where ρ_c is the core density and $\Theta(r)$ is a dimensionless parameter than spans from $\Theta(0) = 1$ to $\Theta(R) = 0$. Substituting this into our Poisson's equation:

$$\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{r^2}{\rho_c}\frac{P_c}{\Theta^n}\frac{\mathrm{d}}{\mathrm{d}r}\Theta^{1+n}\right) = -4\pi G\rho_c\Theta^n \tag{12.45}$$

$$\left(\frac{n+1}{4\pi G}\frac{P_c}{\rho_c^2}\right)\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}\Theta}{\mathrm{d}r}\right) = -\Theta^n \tag{12.46}$$

Now we make some definitions

$$a^2 \equiv \frac{n+1}{4\pi G}\frac{P_c}{\rho_c^2} \quad , \quad \xi \equiv \frac{r}{a} \tag{12.47}$$

Such that we can rewrite the Poisson's equation in a purely dimensionless form

$$\boxed{\frac{1}{\xi^2}\frac{\mathrm{d}}{\mathrm{d}\xi}\left(\xi^2\frac{\mathrm{d}\Theta}{\mathrm{d}\xi}\right) = -\Theta^n} \tag{12.48}$$

This is the **Lane-Emden equation**. To recap, this is a dimensionless version of a Poisson's equation that describes the interior of a polytropic fluid in hydrostatic equilibrium. It may or may not produce physically realistic solutions depending on the choice of n, but it is typically easier to work with than the set of 4 coupled differential stellar structure equations. Some typical solutions for real objects are:

$$n = \begin{cases} 0 & \text{rocky planets} \\ 0-1 & \text{neutron stars} \\ 3/2 & \text{convective zone / non-relativistic degenerate WD } (\gamma = 5/3) \\ 3 & \text{radiative zone / relativistic degenerate WD } (\gamma = 4/3) \\ \infty & \text{isothermal sphere} \end{cases} \tag{12.49}$$

12.11 Radiative Transfer

12.11.1 Specific Intensity

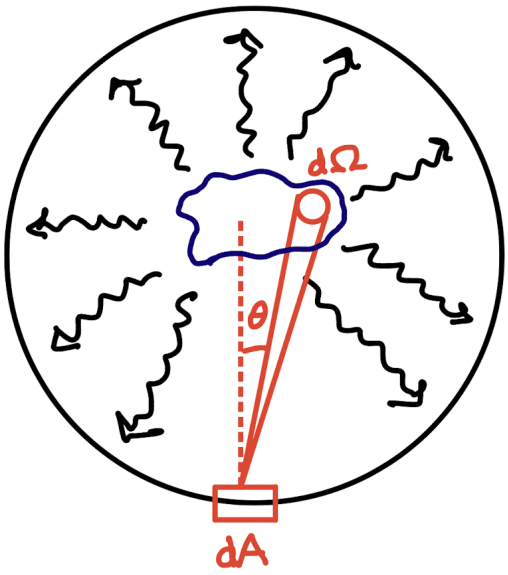


Figure 12.6: Basic radiative transfer from a blob of material.

The most fundamental quantity in radiative transfer is the **specific intensity** (also sometimes called **surface brightness**) I_ν, which is defined as the energy per unit time per unit area per unit solid angle per unit frequency of light that we receive from some object. There is an additional factor of $\cos\theta$ that is present such that we only account for the component of radiation that travels parallel to our infinitesimal area. Thus, I_ν is defined by:

$$\boxed{\mathrm{d}E = I_\nu \cos\theta \, \mathrm{d}t \, \mathrm{d}A \, \mathrm{d}\Omega \, \mathrm{d}\nu} \quad , \quad [I_\nu] = \mathrm{erg\,s^{-1}\,cm^{-2}\,sr^{-1}\,Hz^{-1}} \tag{12.50}$$

Where does each term come from?

- The area term is there because, as light radiates outward from an object, it expands in all directions, creating this large "shell of light"—the amount of light received at some location away from the source will only be the light that traveled in that direction, so it only collects the light that covers a small infinitesimal area ($\mathrm{d}A$) of this expanding shell.
- The solid angle term is there because the incoming light rays onto the infinitesimal area may be different based on what direction they are coming in from. So the solid angle measures the infinitesimal angle ($\mathrm{d}\Omega$) that we collect light from on the plane of the sky. This becomes particularly important when considering the radiation field *inside* certain objects (for example, in the core of a star) where there is light coming from all directions.
- The time term is there because we want to measure the rate of energy we are receiving, which can change over time and tells us a lot about the physical processes going on inside an object.
- The frequency term is there because we may receive different amounts of radiation at different frequencies, which again is a useful tool to tell us what kinds of physical processes are going on to produce this radiation.

12.11.2 Specific Flux

The **specific flux** F_ν is defined as the net energy per unit time per unit area per unit frequency coming from all directions. As such, it is related to the specific intensity by integrating over the solid angle.

$$\mathrm{d}F_\nu = \frac{\mathrm{d}E}{\mathrm{d}t\,\mathrm{d}A\,\mathrm{d}\nu} = I_\nu \cos\theta\,\mathrm{d}\Omega \quad \longrightarrow \quad \boxed{F_\nu = \int \mathrm{d}\Omega\, I_\nu \cos\theta} \quad , \quad [F_\nu] = \mathrm{erg\,s^{-1}\,cm^{-2}\,Hz^{-1}} \tag{12.51}$$

12.11.3 Specific Luminosity

Then we have the **specific luminosity** L_ν, which is defined as the energy per unit time per unit frequency. This is the most intrinsic measurement of an object's overall brightness as we integrate over both the solid angle and the area to remove the effects of the object's distance from the point where it is being measured (which reduces the specific flux by $1/r^2$):

$$\mathrm{d}L_\nu = \frac{\mathrm{d}E}{\mathrm{d}t\,\mathrm{d}\nu} = I_\nu \cos\theta\,\mathrm{d}A\,\mathrm{d}\Omega \quad \longrightarrow \quad \boxed{L_\nu = \int \mathrm{d}A\, F_\nu} \quad , \quad [L_\nu] = \mathrm{erg\,s^{-1}\,Hz^{-1}} \tag{12.52}$$

$$\boxed{L_\nu = 4\pi r^2 F_\nu} \text{ (if the source is isotropic)} \tag{12.53}$$

12.11.4 Specific Mean Intensity

Note there is also the **specific mean intensity** J_ν which is obtained by averaging I_ν over the solid angle:

$$\boxed{J_\nu = \frac{1}{4\pi}\int \mathrm{d}\Omega\, I_\nu} \tag{12.54}$$

such that for an isotropic source, $J_\nu = I_\nu$.

12.11.5 Specific Energy Density

The **specific energy density** u_ν is the energy per unit volume per unit frequency. Recall that photons travel at the speed c, so we have to take the specific intensity (integrated over the solid angle), which gives the energy per unit area per unit time per unit frequency, and divide by the speed to get the energy per unit volume per unit frequency. Therefore, it's given by

$$\boxed{u_\nu = \frac{1}{c}\int \mathrm{d}\Omega\, I_\nu = \frac{4\pi}{c} J_\nu} \quad , \quad [u_\nu] = \mathrm{erg\,cm^{-3}\,Hz^{-1}} \tag{12.55}$$

Thus, in an isotropic radiation field, $u_\nu = (4\pi/c) I_\nu$.

12.11.6 Radiation Pressure

Since photons have a momentum E/c, the momentum per unit area, time, solid angle, and frequency is I_ν/c. If we imagine these photons bouncing off a mirror at angle θ, they transfer a momentum $2p\cos\theta$. Therefore, we need another factor of $\cos\theta$ to get the radiation pressure:

$$\boxed{P_\nu = \frac{1}{c}\int \mathrm{d}\Omega\, I_\nu \cos^2\theta} \quad , \quad [P_\nu] = \mathrm{dyne\,cm^{-2}\,Hz^{-1}} \tag{12.56}$$

The factor of 2 was only relevant for the mirror example—this expression just gives the radiation pressure along a ray. Once again, it is useful to examine the case of isotropic radiation, where we can pull I_ν out of the integral. Substituting in the specific energy density, we find

$$\boxed{P_\nu = \frac{1}{3}u_\nu} \quad \text{(if the source is isotropic)} \tag{12.57}$$

12.11.7 Bolometric Quantities

By integrating over the frequency, we can get the **intensity**, **flux**, and **luminosity** (also sometimes referred to as the **bolometric** intensity/flux/luminosity to clarify that it is over all frequencies):

$$\boxed{I = \int \mathrm{d}\nu\, I_\nu} \quad , \quad [I] = \mathrm{erg\,s^{-1}\,cm^{-2}\,sr^{-1}} \tag{12.58}$$

$$\boxed{F = \int \mathrm{d}\nu\, F_\nu} \quad , \quad [F] = \mathrm{erg\,s^{-1}\,cm^{-2}} \tag{12.59}$$

$$\boxed{L = \int \mathrm{d}\nu\, L_\nu} \quad , \quad [L] = \mathrm{erg\,s^{-1}} \tag{12.60}$$

...etc for all of the other quantities.

12.11.8 Per Unit Wavelength Quantities

So far we have been working exclusively in per unit frequency units, but we can easily convert to per unit wavelength units if desired by enforcing the rule that the integrated specific flux F_ν between ν and $\nu + \mathrm{d}\nu$ should be equal to the equivalent per unit wavelength flux F_λ between λ and $\lambda + \mathrm{d}\lambda$, i.e.

$$F_\nu\,\mathrm{d}\nu = F_\lambda\,\mathrm{d}\lambda \quad \longrightarrow \quad F_\lambda = F_\nu\left|\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\right| \tag{12.61}$$

Using $\lambda\nu = c$, we have $\nu = c/\lambda$ and $\mathrm{d}\nu/\mathrm{d}\lambda = -c/\lambda^2$, so:

$$\boxed{F_\lambda = F_\nu\frac{c}{\lambda^2} = F_\nu\frac{\nu^2}{c}} \quad , \quad [F_\lambda] = \mathrm{erg\,s^{-1}\,cm^{-2}\,\mathring{A}^{-1}} \tag{12.62}$$

where the wavelength λ is most often measured in ångströms ($1\,\text{Å} = 10^{-10}\,\text{m}$).

12.11.9 In the context of distant objects

Now, in the case of observing distant objects (i.e. with imaging or spectroscopy from detectors on Earth), the radiation field we observe becomes highly non-isotropic. We point our detectors at the object of interest, so it's basically all parallel light rays coming from one localized direction (note in most cases the radiation field *emitted* by the object is still isotropic). Taking a look at Figure 12.6 again, for distant objects, we can consider $\theta \approx 0$ such that $\cos\theta \approx 1$, since the light rays will be parallel. Even for extended objects that cover multiple pixels, each individual pixel will have its own incident radiation that is essentially parallel. Therefore, we have

$$\mathrm{d}E \approx I_\nu \,\mathrm{d}t\,\mathrm{d}A\,\mathrm{d}\Omega\,\mathrm{d}\nu \tag{12.63}$$

And

$$F_\nu \approx \int \mathrm{d}\Omega\, I_\nu \tag{12.64}$$

We can then think of I_ν essentially like a δ-function that's equal to $I_{\nu,0}$ when pointing at the object of interest, and 0 when in any other direction. Let's say the object is at position (0,0) and has extent $(\Delta x, \Delta y)$. Then we have

$$I_\nu(x,y) = I_{\nu,0} \times \delta(x,y) \;\text{ where }\; \delta(x,y) = \begin{cases} 1 & x \leqslant \Delta x \;\text{ and }\; y \leqslant \Delta y \\ 0 & \text{otherwise} \end{cases} \tag{12.65}$$

$$F_\nu \approx I_{\nu,0} \int \mathrm{d}\Omega\, \delta(x,y) = I_{\nu,0} \frac{1}{r^2} \int \mathrm{d}A\, \delta(x,y) \tag{12.66}$$

$$\approx I_{\nu,0} \frac{1}{r^2} \int_0^{\Delta x} \mathrm{d}x \int_0^{\Delta y} \mathrm{d}y = I_{\nu,0} \times \frac{\Delta x \Delta y}{r^2} \tag{12.67}$$

$$\tag{12.68}$$

$$\therefore\; F_\nu \approx I_{\nu,0} \times \Delta\Omega \tag{12.69}$$

where $\Delta\Omega = \Delta x \Delta y / r^2$ is the angular size of the object of interest on the sky (or, if converting on a per-pixel basis, $\Delta\Omega$ is the angular size per pixel, and assuming square pixels, $\Delta\theta = \Delta x / r = \Delta y / r$ such that $\Delta\Omega = \Delta\theta^2$). Then for the luminosity

$$L_\nu = 4\pi r^2 F_\nu \approx 4\pi r^2 I_{\nu,0} \Delta\Omega \tag{12.70}$$

where r is the distance from us to the object. Remember, we can do this because the radiation field *emitted* by the object is still isotropic (independent of angle), even though the radiation field we see is not.

12.11.10 A note on the units of I_ν

The per solid angle (sr^{-1}) units may be eliminated in favor of measuring the per area units in the context of the area on the object rather than the area at the location of measurement. For example, using the approximation above, say we have an angle $\Delta\theta = D/r$ where r is the distance from us to the object and D is the length on the object that the angle $\Delta\theta$ sweeps out. Then we can rewrite I_ν as

$$I_\nu \approx \frac{F_\nu}{\Delta\theta^2} = F_\nu\left(\frac{r}{D}\right)^2 = \frac{L_\nu}{4\pi r^2}\left(\frac{r}{D}\right)^2 = \frac{L_\nu}{4\pi D^2} \tag{12.71}$$

Therefore the units of I_ν become $\mathrm{erg\,s^{-1}\,cm^{-2}\,Hz^{-1}}$ with the cm now referring to cm on the object instead of on our detector, and one can integrate I_ν over the size of the object to obtain the luminosity:

$$L_\nu = \int \mathrm{d}A_{\mathrm{obj}}\, I_\nu \tag{12.72}$$

12.11.11 Emissivity

Now we'll start getting into radiative transfer. A ray passing through matter may gain or lose energy due to emission, absorption, and/or scattering. The **spontaneous emission coefficient** j_ν is defined as the energy emitted per unit time per unit volume per unit solid angle per unit frequency.

$$\boxed{\mathrm{d}E = j_\nu \mathrm{d}t\,\mathrm{d}V\,\mathrm{d}\nu\,\mathrm{d}\Omega} \quad , \quad [j_\nu] = \mathrm{erg\,s^{-1}\,cm^{-3}\,sr^{-1}\,Hz^{-1}} \tag{12.73}$$

Thus, the intensity added to the beam $\mathrm{d}I_\nu$ is

$$\mathrm{d}I_\nu = j_\nu \mathrm{d}\ell \tag{12.74}$$

The emission is also sometimes described by the **emissivity** ε_ν which is integrated over the solid angle

$$\boxed{\varepsilon_\nu = \int \mathrm{d}\Omega\, j_\nu} \quad , \quad [\varepsilon_\nu] = \mathrm{erg\,s^{-1}\,cm^{-3}\,Hz^{-1}} \tag{12.75}$$

In an isotropic distribution, this is just

$$\varepsilon_\nu = 4\pi j_\nu \tag{12.76}$$

12.11.12 Absorption

The **absorption coefficient** defines the fraction of intensity expected to be lost to absorption per unit length. Therefore we can quantify it in terms of the number density n and cross section σ, OR equivalently the mass density ρ and the opacity κ:

$$\boxed{\alpha_\nu = n\sigma_\nu = \rho\kappa_\nu} \quad , \quad [\alpha_\nu] = \mathrm{cm^{-1}} \tag{12.77}$$

Such that the change in intensity due to absorption is

$$\mathrm{d}I_\nu = -\alpha_\nu I_\nu \mathrm{d}\ell = -I_\nu \mathrm{d}\tau_\nu \tag{12.78}$$

The **optical depth** τ_ν is the absorption coefficient integrated over the path length, and thus it is unitless

$$\tau_\nu = \int \mathrm{d}\ell\, \alpha_\nu \tag{12.79}$$

12.11.13 Equation of Radiative Transfer

Combining the effects of emission and absorption, the change in intensity over some path length ℓ is given by the **equation of radiative transfer (RTE)**

$$\boxed{\frac{\mathrm{d}I_\nu}{\mathrm{d}\ell} = j_\nu - \alpha_\nu I_\nu} \tag{12.80}$$

In the simple case where we have no absorption, we have

$$I_\nu(\ell) = I_\nu(0) + \int_0^\ell \mathrm{d}\ell\, j_\nu(\ell) \tag{12.81}$$

And in the other simple case where we have no emission, we have

$$I_\nu(\ell) = I_\nu(0) \exp\left[-\int_0^\ell \mathrm{d}\ell\, \alpha_\nu(\ell) \right] = I_\nu(0) e^{-\tau_\nu} \tag{12.82}$$

And in the generic case, defining the **source function** $S_\nu \equiv j_\nu/\alpha_\nu$, we have

$$I_\nu(\tau_\nu) = I_\nu(0) e^{-\tau_\nu} + \int_0^{\tau_\nu} \mathrm{d}\tau'_\nu\, e^{-(\tau_\nu - \tau'_\nu)} S_\nu(\tau'_\nu) \tag{12.83}$$

12.11.14 Blackbody Radiation

In this section we're going to do a quick derivation of Planck's Law (ooh, fun)![273]

First, what is **blackbody radiation**? Well, a **blackbody** is an idealized body that absorbs all incident radiation and does not reflect anything. Therefore, blackbody radiation is the radiation that such a body emits when it's in thermal equilibrium with itself and its environment.

We can start by imagining the boundaries of the black body as the walls of a box. Electromagnetic waves propagating inside the box have the boundary condition that they must go to 0 at the edges of the box (i.e. standing waves). Therefore, they must have the form

$$\psi(\mathbf{r}, t) = A \sin\left(k_x x\right) \sin\left(k_y y\right) \sin\left(k_z z\right) \sin\left(\omega t\right) \tag{12.84}$$

Where we had to take sin in each dimension due to the boundary conditions. The wavevector $\mathbf{k} \equiv$

(k_x, k_y, k_z) is quantized by the size of the box L (assuming the same size in every dimension).

$$n_i \frac{\lambda_i}{2} = L \quad \longrightarrow \quad k_i = \frac{\pi}{L} n_i \tag{12.85}$$

(recall $k = 2\pi/\lambda$). Here n_i are integers that enumerate the states in each dimension. Now consider the 3D space of $\mathbf{n} \equiv (n_x, n_y, n_z)$. We have one possible mode per unit volume in this space, so if we consider a spherical shell with a volume $\mathrm{d}^3\mathbf{n} = 4\pi n^2 \mathrm{d}n$, we have a total number of modes

$$\mathrm{d}N = N(n)\mathrm{d}n = \frac{1}{8} 4\pi n^2 \mathrm{d}n = \frac{L^3}{2\pi^2} k^2 \mathrm{d}k \tag{12.86}$$

where the factor of $1/8$ is because we only consider positive values of n_x, n_y, and n_z, so we only cover $1/8$th of the sphere. In the second step we replaced n^2 and $\mathrm{d}n$ with $k^2 L^2/\pi^2$ and $L\mathrm{d}k/\pi$ respectively.

Now we want to relate k to the frequency, which we can do using the classical wave equation

$$\nabla^2 \psi = \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \psi \quad \longrightarrow \quad k^2 = k_x^2 + k_y^2 + k_z^2 = \frac{\omega^2}{c^2} \quad \longrightarrow \quad k = \frac{2\pi\nu}{c} \tag{12.87}$$

Also notice that $L^3 = V$ is the volume of the box. Therefore, we can write $\mathrm{d}N$ as

$$\mathrm{d}N = \frac{V}{2\pi^2} \frac{8\pi^3 \nu^2}{c^3} \mathrm{d}\nu = \frac{4\pi\nu^2 V}{c^3} \mathrm{d}\nu \tag{12.88}$$

We need one more modification here: each wave can have 2 polarization states, so we multiply by a factor of 2. Then, the number of states per unit volume $\mathrm{d}\mathcal{N}$ is

$$\mathrm{d}\mathcal{N} = \frac{\mathrm{d}N}{V} = \frac{8\pi\nu^2}{c^3} \mathrm{d}\nu \tag{12.89}$$

And the energy density u is related to the number density by

$$\mathrm{d}u = u(\nu)\mathrm{d}\nu = \langle E \rangle \mathrm{d}\mathcal{N} \tag{12.90}$$

$$u(\nu) = \frac{8\pi\nu^2}{c^3} \langle E \rangle \tag{12.91}$$

Now, the average energy of a harmonic oscillator in thermal equilibrium is $\langle E \rangle = kT$, so we would expect the energy density to become

$$\boxed{u(\nu) = \frac{8\pi\nu^2}{c^3} kT} \tag{12.92}$$

This is the **Rayleigh-Jeans formula**. In general, it matches observed blackbody spectra well for low frequencies, but it becomes problematic at high frequencies where it diverges. This becomes even more of a problem when we consider the integral over all frequencies, which also diverges

$$\int_0^\infty \mathrm{d}\nu\, u(\nu) \to \infty \tag{12.93}$$

This seemingly implies that a blackbody has infinite energy. This problem was known as the *ultraviolet catastrophe*. The conceptual problem here is that, if an electromagnetic wave can have any arbitrary energy, then we should be able to increase the frequency ν without bound, meaning we decrease the

wavelength λ to infinitessimally short values. Then, if each of these infinitessimally spaced modes gets an equal amount of energy kT, obviously we are going to end up with an infinite amount of total energy as $\lambda \to 0$.

The solution to this problem is that the energy of the light is *quantized*—that is, it's split up into discrete units of energy called *photons*. This eliminates our infinite energy problem because the electromagnetic modes can no longer have any arbitrary frequency/wavelength. Each photon has an energy $h\nu = hc/\lambda$, and each mode has n photons, where n is an integer, corresponding to an energy $E_n = nh\nu$. If we then assume all of these modes are in thermal equilibrium, as we did before, then the probability of a mode having an energy E_n is given by Boltzmann statistics:

$$p_n = \frac{e^{-E_n/kT}}{\sum_{n=0}^{\infty} e^{-E_n/kT}} \tag{12.94}$$

And the average energy is now given by

$$\langle E \rangle = \sum_{n=0}^{\infty} E_n p_n = \frac{\sum_{n=0}^{\infty} nh\nu e^{-nh\nu/kT}}{\sum_{n=0}^{\infty} e^{-nh\nu/kT}} \tag{12.95}$$

We can define $x \equiv e^{-h\nu/kT}$ such that

$$\langle E \rangle = h\nu \frac{\sum_{n=0}^{\infty} nx^n}{\sum_{n=0}^{\infty} x^n} = h\nu x \frac{1 + 2x + 3x^2 + \ldots}{1 + x + x^2 + \ldots} \tag{12.96}$$

And we can use these known Taylor series expansions

$$\frac{1}{1-x} = 1 + x + x^2 + \ldots \tag{12.97}$$

$$\frac{1}{(1-x)^2} = 1 + 2x + 3x^2 + \ldots \tag{12.98}$$

To find

$$\langle E \rangle = \frac{h\nu x}{1-x} = \frac{h\nu}{x^{-1} - 1} = \frac{h\nu}{e^{h\nu/kT} - 1} \tag{12.99}$$

Then we just replace this in (12.91) to find

$$u(\nu) = \frac{8h\pi\nu^3}{c^3} \frac{1}{e^{h\nu/kT} - 1} \tag{12.100}$$

From here it's a simple matter to convert this to the intensity. Assuming the radiation is isotropic, we divide out the solid angle of 4π, and then we convert the energy density into a flux by multiplying by c: $B_\nu = cu(\nu)/4\pi$. This gives

$$\boxed{B_\nu(\nu, T) = \frac{2h\nu^3}{c^2} \frac{1}{e^{h\nu/kT} - 1}} \tag{12.101}$$

This curve now has a well defined total energy since $B_\nu \to 0$ for both limits as $\nu \to 0$ and $\nu \to \infty$.

See an example in Figure 12.7. We can find the total energy by integrating

$$B = \int_0^\infty \mathrm{d}\nu B_\nu(\nu, T) = \frac{2h}{c^2}\int_0^\infty \frac{\nu^3 \mathrm{d}\nu}{e^{h\nu/kT}-1} \tag{12.102}$$

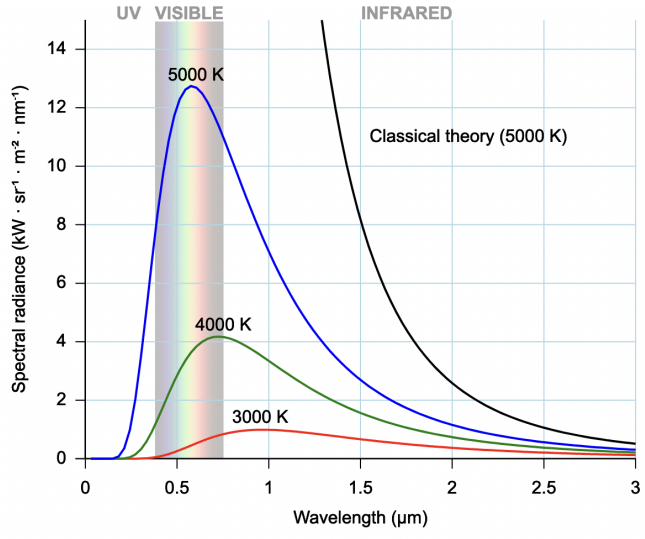


Figure 12.7: The Planck function for a series of different temperatures compared to the Rayleigh-Jeans function at 5000 K.

Doing a substitution of $x = h\nu/kT$ gives

$$B = \frac{2h}{c^2}\left(\frac{kT}{h}\right)^4 \int_0^\infty \frac{x^3 \mathrm{d}x}{e^x - 1} \tag{12.103}$$

where the integral gives $\pi^4/15$. Therefore

$$B = \frac{2\pi^4 k^4}{15h^3c^2}T^4 \tag{12.104}$$

If we say that we have a uniformly bright spherical blackbody, then the flux just at the surface is

$$F = \int \mathrm{d}\Omega \cos\theta B = \int_0^{2\pi} \mathrm{d}\phi \int_0^{\pi/2} \mathrm{d}\theta \sin\theta\cos\theta B = \pi B \tag{12.105}$$

Where we integrate θ only up to $\pi/2$ so that we're only counting radiation that is going outwards from the blackbody. Therefore, we can define the **Stefan-Boltzmann Constant**

$$\sigma_{\mathrm{SB}} \equiv \frac{2\pi^5 k^4}{15h^3c^2} = \frac{\pi^2 k^4}{60\hbar^3 c^2} \approx 5.6705 \times 10^{-5}\,\mathrm{erg\,s^{-1}\,cm^{-2}\,K^{-4}} \tag{12.106}$$

such that

$$\boxed{F = \sigma_{\mathrm{SB}} T^4} \tag{12.107}$$

Note: if we observe the flux further away from the surface, then the specific intensity is B only if a ray from our line of sight intersects the sphere and 0 otherwise, so we must only integrate up to a

polar angle $\theta_c = \arcsin(R/r)$ where R is the radius of the object and r is our distance away:

$$F = \int_0^{2\pi} \mathrm{d}\phi \int_0^{\theta_c} \mathrm{d}\theta \, \sin\theta \cos\theta B = \pi B \sin^2\theta_c = B\frac{\pi R^2}{r^2} \tag{12.108}$$

Now let's see what this gives us in the limit that $r \gg R$, which is certainly appropriate when we're observing, for example, stars in the Galaxy. Notice that the solid angle as a function of θ_c is given by

$$\int \mathrm{d}\Omega = \int_0^{2\pi} \mathrm{d}\phi \int_0^{\theta_c} \mathrm{d}\theta \, \sin\theta = 2\pi(1-\cos\theta_c) \tag{12.109}$$

In the limit $r \gg R$, which corresponds to $\theta_c \ll 1$, we have

$$\Omega \approx 2\pi\left(1 - 1 + \frac{\theta_c^2}{2}\right) = \pi\theta_c^2 \tag{12.110}$$

Where we used the Taylor series $\cos\theta \approx 1 - \theta^2/2$. We can take a similar limit in our expression for the flux using the first-order series for $\sin\theta \approx \theta$

$$F = \pi B \sin^2\theta_c \approx B\pi\theta_c^2 \tag{12.111}$$

Thus, when θ_c is small we can write

$$F \approx B\Omega \tag{12.112}$$

In agreement with what we found earlier in (12.69). Also note that for blackbody radiation, we can write the radiation pressure by integrating equation (12.57) over frequency, which gives

$$P = \frac{4\pi}{3c}B = \frac{4\sigma_{\mathrm{SB}}}{3c}T^4 \tag{12.113}$$

Defining the **radiation constant** as $a \equiv 4\sigma_{\mathrm{SB}}/c$, this becomes simply

$$\boxed{P = \frac{1}{3}aT^4} \tag{12.114}$$

One more important relationship to know related to blackbody radiation is **Wein's displacement law**, which gives the peak frequency/wavelength as a function of temperature. This must be solved for numerically, which gives:

$$\boxed{\lambda_{\mathrm{peak}} \approx \frac{2898\,\mu\mathrm{m\,K}}{T}} \quad , \quad \boxed{\nu_{\mathrm{peak}} \approx (5.879 \times 10^{10}\,\mathrm{Hz\,K^{-1}})\,T} \tag{12.115}$$

Also notice $\nu_{\mathrm{peak}} \neq c/\lambda_{\mathrm{peak}}$ since $B_\nu \neq B_\lambda$.

12.11.15 Kirchhoff's Law of Radiation

Say we have an enclosed box at a temperature T that is in thermal equilibrium. Then, anywhere inside the box we must have blackbody radiation $I_\nu = B_\nu(T)$, as proven in the last section. Now say that we have some small opening in the box, and we place some material at the opening with the

same temperature T (but the material may have different absorption/emission properties than the material within the box). By the first law of thermodynamics, the material is in thermal equilibrium with the box, so there can be no net exchange of energy through the opening. This means that any radiation passing through the opening also cannot change, since radiation carries energy. This requires

$$\frac{\mathrm{d}I_\nu}{\mathrm{d}\ell} = j_\nu - \alpha_\nu B_\nu(T) = 0 \tag{12.116}$$

Therefore, we obtain **Kirchhoff's Law of Radiation**

$$\boxed{j_\nu = \alpha_\nu B_\nu(T)} \tag{12.117}$$

Which means that our source function

$$\boxed{S_\nu = \frac{j_\nu}{\alpha_\nu} = B_\nu(T)} \tag{12.118}$$

We could have placed the material anywhere, so this relationship guarantees that the source function is the Planck function *everywhere* inside the box. This also guarantees that $I_\nu = B_\nu(T)$ everywhere inside the box, as we expected. However, this does *not* mean that we will always observe $I_\nu = B_\nu(T)$ from blackbody sources. In fact, this is only the case if the source is optically thick. Using equation (12.83) and assuming that $S_\nu = B_\nu(T)$ is approximately constant over the integral, we obtain

$$\boxed{I_\nu = I_\nu(0)e^{-\tau_\nu} + (1 - e^{-\tau_\nu})B_\nu(T)} \tag{12.119}$$

Consider a cloud of gas in thermal equilibrium at temperature T. If the cloud is optically thick ($\tau_\nu \gg 1$), then the intensity becomes

$$I_\nu \approx B_\nu(T) \quad \text{(optically thick)} \tag{12.120}$$

However, if the cloud is optically thin ($\tau_\nu \ll 1$), then we instead get

$$I_\nu \approx I_\nu(0)(1 - \tau_\nu) + \tau_\nu B_\nu(T) \quad \text{(optically thin)} \tag{12.121}$$

Confusingly, people often define dimensionless quantities called the **emissivity** ε_ν and **absorptivity** α_ν. Despite having the same name and same symbol, this emissivity is *not* the same thing as what was described in §12.11.11. It is simply the ratio of the observed intensity to the intensity that would be expected from a pure blackbody:

$$I_\nu = \varepsilon_\nu B_\nu(T) \tag{12.122}$$

Similarly, the absorptivity is *not* the same thing as the absorption coefficient. It is simply the ratio of the absorbed intensity to the incident intensity. In general we know that after being extincted, intensity goes as $e^{-\tau_\nu}$, so the absorptivity must be

$$\alpha_\nu = 1 - e^{-\tau_\nu} \tag{12.123}$$

On the walls of a blackbody, the incident intensity is $B_\nu(T)$ and the absorbed intensity is $\alpha_\nu B_\nu(T)$. For there to be no net energy transfer, the emitted intensity then must also be $\alpha_\nu B_\nu(T)$. But we just said above that the emitted intensity is $\varepsilon_\nu B_\nu(T)$. Therefore, in this language, Kirchhoff's Law is recast

as

$$\boxed{\varepsilon_\nu = \alpha_\nu = 1 - e^{-\tau_\nu}} \tag{12.124}$$

12.11.16 Magnitudes are scuffed

A magnitude is a logarithmic measure of brightness that dates all the way back to the Greek astronomer Hipparchus and for some reason is still in use today. It is defined such that a difference of 5 magnitudes corresponds to a factor of 100 in brightness, so a difference of 1 magnitude is a factor of $100^{1/5} \simeq 2.5$.

The **apparent magnitude** m of an object is analogous to the specific flux within some filter since it measures the brightness as it appears to us. Now, at this point we still haven't defined what the zero point of our magnitude scale is, so for now we'll work exclusively in relative quantities. If we have a reference object with magnitude m_{ref} and flux $F_{\nu,\text{ref}}$, then

$$\boxed{m - m_{\text{ref}} = -2.5 \log_{10}\left(\frac{F_\nu}{F_{\nu,\text{ref}}}\right)} \tag{12.125}$$

Notice that because of the negative sign, stupidly, a **brighter** object corresponds to a *smaller* magnitude. And yes, magnitudes *can* be negative if an object is bright enough. Makes perfect sense, right?

Similarly, the **absolute magnitude** M of an object is analogous to the specific luminosity within some filter, and is defined as the apparent magnitude that object would have *if it were a distance of 10 pc away* (10 pc is chosen arbitrarily). If we do a similar relative relationship with a reference object at M_{ref}, we have

$$\boxed{M - M_{\text{ref}} = -2.5 \log_{10}\left(\frac{L_\nu}{L_{\nu,\text{ref}}}\right)} \tag{12.126}$$

Since everything is defined to be measured as if it were at 10 pc, we effectively remove the effects of the distance and have a measure of the object's intrinsic brightness:

$$m - M = 2.5 \log_{10}\left(\frac{F_{\nu,10\,\text{pc}}}{F_\nu}\right) \quad \longrightarrow \quad \boxed{m - M = 5 \log_{10}\left(\frac{d}{10\,\text{pc}}\right) = 5 \log_{10} d - 5} \tag{12.127}$$

Therefore, the quantity $m - M$ is effectively a measurement of the distance to an object, so we call it the **distance modulus**.

Now, if we have an extinction of A magnitudes (see §5.Q83), this is modified by

$$m - M = 5 \log_{10} d - 5 + A \tag{12.128}$$

Notice that we still haven't defined a zero point for our magnitude scale. That's for good reason: there are many different magnitude scales in use with different zero points and conventions, and nobody ever specifies which magnitude scale they're using, so it's always a great fun time trying to figure out how to compare magnitudes between studies. Two of the most common scales are:

- **Vega magnitudes**: the zero point is defined to be the brightness of Vega. Note: the brightness of Vega is different in different filters (duh, it's a blackbody) AND it's also not constant in time (...oopsies teehee).
- **AB mangitudes**: the zero point in *every* filter is defined to be 3631 Jy.

  $m_{\rm AB} = -2.5\log_{10}(F_\nu/3631\,{\rm Jy}) = -2.5\log_{10}F_\nu - 48.60$

We also have **bolometric magnitudes**, which are magnitudes (either apparent or absolute) that are integrated over all frequencies/wavelengths, analogous to bolometric fluxes and luminosities:

$$m_{\rm bol} - m_{\rm bol,ref} = -2.5\log_{10}\left(\frac{F_{\rm bol}}{F_{\rm bol,ref}}\right) \tag{12.129}$$

$$M_{\rm bol} - M_{\rm bol,ref} = -2.5\log_{10}\left(\frac{L_{\rm bol}}{L_{\rm bol,ref}}\right) \tag{12.130}$$

Often we only have measurements of an object within a few select filters rather than across the whole spectrum, so astronomers have developed **bolometric correction** factors to convert these. For example, to convert from a magnitude in the V filter:

$$\boxed{BC \equiv m_{\rm bol} - m_V = M_{\rm bol} - M_V} \tag{12.131}$$

Another "correction" term often applied to magnitudes is the ***K*-correction**. This correction accounts for the fact that, if we observe a sample of objects at different redshifts in the same filter, the rest-frame wavelengths that we cover in each object will be different. Thus, depending on the SED shapes of the objects, we could get completely different values for the magnitudes than would be expected if we were measuring them all in the same wavelength range. Not only that, but the bandpass of the filter effectively shrinks by a factor of $1+z$. In general, K-corrections are *hard* to compute because you need to have some idea of the underlying SED shape of the source you're looking at.

Say the SED in the *rest* frame is given by $F(\nu)$ and the response function of your filter is $S(\nu)$. We observe at a frequency $\nu_{\rm obs}$ that corresponds to a rest-frame frequency of $\nu_{\rm emit} = \nu_{\rm obs}(1+z)$, so the observed flux will be the integral of $F(\nu_{\rm obs}(1+z))S(\nu_{\rm obs})$ over $\nu_{\rm obs}$. Notice the arguments for F and S are different here, which captures the effect of the redshift shifting the bandpass that we observe the source at. The flux that we *would've* observed if the source was not redshifted would be $F(\nu_{\rm obs})S(\nu_{\rm obs})$ integrated over $\nu_{\rm obs}$. Since we're integrating over all frequencies in both cases, we can just drop the subscripts on the ν's. Therefore, remembering the additional factor of $1+z$ to account for the difference in bandpass size, the K-correction is defined by

$$\boxed{K = -2.5\log_{10}\left[(1+z)\frac{\int_0^\infty F(\nu(1+z))S(\nu)\,{\rm d}\nu}{\int_0^\infty F(\nu)S(\nu)\,{\rm d}\nu}\right]} \tag{12.132}$$

such that we can convert our observed magnitude $m_{\rm obs}$ to a "rest-frame magnitude" $m_{\rm rest}$ by

$$\boxed{m_{\rm rest} = m_{\rm obs} - K} \tag{12.133}$$

See Refs. 274,275.

12.12 Cosmological Distance Measures

There are a lot of different measurements of "distance" in cosmology, so this section is dedicated to going over the basics of what we mean when we say "distance" and how all of these different distance measures are related to each other. See Ref. 276.

12.12.1 Expansion of the Universe and Coordinates

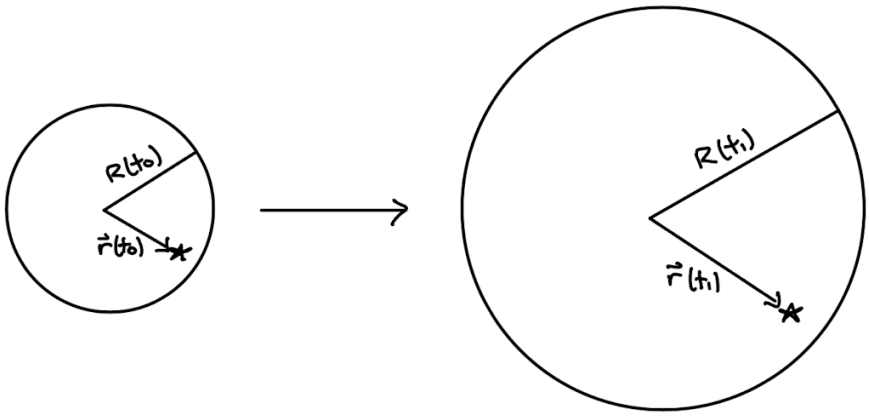


Figure 12.8: The measurement of distance to an object in an expanding universe.

Let's imagine the universe as an expanding sphere as in Figure 12.8 with a radius $R(t)$ as a function of time. Then we can define the dimensionless **scale factor** as the ratio of the size of the universe at some time t relative to some reference time t_0:

$$a(t) \equiv \frac{R(t)}{R(t_0)} \tag{12.134}$$

We can see from here that $a(t_0) = 1$. Now, say we observe an object at a distance of $\boldsymbol{\chi} \equiv \mathbf{r}(t_0)$ at time t_0. Then at some other time (assuming the object itself is not moving) the distance will just be

$$\mathbf{r}(t) = a(t)\boldsymbol{\chi} \tag{12.135}$$

Thus, $\boldsymbol{\chi}$ is a scale-independent (and thus time-independent) measure of the distance, which we call the **comoving distance**. Now we can derive the recessional velocity due to the expansion by just taking the derivative of (12.135):

$$\dot{\mathbf{r}}(t) = \dot{a}\boldsymbol{\chi} = \frac{\dot{a}}{a}\mathbf{r}(t) \tag{12.136}$$

Now we can use this to generalize the Hubble law ($v = H_0 d$) over all of cosmic time. We just replace $v \to |\dot{\mathbf{r}}|$ and $d \to |\mathbf{r}|$. Then the Hubble "constant" is just

$$H(t) = \frac{|\dot{\mathbf{r}}|}{|\mathbf{r}|} = \frac{\dot{a}}{a} \tag{12.137}$$

12.12.2 The FLRW Metric

To find a metric that describes the expanding and potentially curved spacetime of the universe accurately, let's build up one step at a time, starting with the basic metric in flat polar coordinates:

$$ds^2 = d\rho^2 + \rho^2 d\phi^2 \tag{12.138}$$

Where ρ and ϕ are the usual radial and azimuthal angle coordinates.

How does this change if our "2D" distance is actually on a curved surface like a sphere? Let's call our spherical coordinates (R, θ, ϕ), where the radial coordinate R is always the full radius of the sphere since we are considering a 2D surface. Then, a metric on the surface of this sphere should look like

$$ds^2 = R^2 d\theta^2 + \rho^2 d\phi^2 \tag{12.139}$$

See Figure 12.9. The relationship between the cylindrical radial coordinate ρ and the spherical radial coordinate R is

$$\rho = R\sin\theta \quad \longrightarrow \quad d\rho = R\cos\theta\, d\theta \tag{12.140}$$

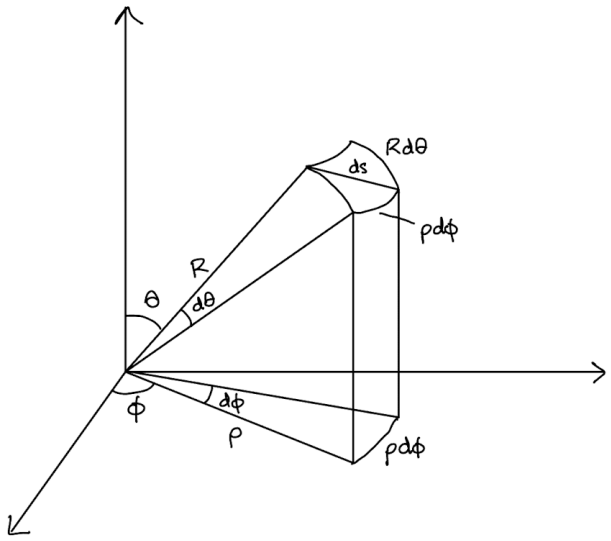


Figure 12.9: The differential distance on a 2D spherical surface.

And

$$\cos\theta = \frac{\sqrt{R^2-\rho^2}}{R} \quad \longrightarrow \quad R d\theta = \frac{d\rho}{\cos\theta} = \frac{R d\rho}{\sqrt{R^2-\rho^2}} = \frac{d\rho}{\sqrt{1-\rho^2/R^2}} \tag{12.141}$$

Therefore, if we want to express this metric in terms of the flat radial coordinate ρ, we have the following, defining the curvature $K \equiv 1/R^2$

$$ds^2 = \frac{d\rho^2}{1-K\rho^2} + \rho^2 d\phi^2 \tag{12.142}$$

Now all we have to do is generalize this to a 3D curved space. This is hard to visualize since we have

some curvature in addition to the 3 spatial dimension we already have. Nonetheless, it's easy to just write out what the metric should be if we don't think about it too hard:

$$\mathrm{d}s^2 = \frac{\mathrm{d}r^2}{1-Kr^2} + r^2\mathrm{d}\theta^2 + r^2\sin^2\theta\mathrm{d}\phi^2 \tag{12.143}$$

And voilà! The hard part is over. Now we just have to generalize to 4D spacetime with the Mike Wazowski metric, which we can do by simply adding a negative time coordinate $-c^2\mathrm{d}t^2$, and the expansion of the universe. We can also simplify the notation by defining $\mathrm{d}\Omega^2 = \mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2$.

Now we'll replace our radial coordinate with $r \to a(t)r$ such that r is now a comoving coordinate. We'll also define the scale-independent curvature k such that $K(t) \equiv k/a^2(t)$. This allows us to write the metric as

$$\boxed{\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2\mathrm{d}\Omega^2\right]} \tag{12.144}$$

This is the **Friedmann-Lemaître-Robertson-Walker (FLRW) Metric**. In this formulation, k has units of $1/\text{length}^2$ and a is dimensionless. Sometimes it is alternatively written with the radius of curvature $R(t) = a(t)R(t_0)$ (from above), dimensionless distance $\bar{r} = r/R(t_0)$, and dimensionless curvature $k' = kR(t_0)^2$. This has the advantage of being able to absorb the normalization of the curvature into $R(t)$, so k' only takes values in the set $\{-1, 0, 1\}$, representing an open, flat, and closed universe, respectively.

$$\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + R^2(t)\left[\frac{\mathrm{d}\bar{r}^2}{1-k'\bar{r}^2} + \bar{r}^2\mathrm{d}\Omega^2\right] \tag{12.145}$$

However, most of the time you will see it written following equation (12.144). We've built this metric up from simple cases and adding complexity using mostly heuristic arguments, but it can be constructed more formally using the Riemann curvature tensor in a manifold that is maximally symmetric.

12.12.3 Redshift and the Scale Factor

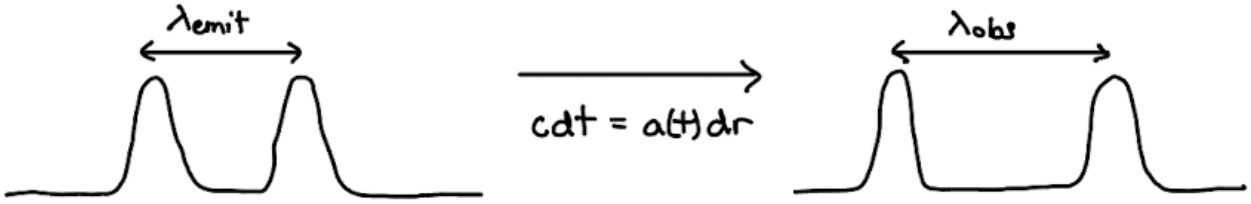


Figure 12.10: Diagram showing the redshifting of photons due to the expansion of the universe.

Redshift is defined as

$$z = \frac{\lambda_{\text{obs}} - \lambda_{\text{emit}}}{\lambda_{\text{emit}}} = \frac{\lambda_{\text{obs}}}{\lambda_{\text{emit}}} - 1 \tag{12.146}$$

We can find the relationship between the cosmic scale factor a and the redshift z by considering two pulses of light separated by a wavelength λ_{emit}. They travel a distance $c\mathrm{d}t = a(t)\mathrm{d}r$, and at the end of their journey they are now separated by a distance λ_{obs}. See this depicted in Figure 12.10. We can integrate to find the **comoving** distance r that the first pulse travels:

$$\int_{r_{\text{emit}}}^{r_{\text{obs}}} \mathrm{d}r = \int_{t_{\text{emit}}}^{t_{\text{obs}}} \frac{c\mathrm{d}t}{a(t)} \tag{12.147}$$

The second pulse travels the same distance but starts and arrives a bit later, so

$$\int_{t_{\text{emit}}+\delta t_{\text{emit}}}^{t_{\text{obs}}+\delta t_{\text{obs}}} \frac{c\mathrm{d}t}{a(t)} = \int_{t_{\text{emit}}}^{t_{\text{obs}}} \frac{c\mathrm{d}t}{a(t)} + \int_{t_{\text{obs}}}^{t_{\text{obs}}+\delta t_{\text{obs}}} \frac{c\mathrm{d}t}{a(t)} - \int_{t_{\text{emit}}}^{t_{\text{emit}}+\delta t_{\text{emit}}} \frac{c\mathrm{d}t}{a(t)} \tag{12.148}$$

Then we set these two distances equal and approximate the small integrals by assuming $a(t)$ does not change over the short integration time

$$\int_{t_{\text{emit}}}^{t_{\text{obs}}} \frac{c\mathrm{d}t}{a(t)} \simeq \int_{t_{\text{emit}}}^{t_{\text{obs}}} \frac{c\mathrm{d}t}{a(t)} + \frac{c}{a_{\text{obs}}}\delta t_{\text{obs}} - \frac{c}{a_{\text{emit}}}\delta t_{\text{emit}} \tag{12.149}$$

The left two terms cancel with each other, leaving only the differential terms. Notice that we have essentially said the quantity $c\delta t/a$, which is like a comoving distance between the wave crests, should be constant from the time of emission and the time of observation. Then, with $\delta t = 1/\nu = \lambda/c$, we have

$$\frac{\lambda_{\text{obs}}}{a_{\text{obs}}} = \frac{\lambda_{\text{emit}}}{a_{\text{emit}}} \quad \longrightarrow \quad \frac{a_{\text{obs}}}{a_{\text{emit}}} = \frac{\lambda_{\text{obs}}}{\lambda_{\text{emit}}} \tag{12.150}$$

Now we set $a_{\text{obs}} = 1$ for today, and we use the definition of redshift to obtain

$$\boxed{\frac{1}{a} = 1 + z} \tag{12.151}$$

12.12.4 Hubble Distance

This is just a constant, defined as

$$\boxed{d_H \equiv c/H_0 \sim 4300 h_{70}^{-1}\,\text{Mpc}} \tag{12.152}$$

12.12.5 Comoving Distance

We can already see our first important distance measure: the **comoving distance** $d_C \equiv \boldsymbol{\chi}$. As mentioned, this represents the actual physical distance between two objects that would be measured by a ruler **at the present epoch** (t_0). To find an expression for the comoving distance in terms of observable parameters, let's consider a photon traveling from some distant object towards us along a radial line of sight. Then, using the FLRW metric (12.144) in the case where $\mathrm{d}s^2 = 0$ (light-like), and

choosing $\mathrm{d}\theta^2 = \mathrm{d}\phi^2 = 0$ (radial path), we have

$$\chi = \int_0^r \frac{\mathrm{d}r'}{\sqrt{1-kr'^2}} = \int_{t_e}^{t_0} \frac{c\mathrm{d}t'}{a(t')} \tag{12.153}$$

The **comoving coordinate** r here must be distinguished from the **comoving distance** χ—r here is just the radial comoving coordinate, but χ is the actual comoving distance that would be measured by a ruler. In a spatially flat universe, these are the same ($k = 0$), but they differ in curved universes. To see why, think of an ant on the surface of a sphere (Figure 12.11). It walks down some distance D from the top and traces out a circle. The circle has a circumference $2\pi\rho$ where $\rho = R\sin\theta$ is the perpendicular distance from the vertical axis, but the ant expects it to have a circumference of $2\pi D$ since that's how far it walked from the center. The difference is due to the curved nature of the surface. The "coordinate distance" here is $r = \rho$, but the actual distance from the center is $\chi = D$. We see that $\chi > r$ in this case, since we have positive curvature. Likewise, $\chi < r$ if we have negative curvature.

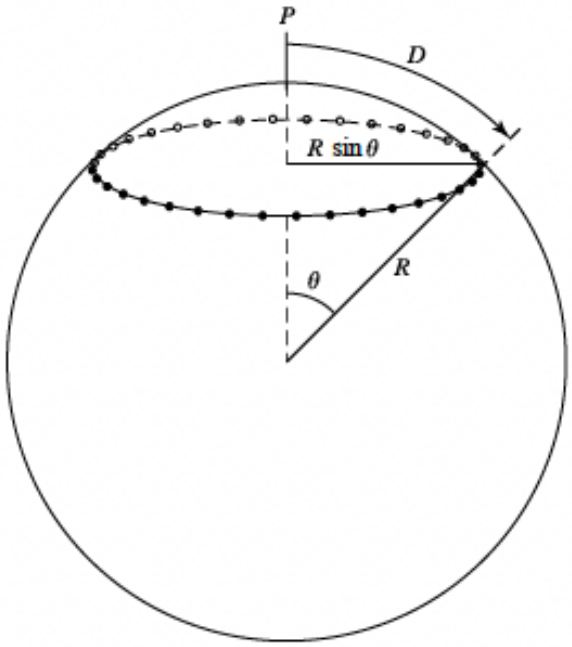


Figure 12.11: Demonstration of the effect of curvature of measured coordinates[11].

We can get the full relationship between r and χ by integrating the LHS of (12.153). We see

$$\chi = \begin{cases} |k|^{-1/2}\mathrm{arcsinh}(r\sqrt{|k|}) & k < 0 \quad \text{open universe} \\ r & k = 0 \quad \text{flat universe} \\ |k|^{-1/2}\arcsin\left(r\sqrt{|k|}\right) & k > 0 \quad \text{closed universe} \end{cases} \tag{12.154}$$

Note, we can also write the integral in terms of the redshift by noticing that

$$\frac{\mathrm{d}t}{a}\frac{\dot{a}}{\dot{a}} = \frac{\mathrm{d}a}{Ha^2} = -\frac{\mathrm{d}z}{(1+z)^2}\frac{1}{Ha^2} = -\frac{\mathrm{d}z}{H} \tag{12.155}$$

$$\therefore \quad \chi = \int_0^z \frac{c\mathrm{d}z'}{H(z')} = \frac{c}{H_0}\int_0^z \frac{\mathrm{d}z'}{E(z')} \tag{12.156}$$

In summary, the **comoving distance** is

$$\boxed{d_C(z) = \int_{t_e}^{t_0} \frac{c\mathrm{d}t'}{a(t')} = d_H \int_0^z \frac{\mathrm{d}z'}{E(z')}} \tag{12.157}$$

The interpretation here is that for each small timestep $\mathrm{d}t$, the photon travels a distance $c\mathrm{d}t$. By dividing this by the scale factor at the time $a(t)$, we convert this into the distance it **would be** today.

Note: sometimes people write the FLRW metric in terms of the comoving distance χ rather than the coordinate distance r. In this case, it takes the form

$$\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + a^2(t)\Big[\mathrm{d}\chi^2 + S^2(\chi)\mathrm{d}\Omega^2\Big] \tag{12.158}$$

where $S(\chi)$ is the inverse of equation (12.154).

12.12.6 Proper Distance

Next up, we have the **proper distance**. As the name suggests, this is literally just the actual distance between two objects that would be measured by a ruler at any point in cosmic time. Therefore, at the present epoch, this is the same thing as the comoving distance, but at different epochs it will differ by the scale factor, i.e.

$$\boxed{d_P(z) = a(t)d_C(z) = \frac{1}{1+z}d_C(z)} \tag{12.159}$$

12.12.7 Transverse Comoving Distance / Proper Motion Distance

Let's look back at the comoving coordinate r. We can try to find this as a function of z by reversing the relationships we derived for the comoving distance d_C. Particularly, the reverse of equation (12.154), substituting $\Omega_K = -kc^2/H_0^2 = -kd_H^2$, gives

$$\boxed{d_M(z) = r = \begin{cases} \dfrac{d_H}{\sqrt{|\Omega_K|}}\sinh\Big(\sqrt{|\Omega_K|}d_C(z)/d_H\Big) & \Omega_K > 0 \\ d_C(z) & \Omega_K = 0 \\ \dfrac{d_H}{\sqrt{|\Omega_K|}}\sin\Big(\sqrt{|\Omega_K|}d_C(z)/d_H\Big) & \Omega_K < 0 \end{cases}} \tag{12.160}$$

As discussed above, this is the **comoving coordinate**, which is NOT the same as the **comoving distance**. However, this quantity is still useful to us because it can help us determine perpendicular or transverse distances. For this reason, it is often given its own special name: the **transverse comoving distance** or the **proper motion distance**, with the symbol $d_M(z)$.

To see why we need to use this instead of the comoving distance, look back at the FLRW metric again. This time, consider that we have 2 objects at the same redshift z separated by an angle θ. To get the proper distance *between* these two objects at the current time t_0, we can set $\mathrm{d}t = \mathrm{d}r = \mathrm{d}\phi = 0$, and with $a(t_0) = 1$ we find

$$\mathrm{d}s = r\mathrm{d}\theta \quad \longrightarrow \quad r = d_M(z) = \mathrm{d}s/\mathrm{d}\theta \tag{12.161}$$

This shows that the **transverse comoving distance** is the correct distance to use when we want to convert angular separations to physical separations (at the current epoch). The reason this is also sometimes called the **proper motion distance** is because this can be equivalently thought of as the ratio of the physical transverse velocity to the angular velocity:

$$d_M(z) = \dot{s}/\dot{\theta} \tag{12.162}$$

since the $\mathrm{d}t$'s cancel.

12.12.8 Angular Diameter Distance

If we want to know the physical separation between two objects at the same redshift z *at the epoch given by* z rather than the current epoch, then we need to correct the transverse comoving distance by the scale factor. This is called the **angular diameter distance**:

$$\boxed{d_A(z) = \frac{D}{\theta} = a(t)d_M(z) = \frac{1}{1+z}d_M(z)} \tag{12.163}$$

This distance is used often to determine the physical sizes of objects (D) based on their angular sizes on the sky (θ), since for a gravitationally bound object like a galaxy, the physical size will not change with the expansion of the universe. Note that for a flat universe, this is equal to the **proper distance**.

Also noteworthy here is that this gives a relationship between the apparent angular size of an object θ as a function of redshift (since D is fixed): $\theta(z) = D/d_A(z)$. The angular diameter distance actually has a turnover, meaning that the angular sizes of objects paradoxically start *increasing* with redshift after this turnaround point (at $z \sim 1$).

12.12.9 Luminosity Distance

The observed luminosity of an object will be affected by the expansion of the universe. Photons emitted will be redshifted, therefore decreasing their energy $h\nu$ by a factor of $1+z$. But in addition, photons will arrive to us less frequently since the expansion increases the timescale δt *between* receiving each photon by another factor of $1+z$. Therefore

$$L_{\mathrm{obs}} \propto \frac{h\nu_{\mathrm{obs}}}{\delta t_{\mathrm{obs}}} = \frac{h\nu_{\mathrm{emit}}}{1+z}\frac{1}{\delta t_{\mathrm{emit}}(1+z)} \propto \frac{L_{\mathrm{emit}}}{(1+z)^2} \tag{12.164}$$

Then this luminosity will also be attenuated by the usual $1/r^2$ law, but what distance do we use here to find the attenuation? The answer is simple if we consider a source that is emitting light at the origin of our coordinate system (so $r = 0$). Using the FLRW metric, we see that at the current time ($a = 1$), the surface area of the sphere that the photons radiate outwards into at a distance r is just $4\pi r^2$ (since the $\mathrm{d}\theta$ and $\mathrm{d}\phi$ components of the metric have no dependence on the curvature), so we

should use the coordinate distance r, AKA the transverse comoving distance d_M, in the calculation of the flux. Note that despite the area being $4\pi r^2$, the proper distance from the origin to the surface of the sphere is NOT r, it is still d_P. This is analogous to the 2D example with the ant on the surface of the sphere where the little guy expects his circle's circumference to be $2\pi D$ when in actuality it's $2\pi\rho$. With this in mind, we have

$$F_{\text{obs}} = \frac{L_{\text{obs}}}{4\pi d_M^2} = \frac{L_{\text{emit}}}{4\pi(1+z)^2 d_M^2} = \frac{L_{\text{emit}}}{4\pi(1+z)^4 d_A^2} \tag{12.165}$$

If we want to relate the observed flux to the emitted luminosity L_{emit}, we define the **luminosity distance** such that

$$F_{\text{obs}} = \frac{L_{\text{emit}}}{4\pi d_L^2} \tag{12.166}$$

Comparing these two equations, we see that

$$\boxed{d_L(z) = (1+z) d_M(z) = (1+z)^2 d_A(z)} \tag{12.167}$$

Notice that if we have a surface brightness (intensity):

$$I_{\text{obs}} = \frac{F_{\text{obs}}}{\Omega} \propto \frac{1/d_L^2}{1/d_A^2} \propto \frac{d_A^2}{(1+z)^4 d_A^2} \longrightarrow \boxed{I_{\text{obs}} \propto \frac{1}{(1+z)^4}} \tag{12.168}$$

It gets dimmed by a factor of $(1+z)^4$.

12.12.10 Light-travel Distance

What if we want to find the total time that it takes a photon to reach us from a distant object? We can rewrite $\mathrm{d}t$ in terms of the redshift

$$\mathrm{d}t = \frac{\mathrm{d}a}{\dot{a}} = -\frac{\mathrm{d}z}{(1+z)^2}\frac{1}{aH} = \frac{1}{H}\frac{\mathrm{d}z}{1+z} \tag{12.169}$$

And from here it is a matter of integrating to find the total time. We can then multiply this by the speed of light to find the **light-travel distance**

$$\boxed{d_L(z) = d_H \int_0^z \frac{\mathrm{d}z'}{(1+z')E(z')}} \tag{12.170}$$

As the name suggests, this gives the total distance traveled by a photon on its journey from the source to us. This is different from the proper distance because when the photon is emitted, the proper distance between us and the source is shorter than the present-day value (the universe has expanded during the photon's travel time). This effect is shown in Figure 12.12. Dividing by c here gives the **lookback time**, which tells us how far back in time we observe the source at.

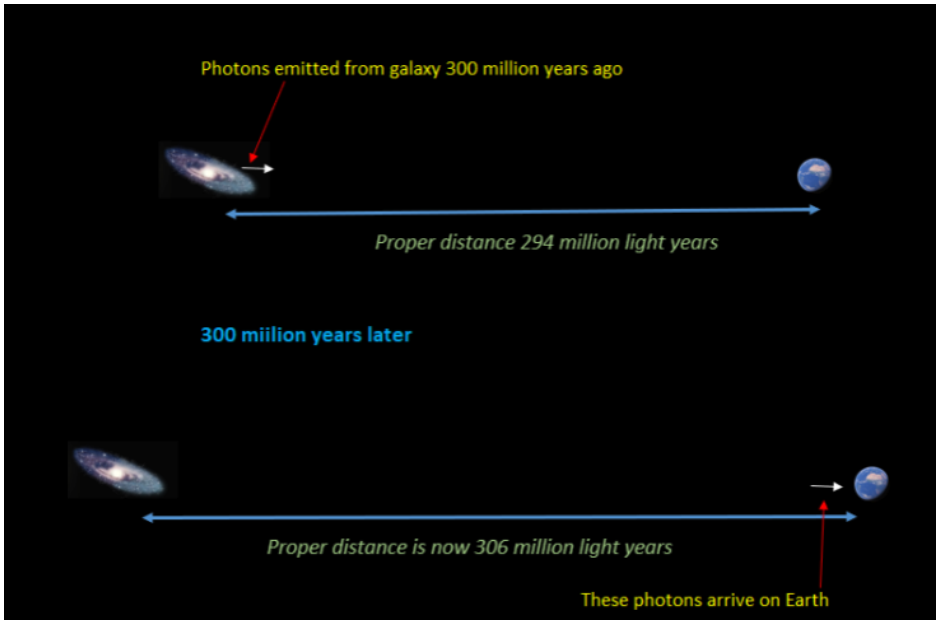


Figure 12.12: Depiction of the light travel distance compared to the proper distance when the photon is emitted and the proper distance today (i.e. the comoving distance)[277].

12.12.11 Horizon Distance

We call the proper distance to the farthest observable point the **horizon distance**. Any two points separated by a larger distance are thus not in causal contact with each other. As such, this may be thought of as the radius of the observable universe as a function of time. We can get it by just taking the integral for the proper distance and extending the limits to reach $t = 0$, corresponding to $z = \infty$:

$$\boxed{d_h(z) = a(t)\int_0^t \frac{c\mathrm{d}t'}{a(t')} = \frac{d_H}{1+z}\int_z^\infty \frac{\mathrm{d}z'}{E(z')}} \tag{12.171}$$

We can approximate this for different eras of the history of the universe. In the radiation-dominated era, we can plug in $a \propto t^{1/2}$ from (8.142)

$$d_h(t) \simeq t^{1/2}\int_0^t \frac{c\mathrm{d}t'}{t'^{1/2}} \longrightarrow \boxed{d_h(t) \simeq 2ct \simeq \frac{c}{H(z)}} \tag{12.172}$$

In the matter-dominated era, we have $a \propto t^{2/3}$ from (8.145)

$$d_h(t) \simeq t^{2/3}\int_0^t \frac{c\mathrm{d}t'}{t'^{2/3}} \longrightarrow \boxed{d_h(t) \simeq 3ct \simeq \frac{2c}{H(z)}} \tag{12.173}$$

And finally, in the Λ-dominated era, $a \propto \exp(t)$ from (8.151)

$$d_h(t) \simeq \exp\left(H_0\sqrt{\Omega_\Lambda}t\right)\int_0^t \frac{c\mathrm{d}t'}{\exp\left(H_0\sqrt{\Omega_\Lambda}t'\right)} \longrightarrow \boxed{d_h(t) \simeq \frac{c}{H_0\sqrt{\Omega_\Lambda}}\left(\exp\left(H_0\sqrt{\Omega_\Lambda}t\right)-1\right)} \tag{12.174}$$

13 Acknowledgements

A big thank-you to the previous generations of MIT grad students whose notes I used extensively when making these notes! Particularly:

- Megan Masterson
- Cian Roche (AstroWiki)
- Meredith Neyer (AstroWiki)
- Stephanie O'Neil
- Xiaowei Ou
- Geoffrey Mo
- Rahul Jayaraman

Shout out to the doggos Bodhi and Maple for providing emotional support.

References

1. Lissauer, J. J. & de Pater, I. *Fundamental Planetary Science: Physics, Chemistry, and Habitability* (Cambridge University Press, Cambridge, UK, 2019), updated edn.

2. Taylor, J. R. *Classical Mechanics* (University Science Books, Melville, NY, 2005).

3. Weingartner, J. C. ASTR 403 Lecture Notes (2022). URL `http://physics.gmu.edu/~joe/ASTR403/Course-materials/Course-materials.html`.

4. Gomes, R., Levison, H. F., Tsiganis, K. & Morbidelli, A. Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. *Nature* **435**, 466–469 (2005).

5. Tsiganis, K., Gomes, R., Morbidelli, A. & Levison, H. F. Origin of the orbital architecture of the giant planets of the Solar System. *Nature* **435**, 459–461 (2005).

6. Morbidelli, A., Levison, H. F., Tsiganis, K. & Gomes, R. Chaotic capture of Jupiter's Trojan asteroids in the early Solar System. *Nature* **435**, 462–465 (2005).

7. Eratosthenes and the Circumference of the Earth. *Nature* **152**, 473 (1943).

8. Berggren, J. L. & Sidoli, N. Aristarchus's on the sizes and distances of the sun and the moon: Greek and arabic texts. *Archive for History of Exact Sciences* **61**, 213–254 (2007).

9. Winn, J. N. Exoplanet Transits and Occultations. In Seager, S. (ed.) *Exoplanets*, 55–77 (2010).

10. Seager, S. (ed.) *Exoplanets* (University of Arizona Press, 2010).

11. Carroll, B. W. & Ostlie, D. A. *An Introduction to Modern Astrophysics* (Cambridge University Press, 2017), 2nd edn.

12. Kalas, P. *et al.* Optical images of an exosolar planet 25 light-years from earth. *Science* **322**, 1345–1348 (2008). URL `https://www.science.org/doi/abs/10.1126/science.1166609`. `https://www.science.org/doi/pdf/10.1126/science.1166609`.

13. Ségransan, D. *et al.* The HARPS search for southern extra-solar planets. XXIX. Four new planets in orbit around the moderatly active dwarfs HD 63765, HD 104067, HD 125595, and HIP 70849. *A&A* **535**, A54 (2011). `1107.0339`.

14. NASA/IPAC. NASA Exoplanet Archive (2024). URL `https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html`.

15. Wolszczan, A. & Frail, D. A. A planetary system around the millisecond pulsar psr1257 + 12. *Nature* **355**, 145–147 (1992). URL `https://doi.org/10.1038/355145a0`.

16. Ananyeva, V. I. *et al.* The mass and orbital-period distributions of exoplanets accounting for the observational selection of the method for measuring radial velocities. a dominant (averaged) structure of planetary systems. *Astronomy Reports* **66**, 886–917 (2022). URL `https://doi.org/10.1134/S106377292210002X`.

17. Wolszczan, A. & Frail, D. A. A planetary system around the millisecond pulsar PSR1257 + 12. *Nature* **355**, 145–147 (1992).

18. Mayor, M. & Queloz, D. A Jupiter-mass companion to a solar-type star. *Nature* **378**, 355–359 (1995).

19. Chen, J. & Kipping, D. Probabilistic forecasting of the masses and radii of other worlds. *The Astrophysical Journal* **834**, 17 (2016). URL http://dx.doi.org/10.3847/1538-4357/834/1/17.

20. Stellar classification. URL https://en.wikipedia.org/wiki/Stellar_classification#:~:text=Most%20stars%20are%20currently%20classified,the%20coolest%20(M%20type).

21. Hertzsprung-russell diagram. URL https://en.wikipedia.org/wiki/Hertzsprung%E2%80%93Russell_diagram.

22. Vanderburg, A. 8.901 Lecture Notes (2023).

23. Baade, W. The Resolution of Messier 32, NGC 205, and the Central Region of the Andromeda Nebula. *ApJ* **100**, 137 (1944).

24. Choudhuri, A. R. *Astrophysics for Physicists* (Cambridge University Press, 2010).

25. Kline, D. 11.2: Equivalent one-body representation for two-body motion. URL https://phys.libretexts.org/Bookshelves/Classical_Mechanics/Variational_Principles_in_Classical_Mechanics_(Cline)/11%3A_Conservative_two-body_Central_Forces/11.02%3A_Equivalent_one-body_Representation_for_two-body_motion.

26. Oswalt, T. D. & Gilmore, G. *Planets, Stars and Stellar Systems: Volume 5: Galactic Structure and Stellar Populations* (2013). URL https://link.springer.com/content/pdf/10.1007/978-94-007-5612-0_17.pdf.

27. Stellar evolution: Cesar's booklet. URL https://cesar.esa.int/upload/201801/stellarevolution_booklet_v2.pdf.

28. Stellar evolution. URL https://www.aavso.org/stellar-evolution.

29. Chakrabarty, D. 8.901 lecture notes (2021).

30. Nave, R. Hyperphysics: Supernovae. URL http://hyperphysics.phy-astr.gsu.edu/hbase/Astro/snovcn.html#:~:text=Supernovae%20are%20classified%20as%20Type%20I%20if%20their%20light%20curves,sharply%20than%20the%20Type%20I.

31. Knapp, J. Radial stellar pulsations (2011). URL https://www.astro.princeton.edu/~gk/A403/pulse.pdf.

32. Pulsating stars: Stars that breathe. URL https://astronomy.swin.edu.au/sao/downloads/HET611-M17A01.pdf.

33. Riess, A. G. *et al.* VizieR Online Data Catalog: Photometry of variables in NGC 3370 (Riess+, 2005). VizieR On-line Data Catalog: J/ApJ/627/579. Originally published in: 2005ApJ...627..579R (2005).

34. Maciel, W. J., Costa, R. D. D. & Idiart, T. E. P. Planetary nebulae and the chemical evolution of the Magellanic Clouds. *Revista Mexicana de Astronomía y Astrofísica* **45**, 127–137 (2009). 0904.2549.

35. Knapp, J. (2011). URL https://www.astro.princeton.edu/~gk/A403/stellar.pdf.

36. Goldberg, D. *The Standard Model in a Nutshell* (Princeton University Press, 2017).

37. Williams, M. 8.701 lecture notes (2022).

38. Borexino Collaboration, M., Agostini *et al.* Experimental evidence of neutrinos produced in the CNO fusion cycle in the Sun. *Nature* **587**, 577–582 (2020). 2006.15115.

39. Weingartner, J. C. ASTR 328 Lecture Notes (2021). URL http://physics.gmu.edu/~joe/ASTR328/Course-materials/Course-materials.html.

40. Kippenhahn, R., Weigert, A. & Weiss, A. *Stellar Structure and Evolution* (Springer), 2nd edn.

41. Nuclear binding energy. URL https://phys.libretexts.org/Bookshelves/University_Physics/University_Physics_%28OpenStax%29/University_Physics_III_-_Optics_and_Modern_Physics_%28OpenStax%29/10%3A__Nuclear_Physics/10.03%3A_Nuclear_Binding_Energy.

42. Adelberger, E. G. *et al.* Solar fusion cross sections. ii. the<mml:math xmlns:mml="http://www.w3.org/1998/math/mathml" display="inline"><mml:mi>p</mml:mi><mml:mi>p</r and cno cycles. *Reviews of Modern Physics* **83**, 195–245 (2011). URL http://dx.doi.org/10.1103/RevModPhys.83.195.

43. Astronomy (2017). URL https://pressbooks.online.ucf.edu/astronomybc/chapter/16-3-the-solar-interior-theory/.

44. The opal opacity code (2001). URL https://opalopacity.llnl.gov/.

45. Iglesias, C. A. & Rogers, F. J. Updated Opal Opacities. *ApJ* **464**, 943 (1996).

46. Fabian, A. C., Iwasawa, K., Reynolds, C. S. & Young, A. J. Broad Iron Lines in Active Galactic Nuclei. *PASP* **112**, 1145–1161 (2000). astro-ph/0004366.

47. Reynolds, C. S. Observational Constraints on Black Hole Spin. *ARA&A* **59**, 117–154 (2021). 2011.08948.

48. Peterson, B. M. & Horne, K. Echo mapping of active galactic nuclei. *Astronomische Nachrichten* **325**, 248–251 (2004). URL http://dx.doi.org/10.1002/asna.200310207.

49. Uttley, P., Cackett, E. M., Fabian, A. C., Kara, E. & Wilkins, D. R. X-ray reverberation around accreting black holes. *The Astronomy and Astrophysics Review* **22** (2014). URL http://dx.doi.org/10.1007/s00159-014-0072-0.

50. Event Horizon Telescope Collaboration *et al.* First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole. *ApJL* **875**, L1 (2019). 1906.11238.

51. Tolman, R. C. Static Solutions of Einstein's Field Equations for Spheres of Fluid. *Physical Review* **55**, 364–373 (1939).

52. Oppenheimer, J. R. & Volkoff, G. M. On Massive Neutron Cores. *Physical Review* **55**, 374–381 (1939).

53. Miller, M. C. *et al.* The Radius of PSR J0740+6620 from NICER and XMM-Newton Data. *ApJL* **918**, L28 (2021). 2105.06979.

54. Miller, C. Overview of gravitational radiation (2008). URL https://www.astro.umd.edu/~miller/teaching/astr498/lecture24.pdf.

55. Wheeler. Gravitational waves (2013). URL https://www.physics.usu.edu/Wheeler/GenRel2013/Notes/GravitationalWaves.pdf.

56. Sticky bead argument. URL https://en.wikipedia.org/wiki/Sticky_bead_argument.

57. Astrophysical sensitivity (2023). URL https://cosmicexplorer.org/sensitivity.html.

58. Agazie, G. *et al.* The NANOGrav 15 yr Data Set: Evidence for a Gravitational-wave Background. *ApJL* **951**, L8 (2023). 2306.16213.

59. Sanders, R. After 15 years, pulsar timing yields evidence of cosmic gravitational wave background (2023). URL https://news.berkeley.edu/2023/06/28/after-15-years-pulsar-timing-yields-evidence-of-cosmic-gravitational-wave-background#:~:text=After%2015%20years%2C%20pulsar%20timing,cosmic%20gravitational%20wave%20background%20%7C%20Berkeley.

60. Crossfield, I. J. M. Accretion (2019). URL https://crossfield.ku.edu/8901_2019A/lec027.pdf.

61. Roche limit. URL https://en.wikipedia.org/wiki/Roche_limit#:~:text=In%20celestial%20mechanics%2C%20the%20Roche,the%20second%20body's%20self%2Dgravitation.

62. Shakura, N. I. & Sunyaev, R. A. Black holes in binary systems. Observational appearance. *A&A* **24**, 337–355 (1973).

63. Hirata, C. Interstellar shocks (2011). URL http://www.tapir.caltech.edu/~chirata/ay102/Shocks.pdf.

64. Draine, B. T. *Physics of the Interstellar and Intergalactic Medium* (Princeton University Press, Princeton, NJ, 2011).

65. Weingartner, J. C. ASTR 480 Lecture Notes (2022). URL http://physics.gmu.edu/~joe/ASTR480.html.

66. Chandrasekhar, S. On stars, their evolution and their stability (1983). URL https://www.nobelprize.org/uploads/2018/06/chandrasekhar-lecture.pdf.

67. Hessels, J. W. T. *et al.* A radio pulsar spinning at 716 hz. *Science* **311**, 1901–1904 (2006). URL http://dx.doi.org/10.1126/science.1123430.

68. Hulse, R. A. & Taylor, J. H. Discovery of a pulsar in a binary system. *ApJL* **195**, L51–L53 (1975).

69. The hulse-taylor pulsar – evidence of gravitational waves. URL https://blogs.cardiff.ac.uk/physicsoutreach/gravitational-physics-tutorial/3-the-hulse-taylor-pulsar-evidence-of-gravitational-waves/.

70. Kulkarni, S. Supernovae in binary systems: Orbital dynamics (2004). URL https://sites.astro.caltech.edu/~srk/ay125/SN-Explosion.pdf.

71. Young, P. J., Corwin, J., H. G., Bryan, J. & de Vaucouleurs, G. Light curve of nova V1500 Cygni 1975. *ApJ* **209**, 882–894 (1976).

72. Crab pulsar. URL https://en.wikipedia.org/wiki/Crab_Pulsar.

73. Becker, P. A. & Wolff, M. T. Thermal and Bulk Comptonization in Accretion-powered X-Ray Pulsars. *ApJ* **654**, 435–457 (2007). astro-ph/0609035.

74. Kalapotharakos, C., Wadiasingh, Z., Harding, A. K. & Kazanas, D. The Multipolar Magnetic Field of the Millisecond Pulsar PSR J0030+0451. *ApJ* **907**, 63 (2021). 2009.08567.

75. Miller, C. Magnetic accretion onto neutron stars. URL https://www.astro.umd.edu/~miller/teaching/astr498/lecture19.pdf.

76. Anomalous x-ray pulsar. URL https://astronomy.swin.edu.au/cosmos/a/Anomalous+X-ray+Pulsar.

77. Piran, T. Gamma-ray bursts and the fireball model. *Physics Reports* **314**, 575–667 (1999). URL http://dx.doi.org/10.1016/S0370-1573(98)00127-6.

78. Knust, F. Applying the fireball model to short gamma-ray burst afterglows: Methods, jet opening angles and plateau phases. URL https://www.imprs-astro.mpg.de/sites/default/files/fabian_knust.pdf.

79. Berger, E. Short-duration gamma-ray bursts. *Annual Review of Astronomy and Astrophysics* **52**, 43–105 (2014). URL http://dx.doi.org/10.1146/annurev-astro-081913-035926.

80. Abbott, B. *et al.* Gw170817: Observation of gravitational waves from a binary neutron star inspiral. *Physical Review Letters* **119** (2017). URL http://dx.doi.org/10.1103/PhysRevLett.119.161101.

81. Salpeter, E. E. The Luminosity Function and Stellar Evolution. *ApJ* **121**, 161 (1955).

82. Kroupa, P. The Initial Mass Function of Stars: Evidence for Uniformity in Variable Systems. *Science* **295**, 82–91 (2002). astro-ph/0201098.

83. Chabrier, G. Galactic Stellar and Substellar Initial Mass Function. *PASP* **115**, 763–795 (2003). astro-ph/0304382.

84. Wiktorowicz, G. *et al.* Populations of stellar-mass black holes from binary systems. *The Astrophysical Journal* **885**, 1 (2019). URL https://dx.doi.org/10.3847/1538-4357/ab45e6.

85. Staubert, R. *et al.* Cyclotron lines in highly magnetized neutron stars. *A&A* **622**, A61 (2019). 1812.03461.

86. Johnston, S. & Karastergiou, A. Pulsar braking and the p–$\dot{P}$ diagram. *Monthly Notices of the Royal Astronomical Society* **467**, 3493–3499 (2017). URL http://dx.doi.org/10.1093/mnras/stx377.

87. Baldini, U. *JET'S POWER IN RADIO-LOUD ACTIVE GALACTIC NUCLEI: A CASE STUDY ON THE NATURE OF BLAZAR CANDIDATES*. Ph.D. thesis (2015).

88. Yang, G. *et al.* x-cigale: fitting agn/galaxy seds from x-ray to infrared. *Monthly Notices of the Royal Astronomical Society* **491**, 740–757 (2019). URL http://dx.doi.org/10.1093/mnras/stz3001.

89. Mountrichas, G. *et al.* Galaxy properties of type 1 and 2 X-ray selected AGN and a comparison among different classification criteria. *A&A* **653**, A70 (2021). 2106.11579.

90. McKee, C. F. & Ostriker, J. P. A theory of the interstellar medium: three components regulated by supernova explosions in an inhomogeneous substrate. *ApJ* **218**, 148–169 (1977).

91. Flynn, C. Lecture 3 : Accelerated charges and bremsstrahlung (2006). URL https://www.astro.utu.fi/~cflynn/astroII/l3.html#bsrad.

92. Ghisellini, G. *Radiative Processes in High Energy Astrophysics* (Springer International Publishing, 2013). URL http://dx.doi.org/10.1007/978-3-319-00612-3.

93. Rybicki, G. B. & Lightman, A. P. *Radiative Processes in Astrophysics* (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 1979).

94. Pohl, M. K. W. Thermal bremsstrahlung (2008). URL https://www-zeuthen.desy.de/~pohlmadq/teach/405-505-f05/sess18.pdf.

95. Young, L. M. Bremsstrahlung radiation (2009). URL http://kestrel.nmt.edu/~lyoung/426/Chap4.pdf.

96. University, S. Synchrotron emission. URL https://astronomy.swin.edu.au/cosmos/s/synchrotron+emission.

97. Eilek, J. A. & Arendt, P. N. The Synchrotron Spectrum of Diffuse Radio Sources: Effects of Particle and Field Distributions. *ApJ* **457**, 150 (1996).

98. Knapp, J. Ionization, saha equation (2011). URL https://www.astro.princeton.edu/~gk/A403/ioniz.pdf.

99. Andersson, B.-G. Interstellar dust grain alignment (2019). URL https://ntrs.nasa.gov/api/citations/20200001522/downloads/20200001522.pdf.

100. Livio, M. & Kamionkowski, M. Commentary: BICEP2's B modes: Big Bang or dust? *Physics Today* **67**, 8–10 (2014). URL https://doi.org/10.1063/PT.3.2598. https://pubs.aip.org/physicstoday/article-pdf/67/12/8/10098737/8_1_online.pdf.

101. Condon, J. J. & Ransom, S. M. Essential radio astronomy (2018). URL https://www.cv.nrao.edu/~sransom/web/Ch5.html.

102. Domínguez, A. *et al.* Detection of the cosmic gamma-ray horizon from multiwavelength observations of blazars. *The Astrophysical Journal* **770**, 77 (2013). URL http://dx.doi.org/10.1088/0004-637X/770/1/77.

103. Martinez, M. Fundamental physics with cosmic gamma rays. In *Journal of Physics Conference Series*, vol. 171 of *Journal of Physics Conference Series*, 012013 (2009).

104. Cardelli, J. A., Clayton, G. C. & Mathis, J. S. The Relationship between Infrared, Optical, and Ultraviolet Extinction. *ApJ* **345**, 245 (1989).

105. Calzetti, D. *et al.* The Dust Content and Opacity of Actively Star-forming Galaxies. *ApJ* **533**, 682–695 (2000). astro-ph/9911459.

106. van Dishoeck, E. F. ISO Spectroscopy of Gas and Dust: From Molecular Clouds to Protoplanetary Disks. *ARA&A* **42**, 119–167 (2004). astro-ph/0403061.

107. Mo, H., van den Bosch, F. & White, S. *Galaxy Formation and Evolution* (Cambridge University Press, Cambridge, UK, 2010).

108. Pogge, R. W. Introduction to the interstellar medium (2023). URL https://www.astronomy.ohio-state.edu/pogge.1/Ast871/Notes/Intro.pdf.

109. van den Bosch, F. Heating & cooling (2022). URL http://www.astro.yale.edu/vdbosch/astro610_lecture15.pdf.

110. Bradt, H. *Astrophysics Processes* (Cambridge University Press, Cambdrige, UK, 2008).

111. Cordes, J. M. & Lazio, T. J. W. NE2001. II. Using Radio Propagation Data to Construct a Model for the Galactic Distribution of Free Electrons. *arXiv e-prints* astro–ph/0301598 (2003). astro-ph/0301598.

112. Cordes, J. M. & Chatterjee, S. Fast Radio Bursts: An Extragalactic Enigma. *ARA&A* **57**, 417–465 (2019). 1906.05878.

113. Ogilvie, G. I. Astrophysical fluid dynamics. *Journal of Plasma Physics* **82**, 205820301 (2016).

114. Murphy, N. Ideal magnetohydrodynamics (2014). URL https://lweb.cfa.harvard.edu/~namurphy/Lectures/Ay253_01_IdealMHD.pdf.

115. Murphy, N. Magnetohydrodynamic waves (2016). URL https://lweb.cfa.harvard.edu/~namurphy/Lectures/Ay253_2016_07_MHDwaves.pdf.

116. Cairns, I. Mhd waves (1999). URL http://www.physics.usyd.edu.au/~cairns/teaching/lecture3/node13.html.

117. Howes, G. Lecture 16: The frozen-in flux theorem. URL https://homepage.physics.uiowa.edu/~ghowes/teach/phys4731/lect/4731lec16.pdf.

118. Meneghetti, M. *Introduction to Gravitational Lensing With Python Examples* (Springer, 2021).

119. McDonald, M. 8.902 Lecture Notes (2022).

120. Schneider, P. *Extragalactic Astronomy and Cosmology: An Introduction* (Springer Berlin, Heidelberg, 2015), 2nd edn.

121. Wong, K. C. *et al.* H0licow – xiii. a 2.4 per cent measurement of h0 from lensed quasars: 5.3sigma tension between early- and late-universe probes. *Monthly Notices of the Royal Astronomical Society* **498**, 1420–1439 (2019). URL http://dx.doi.org/10.1093/mnras/stz3094.

122. Shapiro, I. I. *et al.* Fourth Test of General Relativity: Preliminary Results. *PhRvL* **20**, 1265–1269 (1968).

123. Shapiro, I. I. *et al.* Fourth Test of General Relativity: New Radar Result. *PhRvL* **26**, 1132–1135 (1971).

124. Bertone, G. & Hooper, D. History of dark matter. *Reviews of Modern Physics* **90**, 045002 (2018). 1605.04909.

125. Sarazin, C. L. *X-ray emission from clusters of galaxies* (1988).

126. Binney, J. & Tremaine, S. *Galactic Dynamics* (Princeton University Press, Princeton, NJ, 2008), 2nd edn.

127. Vogelsberger, M. 8.902 lecture notes (2021).

128. Nave, R. Hyperphysics: Freefall time and star formation time. URL http://hyperphysics.phy-astr.gsu.edu/hbase/Astro/starff.html#c2.

129. Li, C., Zhao, G. & Yang, C. Galactic Rotation and the Oort Constants in the Solar Vicinity. *ApJ* **872**, 205 (2019).

130. Press, W. H. & Schechter, P. Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. *ApJ* **187**, 425–438 (1974).

131. Planck Collaboration *et al.* Planck 2018 results. VIII. Gravitational lensing. *A&A* **641**, A8 (2020). 1807.06210.

132. McKee, C. F., Parravano, A. & Hollenbach, D. J. Stars, gas, and dark matter in the solar neighborhood. *The Astrophysical Journal* **814**, 13 (2015). URL https://dx.doi.org/10.1088/0004-637X/814/1/13.

133. Bahcall, N. A. Clusters and superclusters of galaxies (1996). astro-ph/9611148.

134. White, S. D. M., Navarro, J. F., Evrard, A. E. & Frenk, C. S. The baryon content of galaxy clusters: a challenge to cosmological orthodoxy. *Nature* **366**, 429–433 (1993). URL https://doi.org/10.1038/366429a0.

135. Hong, T., Han, J. L., Wen, Z. L., Sun, L. & Zhan, H. The correlation function of galaxy clusters and detection of baryon acoustic oscillations. *The Astrophysical Journal* **749**, 81 (2012). URL http://dx.doi.org/10.1088/0004-637X/749/1/81.

136. Eisenstein, D. J. *et al.* Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies. *ApJ* **633**, 560–574 (2005). astro-ph/0501171.

137. Ruhlmann-Kleider, V. (2015). URL http://www.scholarpedia.org/article/Cosmological_constraints_from_baryonic_acoustic_oscillation_measurements.

138. Beckmann, V. & Shrader, C. *Active Galactic Nuclei* (Wiley-VCH, 2012).

139. Schmidt, M. The Rate of Star Formation. *ApJ* **129**, 243 (1959).

140. Kennicutt, J., Robert C. The Global Schmidt Law in Star-forming Galaxies. *ApJ* **498**, 541–552 (1998). astro-ph/9712213.

141. Kormendy, J. & Ho, L. C. Coevolution (Or Not) of Supermassive Black Holes and Host Galaxies. *ARA&A* **51**, 511–653 (2013). 1304.7762.

142. Madau, P. & Dickinson, M. Cosmic star-formation history. *Annual Review of Astronomy and Astrophysics* **52**, 415–486 (2014). URL http://dx.doi.org/10.1146/annurev-astro-081811-125615.

143. Speagle, J. S., Steinhardt, C. L., Capak, P. L. & Silverman, J. D. A Highly Consistent Framework for the Evolution of the Star-Forming "Main Sequence" from z $\sim$0-6. *ApJS* **214**, 15 (2014). 1405.2041.

144. Bullock, J. S. Notes on the missing satellites problem (2010). 1009.4505.

145. de Blok, W. J. G. The core-cusp problem. *Advances in Astronomy* **2010**, 1–14 (2010). URL http://dx.doi.org/10.1155/2010/789293.

146. Del Popolo, A. & Le Delliou, M. Review of Solutions to the Cusp-Core Problem of the ΛCDM Model. *Galaxies* **9** (2021). URL https://www.mdpi.com/2075-4434/9/4/123.

147. Fabian, A. C. Cooling Flows in Clusters of Galaxies. *ARA&A* **32**, 277–318 (1994).

148. Chatzikos, M. *et al.* Implications of coronal line emission in NGC 4696*. *MNRAS* **446**, 1234–1244 (2015). 1410.4566.

149. Fabian, A. C. *et al.* Hidden cooling flows in clusters of galaxies. *Monthly Notices of the Royal Astronomical Society* **515**, 3336–3345 (2022). URL http://dx.doi.org/10.1093/mnras/stac2003.

150. Takahashi, T. 21 cm cosmology (2015). URL https://www2.yukawa.kyoto-u.ac.jp/~ppp.ws/PPP2015/slides/takahashi.pdf.

151. Pritchard, J. R. & Loeb, A. 21 cm cosmology in the 21st century. *Reports on Progress in Physics* **75**, 086901 (2012). URL http://dx.doi.org/10.1088/0034-4885/75/8/086901.

152. Tegmark, M. Doppler peaks and all that: Cmb anisotropies and what they can tell us (1995). astro-ph/9511148.

153. van den Bosch, F. Lecture 5: Newtonian perturbation theory ii. baryonic perturbations (2024). URL http://www.astro.yale.edu/vdbosch/astro610_lecture5.pdf.

154. Hu, W. Wandering in the background: A cmb explorer (2009). astro-ph/9508126.

155. Aghanim, N. *et al.* Planck2018 results: Vi. cosmological parameters. *Astronomy & Astrophysics* **641**, A6 (2020). URL http://dx.doi.org/10.1051/0004-6361/201833910.

156. Hu, W. Intermediate guide to the acoustic peaks and polarization (2001). URL http://background.uchicago.edu/~whu/intermediate/baryons.html.

157. The cycloid. URL http://mathonline.wikidot.com/the-cycloid.

158. Hubble, E. A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. *Proceedings of the National Academy of Science* **15**, 168–173 (1929).

159. Wikipedia. Cosmic distance ladder. URL https://en.wikipedia.org/wiki/Cosmic_distance_ladder.

160. Suyu, S. H. *et al.* H0licow – i. h0 lenses in cosmograil's wellspring: program overview. *Monthly Notices of the Royal Astronomical Society* **468**, 2590–2604 (2017). URL http://dx.doi.org/10.1093/mnras/stx483.

161. Abbott, B. P. *et al.* A gravitational-wave standard siren measurement of the Hubble constant. *Nature* **551**, 85–88 (2017). 1710.05835.

162. Di Valentino, E. A combined analysis of the h0 late time direct measurements and the impact on the dark energy sector. *Monthly Notices of the Royal Astronomical Society* **502**, 2065–2073 (2021). URL http://dx.doi.org/10.1093/mnras/stab187.

163. Bolte, M., Waters, R. & Wilden, B. Measuring stellar ages (2000). URL https://www.ucolick.org/~bolte/AY4_00/week7/cluster_ages.html.

164. Morris, J. M. Mapping the friedmann equations (2023). URL https://johnmarkmorris.com/2023/05/29/mapping-the-friedmann-equations/.

165. Cortês, M. & Liddle, A. R. Interpreting desi's evidence for evolving dark energy (2024). 2404.08056.

166. Madau, P. Radiative Transfer in a Clumpy Universe: The Colors of High-Redshift Galaxies. *ApJ* **441**, 18 (1995).

167. (2013). URL https://astrobites.org/2013/07/21/astrophysical-classics-neutral-hydrogen-in-t?__cf_chl_tk=4ehOAO7X6i1zTSsPla3BGBYrG_Q1vS_vVRTQ5LSEO0M-1708211463-0.0-4903.

168. Tsujikawa, S. Introductory review of cosmic inflation (2003). hep-ph/0304257.

169. Guth, A. H. Inflationary universe: A possible solution to the horizon and flatness problems. *Phys. Rev. D* **23**, 347–356 (1981). URL https://link.aps.org/doi/10.1103/PhysRevD.23.347.

170. Nicastro, F. *et al.* Observations of the missing baryons in the warm-hot intergalactic medium. *Nature* **558**, 406–409 (2018). 1806.08395.

171. Balazs, C. Baryogenesis: A small review of the big picture (2014). URL https://arxiv.org/abs/1411.3398.

172. Sampoorna, M. (2003). URL https://web.iucaa.in/~dipankar/ph217/contrib/compton.pdf.

173. URL http://user.astro.columbia.edu/~jules/GR6001_17/SZ_effect.pdf.

174. NASA. Cosmic background explorer. URL https://science.nasa.gov/mission/cobe/.

175. Wilkinson microwave anisotropy probe. URL https://map.gsfc.nasa.gov/.

176. Larson, D. *et al.* Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Power Spectra and WMAP-derived Parameters. *ApJS* **192**, 16 (2011). 1001.4635.

177. ESA. Planck. URL https://www.esa.int/Enabling_Support/Operations/Planck.

178. Aghanim, N. *et al.* Planck 2018 results: V. cmb power spectra and likelihoods. *Astronomy & Astrophysics* **641**, A5 (2020). URL http://dx.doi.org/10.1051/0004-6361/201936386.

179. South pole telescope. URL https://pole.uchicago.edu/public/Home.html.

180. Balkenhol, L. *et al.* Measurement of the cmb temperature power spectrum and constraints on cosmology from the spt-3g 2018 dataset. *Physical Review D* **108** (2023). URL http://dx.doi.org/10.1103/PhysRevD.108.023510.

181. NASA. Atacama cosmology telescope. URL https://lambda.gsfc.nasa.gov/product/act/#:~:text=It%20is%20designed%20to%20make,'dovich%20(tSZ)%20effect.

182. Choi, S. K. *et al.* The atacama cosmology telescope: a measurement of the cosmic microwave background power spectra at 98 and 150 ghz. *Journal of Cosmology and Astroparticle Physics* **2020**, 045–045 (2020). URL http://dx.doi.org/10.1088/1475-7516/2020/12/045.

183. Carlstrom, J. Cmb stage 4 progress. URL https://indico.fnal.gov/event/12206/contributions/12877/attachments/8517/10858/CMB-S4_PAC_160621v2.pdf.

184. JPL, N. URL https://science.jpl.nasa.gov/projects/BICEP3/.

185. Ellis, R. S. The formation and evolution of galaxies (1998). astro-ph/9807287.

186. KIT. URL https://www.katrin.kit.edu/index.php.

187. Aker, M. *et al.* Direct neutrino-mass measurement with sub-electronvolt sensitivity. *Nature Physics* **18**, 160–166 (2022). URL https://doi.org/10.1038/s41567-021-01463-1.

188. Armitage, P. (2004). URL https://jila.colorado.edu/~pja/astr3830/lecture02.pdf.

189. Group, S. URL https://aether.lbl.gov/www/projects/neutrino/grb/logn_logs.html.

190. Padmanabhan, T. Advanced topics in cosmology: A pedagogical introduction. In *AIP Conference Proceedings* (AIP, 2006). URL http://dx.doi.org/10.1063/1.2219327.

191. MacGregor, M. Interferometry basics (2018). URL https://science.nrao.edu/facilities/alma/naasc-workshops/nrao-cd-jhu18/Interferometry_Basics.pdf.

192. Francesco, J. D. A crash course in radio astronomy and interferometry: 2. aperture synthesis. URL https://science.nrao.edu/science/meetings/presentation/jdf.webinar.2.pdf.

193. Ricci, L., Peck, A., Braatz, J., Bemis, A. & Stierwalt, S. Introduction to radio interferometry. URL https://science.nrao.edu/facilities/alma/naasc-workshops/nrao-cd-ru17/InterfBasics_Rice.pdf.

194. Event horizon telescope (2023). URL https://eventhorizontelescope.org/.

195. Griffiths, D. J. *Introduction to Electrodynamics* (Cambridge University Press, 2017), 4th edn.

196. Christensen, F. E. & Ramsey, B. D. *X-Ray Optics for Astrophysics: A Historical Review*, 1–42 (Springer Nature Singapore, 2022). URL http://dx.doi.org/10.1007/978-981-16-4544-0_1-1.

197. Maughan, B. Introduction to x-ray astronomy. URL https://www.star.bris.ac.uk/bjm/lectures/x-ray/x-ray.pdf.

198. Bolte, M. Signal-to-noise in optical astronomy (2006). URL https://www.ucolick.org/~bolte/AY257/s_n.pdf.

199. Pedrotti, F. L., Pedrotti, L. M. & Pedrotti, L. S. *Introduction to Optics* (Addison-Wesley, 2006), 3rd edn.

200. Greene, J. Homework #4 solutions, ast 303 (2018). URL https://www.astro.princeton.edu/~jgreene/ast303/Homework_4_soln.pdf.

201. Schmidt camera. URL https://en.wikipedia.org/wiki/Schmidt_camera.

202. Ritchey-chretien telescope. URL https://en.wikipedia.org/wiki/Ritchey%E2%80%93Chr%C3%A9tien_telescope.

203. Djorgovski, G. Spectrographs and spectroscopy (2012). URL https://sites.astro.caltech.edu/~george/ay122/Ay122a_Spectroscopy.pdf.

204. How does a spectrograph work? URL https://www.atnf.csiro.au/outreach/education/senior/astrophysics/spectrographs.html.

205. Slitless spectrograph. URL http://astro.vaporia.com/start/slitlessspectrograph.html.

206. Integral field spectrograph. URL https://en.wikipedia.org/wiki/Integral_field_spectrograph.

207. Coronagraph. URL https://spaceweather.com/glossary/coronagraph.html.

208. URL https://spie.org/publications/spie-publication-resources/optipedia-free-optics-information/tt61_541_laser_interferometer#_=_.

209. Adaptive optics. URL https://www.eso.org/public/teles-instr/technology/adaptive_optics/.

210. Voyager. URL https://voyager.jpl.nasa.gov/mission/timeline/#event-a-once-in-a-lifetime-alignment.

211. Hubble space telescope. URL https://science.nasa.gov/mission/hubble/.

212. Hubble's instruments. URL https://esahubble.org/about/general/instruments/.

213. Magellan telescopes. URL https://www.cfa.harvard.edu/facilities-technology/telescopes-instruments/magellan-telescopes.

214. Chandra x-ray observatory. URL https://chandra.harvard.edu/about/science_instruments.html.

215. Murchison widefield array. URL https://www.mwatelescope.org/.

216. Neil gehrels swift observatory. URL https://swift.gsfc.nasa.gov/about_swift/#science.

217. The rosat mission. URL https://heasarc.gsfc.nasa.gov/docs/rosat/rosat3.html.

218. Spitzer. URL https://science.nasa.gov/mission/spitzer/.

219. Tess mission. URL https://tess.mit.edu/science/.

220. Ligo. URL https://www.ligo.caltech.edu/page/about.

221. Lisa. URL https://lisa.nasa.gov/.

222. Nicer. URL https://heasarc.gsfc.nasa.gov/docs/nicer/index.html.

223. The canadian hydrogen intensity mapping experiment is a revolutionary new canadian radio telescope designed to answer major questions in astrophysics and cosmology. URL https://chime-experiment.ca/en.

224. Hera: Hydrogen epoch of reionization array. URL https://reionization.org/.

225. Lourie, N. P. *et al.* The wide-field infrared transient explorer (winter). In Evans, C. J., Bryant, J. J. & Motohara, K. (eds.) *Ground-based and Airborne Instrumentation for Astronomy VIII* (SPIE, 2020). URL http://dx.doi.org/10.1117/12.2561210.

226. Webb's scientific instruments. URL https://webbtelescope.org/news/webb-science-writers-guide/webbs-scientific-instruments.

227. The nancy grace roman space telescope. URL https://www.jpl.nasa.gov/missions/the-nancy-grace-roman-space-telescope.

228. Vera c. rubin observatory: About. URL https://rubinobservatory.org/about.

229. Atmospheric windows. URL http://earthguide.ucsd.edu/eoc/special_topics/teach/sp_climate_change/p_atmospheric_window.html.

230. Bouchy, F., Pepe, F. & Queloz, D. Fundamental photon noise limit to radial velocity measurements. *A&A* **374**, 733–739 (2001).

231. Figueira, P., Curto, G. L. & Mehner, A. Very large telescope paranal science operations espresso user manual (2020). URL https://www.eso.org/sci/facilities/paranal/instruments/espresso/ESPRESSO_User_Manual_P106_3.pdf.

232. Guyon, O. Radial velocity measurements and exoplanets (2012). URL https://subarutelescope.org/staff/guyon/15teaching.web/01AstrOptics2012.web/AstOpt2012_06Spectroscopy.pdf.

233. Boquien, M. *et al.* CIGALE: a python Code Investigating GALaxy Emission. *A&A* **622**, A103 (2019). 1811.03094.

234. Galaxy spectra. URL http://astronomy.nmsu.edu/nicole/teaching/ASTR505/lectures/lecture26/slide01.html.

235. Hartigan, P. HE II Line Emission as a Tracer of Nonradiative Preionized Shocks: Application to Winds and Circumstellar Disks in the Orion Nebula. *ApJ* **526**, 274–278 (1999).

236. Orion nebula. URL https://en.wikipedia.org/wiki/Orion_Nebula.

237. Messier 13. URL https://en.wikipedia.org/wiki/Messier_13.

238. Andromeda galaxy. URL https://en.wikipedia.org/wiki/Andromeda_Galaxy.

239. Virgo cluster. URL https://en.wikipedia.org/wiki/Virgo_Cluster.

240. Coma cluster. URL https://en.wikipedia.org/wiki/Coma_Cluster.

241. 3c 273. URL https://en.wikipedia.org/wiki/3C_273.

242. Cassiopeia a. URL https://en.wikipedia.org/wiki/Cassiopeia_A.

243. Neugebauer, G. & Becklin, E. E. The brightest infrared sources. *Scientific American* **228**, 28–42 (1973). URL http://www.jstor.org/stable/24923023.

244. What are the (apparently) brightest x-ray sources in the sky as seen from the earth? URL https://heasarc.gsfc.nasa.gov/docs/heasarc/headates/brightest.html.

245. Crab nebula. URL https://en.wikipedia.org/wiki/Crab_Nebula.

246. Messier 3. URL https://en.wikipedia.org/wiki/Messier_3.

247. Messier 87. URL https://en.wikipedia.org/wiki/Messier_87.

248. Ginsburg, A. A review of the w51 cloud (2017). 1702.06627.

249. Cygnus x-1. URL https://en.wikipedia.org/wiki/Cygnus_X-1.

250. Cygnus a. URL https://en.wikipedia.org/wiki/Cygnus_A.

251. Ss 433. URL https://en.wikipedia.org/wiki/SS_433.

252. Bowler, M. G. SS 433: Two robust determinations fix the mass ratio. *A&A* **619**, L4 (2018).

253. First observation of gravitational waves. URL https://en.wikipedia.org/wiki/First_observation_of_gravitational_waves.

254. Gw170817. URL https://en.wikipedia.org/wiki/GW170817.

255. Large magellanic cloud. URL https://en.wikipedia.org/wiki/Large_Magellanic_Cloud.

256. Boötes void. URL https://en.wikipedia.org/wiki/Bo%C3%B6tes_Void.

257. Trappist-1. URL https://en.wikipedia.org/wiki/TRAPPIST-1.

258. Proxima centauri b. URL https://en.wikipedia.org/wiki/Proxima_Centauri_b.

259. Sagittarius a*. URL https://en.wikipedia.org/wiki/Sagittarius_A*.

260. Postman, M. *et al.* A brightest cluster galaxy with an extremely large flat core. *The Astrophysical Journal* **756**, 159 (2012). URL http://dx.doi.org/10.1088/0004-637X/756/2/159.

261. Lauer, T. R. *et al.* The Masses of Nuclear Black Holes in Luminous Elliptical Galaxies and Implications for the Space Density of the Most Massive Black Holes. *ApJ* **662**, 808–834 (2007). astro-ph/0606739.

262. Quinlan, G. D. The dynamical evolution of massive black hole binaries i. hardening in a fixed stellar background. *New Astronomy* **1**, 35–56 (1996). URL http://dx.doi.org/10.1016/S1384-1076(96)00003-6.

263. Milosavljević, M. & Merritt, D. Formation of Galactic Nuclei. *ApJ* **563**, 34–62 (2001). astro-ph/0103350.

264. Milosavljević, M. & Merritt, D. The Final Parsec Problem. In Centrella, J. M. (ed.) *The Astrophysics of Gravitational Wave Sources*, vol. 686 of *American Institute of Physics Conference Series*, 201–210 (2003). astro-ph/0212270.

265. Faber, S. M. *et al.* The Centers of Early-Type Galaxies with HST. IV. Central Parameter Relations. *AJ* **114**, 1771 (1997). astro-ph/9610055.

266. Djorgovski, G. Thermal history of the universe (2008). URL https://sites.astro.caltech.edu/~george/ay127/kamionkowski-earlyuniverse-notes.pdf.

267. Davenhall, J. P. . A. The ccd photometric calibration cookbook (2001). URL http://star-www.rl.ac.uk/docs/sc6.htx/sc6se7.html.

268. Ulisse Munari, D. M., Massimo Fiorucci. The asiago database on photometric systems (2002). URL http://ulisse.pd.astro.it/Astro/ADPS/Systems/index.html.

269. Filter profile service. URL http://svo2.cab.inta-csic.es/theory/fps/index.php?mode=browse&gname=Spitzer&asttype=.

270. Fraunhofer lines. URL https://en.wikipedia.org/wiki/Fraunhofer_lines.

271. Term symbol. URL https://en.wikipedia.org/wiki/Term_symbol.

272. Chojnowski, D. Table of galaxy emission lines from 700Å to 11,000Å (2015). URL http://astronomy.nmsu.edu/drewski/tableofemissionlines.html.

273. Chapter 10: Derivation of the planck formula. URL https://edisciplinas.usp.br/pluginfile.php/48089/course/section/16461/qsp_chapter10-plank.pdf.

274. Oke, J. B. & Sandage, A. Energy Distributions, K Corrections, and the Stebbins-Whitford Effect for Giant Elliptical Galaxies. *ApJ* **154**, 21 (1968).

275. Hogg, D. W., Baldry, I. K., Blanton, M. R. & Eisenstein, D. J. The k correction (2002). astro-ph/0210394.

276. Hogg, D. W. Distance measures in cosmology (2000). astro-ph/9905116.

277. URL https://explainingscience.org/2021/03/13/distances-in-cosmology/.