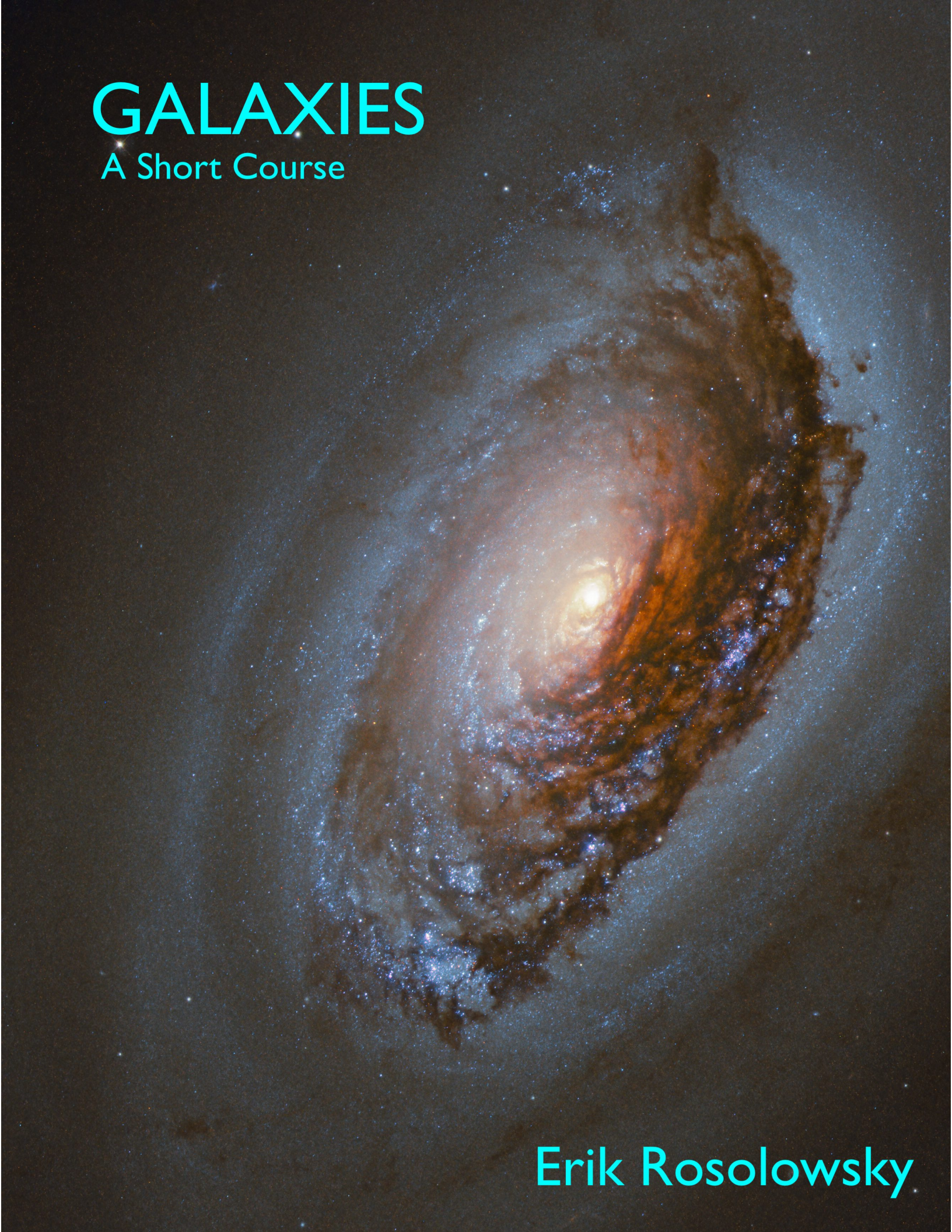
GALAXIES

A Short Course



Erik Rosolowsky

DO THE DUMBEST THING FIRST.

— D. P. FINKBEINER

ERIK ROSOLOWSKY

GALAXIES

A SHORT COURSE

DRAFT DATE: APRIL 22, 2022

PUBLISHED BY DRAFT DATE: APRIL 22, 2022

TUFTE-LATEX.GOOGLECODE.COM

First printing, April 2022

Contents

List of Figures

List of Tables

Acknowledgments

I am grateful to the students of PHYS 495/595 and ASTRO 322 who helped edit these notes to fix the many typos, errors, and conceptual problems. This includes (alphabetically): Sanjina Aurin, Dhananjhay Bansal, Soumen Deb, Ashaduzzaman Joy, Eric Koch, Kenneth Opena, Nicolas Pacholok, Aditya Shah, Kenny Van, Yue Zhao. I welcome suggestions and corrections to these notes, which can earn you a coveted spot in this paragraph. I'm also grateful for my colleague Dr. Greg Sivakoff who provided expert summaries of parts of the literature.

1
Observational Astronomy

Astronomy is the observation of the Universe beyond Earth. Right now, our opportunity to go out into space and actually conduct on-site science is limited. We've sent probes to a handful of planets and visited one (1) celestial body, i.e., the moon. However, we can tell a surprisingly coherent, scientifically motivated explanation of the workings of the Universe over its 14 billion years of evolution. We accomplish this wild feat using messenger particles, which come from the Universe to us here on Earth, and we use the properties of the detected messenger to study the objects that emitted the messengers. This inference is accomplished using the fundamental assumption of astrophysics: *the same physical laws operate throughout the visible Universe as they do here on Earth.* For example, since Coulomb's law in an inverse square law here on Earth, it is an inverse square law everywhere that the messengers come from. A stronger assumption is that the physical constants (compiled in the Appendix) are the same through the Universe. This is not necessarily true but not yet demonstrated to be false. This assumption allows us to activate the infrastructure of physics to interpret the messengers that we observe, transforming astronomy into *astrophysics*. This field is broadly divided into the dialogue between two branches: *observational astrophysics* which collects the information from our messenger species and *theoretical astrophysics* which provides models for interpreting those observations in terms of the objects we study.

In this Chapter, we will give an overview of the astrophysical messengers that we are using to study the Universe, including light, particles and gravitational waves. Since light is the standard tool for astronomical observations, we will then explore how we quantify the measurement of light in terms of its amount and direction. Finally, we will describe the basic properties of telescopes that shape the resulting observations.

1.1 Astronomical Messengers

1.1.1 Light

Nearly all observational astronomy is done by measuring the light from the Universe after it travels through space to arrive at Earth. Thanks to spooky quantum mechanics, light can be effectively described as both as a particle and as a self-propagating electromagnetic wave.

When viewed as a wave, light follows a dispersion relation $c = \lambda\nu$, where the speed of light in a vacuum $c = 299\,792\,458$ m s^{-1} $\approx 3.00 \times 10^8$ m s^{-1}, λ is the wavelength and ν is the frequency of the wave. The wavelength / frequency of the wave specifies the type of light, allowing us to categorize light into the electromagnetic (EM) spectrum. The following Table shows the typical divisions of the electromagnetic spectrum separated into its primary *wavebands* based on the frequencies / wavelengths of the light. These divisions between the different wavebands are approximate, but they represent the major parts of the EM spectrum that are observed by telescopes. Common abbreviations are given next to each of the names of the wavebands in brackets.

Type of Light	Characteristic Wavelengths	Characteristic Frequencies	Characteristic Photon Energies
Radio	4 mm - 100 m	3 MHz - 50 GHz	10 neV - 0.2 meV
Millimetre (mm)	1 mm - 4 mm	70 GHz - 300 GHz	0.2 - 1 mev
Submillimetre (submm)	300 μm - 1 mm	300 GHz-1 THz	1 - 3 meV
Far Infrared (FIR)	70 μm - 300 μm	1-4 THz	3 - 12 meV
Mid Infrared (MIR)	20 μm - 70 μm	4 -15 THz	12-60 meV
Near Infrared (NIR)	700 nm - 20 μm	15 THz - 400 THz	0.06 - 1 eV
Visible	400 nm - 700 nm	400 - 750 THz	1-3 eV
Near Ultraviolet (NUV)	180 - 400 nm	0.75 - 1.5 PHz	3-7 eV
Far Ultraviolet (FUV)	91.2 - 180 nm	1.5 - 3.2 PHz	7-13.6 eV
Extreme Ultraviolet (EUV)	6.2 - 91.2 nm	3-50 PHz	13.6-200 eV
X-ray	0.01 - 6.2 nm	50 - 3000 PHz	0.2-100 keV
Gamma ray	$<$ 10 pm	$>$ 30 EHz	$>$ 100 keV

Table 1.1: The different wavebands in the electromagnetic spectrum.

Light can also be viewed as a particle called a *photon* (commonly abbreviated with a γ), where the energy in a given photon (E_γ) is related to its frequency $E_\gamma = h\nu$. Here, Planck's constant $h = 6.63 \times 10^{-34}$ J s . The notable thing about Planck's constant is that it's quite small, which just means that photon energies are relatively small compared to the metre-kilogram-second scale that are so useful for human-sized measurements. For quantifying energies that are the scale of quantum mechanics, specifically the energies that typify the

electronic transitions of atoms, it is useful to express the energies instead in units of electron-volts (eV) where 1 eV $= 1.60 \times 10^{-19}$ J such that the ionization energy of a hydrogen atom is then $E_R =$ 13.6 eV rather than the dull 2.12×10^{-18} J. [1]

[1] tl;dr: $E = h\nu$; $c = \lambda\nu$; and 1 eV $= 1.60 \times 10^{-19}$ J.

The transition between where it's best to consider light as a particle vs light as a wave is usually governed by how much energy is in the light compared to an individual photon energy. If there are lots of photons (100s or more), then the wave description is usually used. For smaller numbers of photons, the individual photon description can be more appropriate.

Astronomical convention also typically favours the use of one set of units in a given waveband, even though all descriptions of light – wavelength, frequency, and photon energy – are equivalent. Frequencies are typically used to describe light in the radio to the submillimetre, wavelengths are usually used from the Far Infrared to the Far Ultraviolet and energies are used for X-rays and gamma rays. These conventions are not universal and are mostly set by the legacy of the engineering decisions that drove the original detector technologies. For example, radio telescopes were developed by radio-frequency engineers, which traditionally used frequency to describe their radiation (a radio in a car expresses the frequency of the broadcasting station rather than the wavelength). Hence, radio astronomers still favour the use of frequency to describe their light. Optical astronomers typically used instruments that used physical optics to set up wave diffraction and interference. Since these devices depend on the scale of the instrument relative to the wavelength of light (i.e., a diffraction grating), the "middle" wavebands used wavelength. Finally, high energy astronomers inherited their observing techniques from high energy physics, where the energy of the particle is the property best measured by the detectors, so these wavebands favour the use of energies. Regardless, all of these are equivalent and we must be fluent in translation across these boundaries.

LIGHT DETECTORS – Astronomers have used a range of detector systems over time. The nature of these detector systems provides important context for what follows in terms of conventions and definitions as we begin to quantify light. The most obvious detector system is the human eye. Eyes are amazing biological structures but they are pretty crappy scientific instruments. First, they have a limited range of wavelength sensitivity, running from 400 to 700 nm, which is the definition of the optical part of the spectrum. Since the Sun gives of most of its energy in this range, this turns out to be an appropriate evolutionary adaptation. The other 'feature' of the human eye is that its response to the brightness of light is logarithmic.

Lights with twice the flux don't appear twice as bright to humans. This logarithmic scaling is great for operating in variable light and dark conditions in our environment but it doesn't really help in making careful measurements. Finally, eyes are inefficient. Only about 4% of the photons that hit our eyes turn into neurological responses. For astronomy, we want to count all the photons we can get.

Astronomers developed detector systems for recording optical light. The first innovation was photographic film, which had an efficiency in responding to $\sim$ 15% of the photons that landed on the film. Later, astronomers used photomultiplier tubes, which could detect single photons but could only image one object at a time. More recently, the dominant detector used in optical astronomy is the Charged Coupled Device (CCD), a semiconductor chip that populates the conduction band with nearly perfect efficiency, putting one electron into the conduction band for every photon that hits the CCD.

Film, photomultiplier tubes, and CCDs are all sensitive to a broad range of wavelengths, responding nearly equally to light at 400 nm and at 700 nm. Essentially, all detectors are "greyscale" meaning they only detect the amount of light falling on the detector with little sensitivity to its colour. CCDs are highly efficient tools for taking images of extended fields of view. Nearly all modern images that you see in books and on the internet were collected using CCDs as the detector system. We discuss how you can use CCDs to take colour images in Section 1.2.

CCDs are primarily useful in the optical, UV, and X-ray. Infrared Light and lower frequencies require different detection technologies because the photon energies are not enough to excite photons across the band gap in normal semiconductors (typically 0.7 eV). Reviewing all the detector technologies used in astronomy is a bit beyond the scope of this text, but the fundamental theme remains: detectors typically only sense the amount of light falling on them. They are not usually also sensitive to the wavelength of that light.

The only other qualitative thing we need to know about light is that it has a *polarization* state described by the orientation of its electric field vector. Most astronomical light is unpolarized, which just means that there is a uniform distribution of light at all polarization angles. When we detect polarization, it is usually as a partial polarization, namely the strength of the light wave along one axis is larger than when measured along a perpendicular axis. The polarization is usually described in terms of the polarization fraction, which is the amount of this "excess" polarization in one direction compared to the intensity of the light as a whole.

Light is an incredibly versatile messenger and nearly everything we know about the Universe comes from photon messengers. Indeed,

we are a visually-focused species and our eyes are well adapted to see the light from the dominant source of photons in our corner of the Galaxy, namely the Sun. However, several other species are moving through space and we are just starting to see the Universe with the full range of possible tracers.

1.1.2 *Particles*

In addition to light, our study of physics has refined our view of matter and later antimatter. While a proton or an electron is a relatively ordinary thing in our world, receiving the same particles from space is a novel way to see the Universe. Generally, particle messengers that we receive from space must either be locally generated or nearly relativistic so they can cross interstellar (and intergalactic) distances in a reasonable amount of time. The dominant source of particles that are "locally generated" is the particle flux from the Sun called the *solar wind* , which is a magnetohydrodynamically accelerated flux of protons, electrons and helium nuclei (also called α-particles) from the surface of our nearest star. The solar wind has a typical speed of a few hundred km/s, which is fast but nowhere near the speed of light. Despite the apparently high speeds, our solar wind eventually thins out and is distrupted by the violent particle flows in interstellar space. The region of space dominated by the solar wind and its connected magnetic field is called the *heliosphere*, which acts as a shield against low energy particles reaching the inner solar system.

To quantify which particles reach the Earth from interstellar space, we compare the gyroradius of the particle in the solar magnetic field to the size scale of the solar system[2]:

$$r_g = \frac{\gamma m v_\perp}{qB}, \tag{1.1}$$

[2] This formula is derived by setting the Lorentz force law for magnetic fields equal to the force required to put a particle into uniform circular motion.

where γ is the Lorentz factor $\gamma = 1/\sqrt{1 - v^2/c^2}$, m is the particle mass, $v_\perp$ is formally the velocity component perpendicular to the magnetic field but we will take this just as the velocity, q is the particle charge and B is the solar magnetic field. At the location of the Earth, $B \approx 1$ nT and we can compare r_g to the scale of the solar system, for which a useful reference is 1 AU where an AU is an *astronomical unit*, i.e., the mean orbital separation between the Earth and the Sun. 1 AU = 1.49×10^{11} m. For $v_\perp = 10^5$ m s^{-1}, $\gamma \approx 1$ and $B = 10^{-9}$ T, a proton ($m_p \approx m_\mathrm{H} = 1.67 \times 10^{-27}$ kg) has $r_g = 10^6$ m and an electron has r_g 1830 times smaller ($\propto m_e/m_p = 1/1830$). These are both tiny compared to an AU so these low energy particles are strongly steered by the solar magnetic field. Indeed, for particles to reach the inner solar system without being strongly deflected, they need to have γ significantly larger than unity, for which $v \approx c$.

COSMIC RAYS – Several physical phenomena in the Universe act as particle accelerators which hurl particles at relativistic speeds across space, giving them the energy they need to traverse interstellar distances. These particles are broadly referred to as *cosmic rays*. Because of their higher masses, we tend to directly detect cosmic rays that are protons and other positive ions. Such particles are accelerated to $\gamma > 10$ which can push into the inner solar system, frequently colliding with the particles in our upper atmosphere and other times reaching down to the ground. These relativistic collisions in our upper atmosphere lead to a cascade of high energy particle physics yielding exotic particles like muons being produced in the high atmosphere and propagating down to the Earth's surface[3] We typically describe cosmic rays in terms of their total mass energy $E = \gamma mc^2$ where cosmic rays can have energy scales ranging from 1 GeV all the way up to > 1 YeV $= 10^{20}$ eV $= 16$ J. It's disturbing to find the Universe creates particles that have energies comparable to macroscopic physical objects: a tennis ball being being served at 72 km/h also has energy of 10^{20} eV.

[3] These muons are also moving at relativistic speeds and understanding relativity was essential to understanding how these short lifetime particles (half-life of 1.56 μs) could actually reach the surface of the Earth.

We also believe that there is a large population of cosmic ray electrons but these particles are only rarely detected here on Earth because of the low mass of the electron. Such a low mass implies the gyroradius is smaller than for protons and is more readily deflected by the solar magnetic field. We detect these particles through their radio frequency radiation. Charged particles gyrating in a magnetic field are accelerating and all accelerating charges radiate electromagnetic radiation. Our observations of this radiation can constrain the population of cosmic ray electrons.

Because of magnetic deflection, it is challenging to trace cosmic rays back to their origins except for the highest energy particles. However, we use the aggregate information about the population of detected cosmic rays to constrain the properties of their accelerating mechanisms and to estimate how these particles must heat the gas in interstellar space.

NEUTRINOS – More recently, particle physicists have turned their attention to detecting neutrinos. Neutrinos are a subatomic particle which only interacts with other particles through the weak nuclear force and gravity. Because the weak nuclear force is (wait for it) weak, neutrinos interactions with ordinary matter are rare. Thus, neutrinos can propagate through vast amounts of matter before they are likely to have an interaction. The properties of neutrinos are constrained but not completely known because of the challenge of detecting these elusive particles. The standard means of producing neutrinos in astrophysics is through weak reactions like a "beta" de-

cay, so named because early physicists classified this decay based on a "beta" particle which was later identified to also be the electron:

$$n_0 \quad \rightarrow \quad p^+ + e^- + \bar{\nu}_e \tag{1.2}$$

where n_0 is a neutron, p^+ and $\bar{\nu}_e$ is an electron antineutrino. Neutrinos come in three flavours corresponding to the three generations of leptons in the standard model of particle physics: electron, mu, and tau. We do not (yet) know what the mass of the neutrino is. However, thanks in part to the hard work of some of the physicists here at U. Alberta, we know that is does have a non-zero mass and that $m_\nu c^2 \lesssim 3$ eV.

Neutrinos are typically detected in particle physics experiments. The original neutrino detection events search for changes in the chemical content of large reservoirs of detection material. More recently, the most successful neutrino detection experiments have focused more on the detection of the radiation[4] created by the recoil of water molecules after being impacted by particles created through a high energy neutrino interacting with material in the detector. These faint flashes of light are detected by vast arrays of photomultiplier tubes. The hometown-hero here is the Sudbury Neutrino Observatory (SNO)[5], where a lot of U. Alberta faculty work on measuring the properties of neutrinos. Coincidentally, one of the major problems with neutrino detections are the cosmic rays discussed in the previous section. To gain shielding from cosmic rays, neutrino observatories are typically located in old mines under kilometres of rock, to gain shielding from high energy cosmic rays.

[4] Specifically, Cherenkov radiation

[5] https://en.wikipedia.org/wiki/Sudbury_Neutrino_Observatory

While this is all exciting, the utility of neutrinos as an astrophysical messenger comes from its low interaction probability with ordinary matter. This means that these neutrinos can probe parts of space for which light and ordinary particles would have to propagate through massive layers of matter to reach us. For example, neutrinos provide a direct probe of the nuclear fusion processes in the centre of the Sun, the mechanisms behind supernova explosions, and the matter near black holes. Unlike cosmic rays, they are not thought to be physically important in the regulation of aspects of galaxy evolution, though some exotic stellar processes depend on neutrino physics.

1.1.3 Gravitational Waves

The most recent addition to the astronomical toolkit is the first detection of gravitational waves by the Laser Interferometer Gravitational-Wave Observatory (LIGO) in 2015. This pair of observatories, now being joined by similar facilities across the globe, operates by measuring the gravitational "strain" induced by a passing distortion

in spacetime. These gravitational waves were predicted as a consequence of the General Theory of Relativity [Einstein, 1916], where accelerating masses stretch spacetime around them. Celestial objects orbiting each other are the simplest generators of gravitational waves, but only the most massive objects in fast orbits lead to detectable gravitational wave generation. Gravitational waves were known to exist indirectly through observations of massive objects in orbit around each other, but LIGO was the first facility to make a direct detection of the actual passing spacetime distortion. The first detection from LIGO came from a pair of (surprisingly) massive black holes (30 $M_\odot$)[6] orbiting around each other. Because gravitational waves carry energy away from the system, the orbital energy of the orbiting black holes decreases, leading to them falling in toward each other, speeding up, and radiating away gravitational wave radiation at a higher rate. This runaway process causes the black holes to spiral into each other and merge into a single larger black hole. These gravitational wave events reveal the mergers of massive objects.

[6] A solar mass, 1 $M_\odot = 1.99 \times 10^{30}$ kg.

While merging black holes definitely get points for the dramatic, the actual astrophysical consequences are relatively minor: two black holes become a larger black hole. There is little in the way of additional information from a black hole merge. In contrast, the major LIGO discovery with implications for astrophysics came in 2017 when LIGO observed the merger of two *neutron stars*. A neutron star is a remnant of a post-supernova massive star. Neutron stars have typical radii $R_{\rm NS} \approx 12$ km and masses of $M_{\rm NS} \approx 1.4\ M_\odot$, implying absurdly high mass densities. When two neutron stars spiral in and collide, they explode in an event called a *kilonova* and collapse into a black hole. [7] Returning to the 2017 observations, LIGO detected the inspiral and collision of the two neutron stars but electromagnetic observatories also detected radiation from the kilonova event, allowing a detailed characterization of the material created in a neutron star merger. This material turned out to be extremely rich in heavy elements, meaning nuclei heavier than iron. This finally explained the cosmic origin of these heavy elements, leading to a conclusion that these rare heavy elements (including lead, gold, platinum and other things you might actually possess) are mostly byproducts of neutron star mergers.

[7] Two of our professors, Craig Heinke and Rodrigo Fernández study neutron star properties and collisions in details. I got a lot of the details from this paragraph from listening to their research talks.

This result was the major result from *mulitmessenger astrophysics*, meaning the detection of more than one messenger species from an astrophysical sources. It was only the coincidence of the gravitational wave detection, confirming the neutron star merger, combined with the electromagnetic detection to characterize the kilonova material, that confirmed this idea.

Key Points

- The primary astronomical messenger is light but recently we have started using particles like neutrinos and more recently gravitational waves.
- Cosmic rays are relativistic particles accelerated by sources in the Milky Way galaxy. They are confined to the Galaxy by the magnetic field.
- Light is a wave described by its frequency, wavelength and energy. The electromagnetic spectrum is divided into wavebands summarized in Table 1.1.
- The energy of a photon is $E = h\nu$ and the wavelength and frequency are related by a dispersion relationship: $\lambda\nu = c$.

1.2 *Describing Light: Quantity*

While modern astrophysics is developing several new ideas around multimessenger work, most of the study of galactic astrophysics remains primarily oriented around interpreting the properties of light. To make this study more precise, we must start to quantify the radiation field we receive from objects.

1.2.1 *Flux and Magnitudes*

We will describe the light we receive from a source in terms of its *flux*, which we will give the variable f (and the literature will sometimes use F or occasionally S). The flux is the amount of energy received per unit area per unit time, implying units of W m^{-2}. If a detector has an area A, then the power received by the ideal detector would be $P = FA$. For example, the flux of solar radiation at the location of the Earth's orbit is $F_\odot = 1365$ W m^{-2} so that a detector with an area of 4 m^2 would collect $P = FA = 5$ kW of power. This is the motivation behind solar power: there is a lot of energy in sunlight.

We relate the flux received by a detector to the *luminosity*, L of a source, which is the total power emitted in electromagnetic radiation. We typically assume that sources are "point sources" meaning that we are sufficiently far away from them that they are well approximated as a single point, emitting radiation. The flux and luminosity are then related by the distance d in an inverse square law:

$$f = \frac{L}{4\pi d^2}. \tag{1.3}$$

Assuming that the light is radiating isotropically (the same in all

directions), the total amount of power (the luminosity) is always flowing through a spherical shell of radius equal to the distance d away from the source. A detector then collects radiation from a tiny fraction of that the total area of the spherical shell.

The flux and luminosity are good descriptors for point sources but they become less useful for extended objects, particularly those that are not spherical. This requires introducing additional descriptions of the radiation field, notably the (specific) intensity of the light. However, we will avoid these more in-depth approaches at this point but you should be aware that this brief treatment is not the end.

FLUX DENSITY – Describing light in terms of its flux remains incomplete in one key way, namely this does not describe how much light is received at one frequency (wavelength) vs another. This idea is particularly challenging because the frequency / wavelength is a continuous variable and if you want to quantify how much light has a frequency of exactly $\nu = 1.4000...$ GHz the answer will be zero. Instead, we use the idea of a *flux density*, for which we use the variable f_ν or f_λ to designate whether the density is a density per unit frequency or per unit wavelength. Since $E \propto \nu$, we do not also write down expressions for flux per unit energy but these are proportional to the expressions for frequency. The flux density is a density function for the flux, which means that it only makes sense to be integrated before it gives a measurable physical quantity. Specifically, the amount of light F received between frequencies ν_1 and ν_2 is

$$f = \int_{\nu_1}^{\nu_2} f_\nu(\nu) d\nu. \tag{1.4}$$

Like all integrals, the flux density can be well approximated by a rectangle sum over small intervals, e.g., ν_0 to $\nu_0 + \Delta\nu$ provided that $\Delta\nu \ll \nu_0$ so that $f = f_\nu(\nu_0)\Delta\nu$.

A QUICK EXAMPLE of flux density is a common radio astronomy unit named the *Jansky*, so named after the pioneer for using radio waves for astronomical observations. A Jansky is abbreviated as 1 Jy and has a definition of $1\ \text{Jy} = 10^{-26}\ \text{W m}^{-2}\ \text{Hz}^{-1}$. The (strong) radio source affectionately named 3C 48 has a flux density of 13 Jy at a frequency of 1.5 GHz (I've been observing this source a lot at the time of writing this book). If we want to know how much power an 26-m diameter circular antenna ($A = \pi D^2/4$) at the Very Large Array in New Mexico receives from 3C 48 at 1.5 GHz in a bandwidth of 1 MHz (10^6 Hz), we calculate this as $P = FA = F_\nu \Delta\nu A$ or:

$$P = 13\ \text{Jy} \left(\frac{10^{-26}\ \text{W m}^{-2}\ \text{Hz}^{-1}}{1\ \text{Jy}} \right) (10^6\ \text{Hz}) [\frac{\pi}{4} (26\ \text{m})^2] = 6.9 \times 10^{-17}\ \text{W}. \tag{1.5}$$

This is a very tiny number, which tells you how sensitive radio telescopes need to be to detect sources.

$F_\nu \neq F_\lambda$ – In working with flux density units, we need to keep track of whether the power is specified per unit wavelength or per unit frequency. Because of the inverse relationship $\lambda = c/\nu$, there isn't a simple scaling between the flux densities expressed in these two domains. For a frequency interval that is equivalent to a wavelength interval, i.e. from ν_1 to ν_2 so that $\lambda_1 = c/\nu_1$ and $\lambda_2 = c/\nu_2$, the power per unit area in that interval must be the same. This gives us route to change variables and express, e.g., f_λ in terms of a given f_ν

$$\begin{aligned}
\int_{\lambda_1}^{\lambda_2} f_\lambda d\lambda &= \int_{\lambda_1}^{\lambda_2} f_\nu \frac{d\nu}{d\lambda} d\lambda, \\
&= \int_{\lambda_1}^{\lambda_2} f_\nu \frac{d}{d\lambda}\left(\frac{c}{\lambda}\right) d\lambda, \\
&= \int_{\lambda_1}^{\lambda_2} f_\nu \left(-\frac{c}{\lambda^2}\right) d\lambda, \\
&= \int_{\lambda_2}^{\lambda_1} f_\nu \left(\frac{c}{\lambda^2}\right) d\lambda
\end{aligned}$$

where in the last step, we have used the negative sign to switch the bounds of integration, which is a statement that frequency and wavelength are increasing in opposite directions. Taking this switching of bounds as implicit, we often write

$$f_\lambda = f_\nu \frac{c}{\lambda^2} = f_\nu \frac{\nu^2}{c} \tag{1.6}$$

to convert between these two representations of flux density.

Specific Luminosity – Following a similar pattern as flux density, we also introduce the *specific luminosity*, which is the luminosity of a source per unit of frequency or wavelength, usually denoted L_ν or L_λ. The inverse square law holds here so that $L_\nu = 4\pi d^2 f_\nu$. The term "specific" is commonly used to indicate that this is a density function. The specific luminosity can be integrated following a similar set of conventions as the flux density and the conversion between wavelength and frequency is the same. The specific luminosity is sometimes also called the *monochromatic luminosity*.

Magnitudes – The final method for characterizing light that we will need to be aware of is the *magnitude* system, which is arguably the worst astronomical convention that was ever established. However, the use of magnitudes is so pervasive that we cannot engage with the astronomical literature if we don't learn what a magnitude is.

The origins of the magnitude system with ancient Greek astronomers who catalogued the stars in the sky according to brightness, where the brightest stars were first magnitude, the next brightest stars were second magnitude, and so on until the faintest stars that were visible to humans were sixth magnitude. Cataloguing is a great approach but this had two main deficits to haunt us. First, the magnitude system is backwards: small numbers on the magnitude scale correspond to brighter sources. Second: the human eye, which was the instrument used by the Greek astronomers, has the aforementioned logarithmic response to brightness of light.

The heritage of observational astronomy then opted to make sure that the quantitative measurement of astronomical source brightness remained compatible with measurements back to the Greeks. To make the magnitude system into a basis for quantitative measurements, astronomers adopted a convention that a difference of five magnitudes corresponded to a factor a 100 in flux. To compare two sources of fluxes f_1 and f_2, the corresponding magnitudes of the sources m_1 and m_2 are related:

$$m_1 - m_2 = -2.5 \log_{10} \left(\frac{f_1}{f_2} \right) . \tag{1.7}$$

This is the fundamental equation of the magnitude system. If $f_1 = 100 f_2$, the log of the ratio evaluates to 2 and then we get that $m_1 = m_2 - 5$, i.e., the "five magnitudes is a difference of a factor of 100 in brightness." The negative sign expresses the backward nature of the magnitude system. This also means that a difference of a single magnitude corresponds to a flux difference corresponding to a factor of $100^{1/5} = 2.512$.

The major nightmare of the magnitude system is that it is not a single system but rather the magnitudes depend strongly on the telescope and detector system being used. The heart of this difficulty is because flux density cannot be uniquely measured at a single frequency / wavelength. Instead, an actual measurement requires integrating over the spectrum and the details of this integral, namely the spectrum being observed and the interval over which the integration happens, lead to differences in the system. Typically telescope systems work to have excellent internal calibration, so that measurements of different objects with the same telescope are well measured. Once the telescope needs to set an absolute calibration, the accuracy of the measurements is typically much lower.

Astronomers needed to set a zero point of the magnitude system to relate measured fluxes to magnitudes. For most of modern astronomy, this was selected so that the star Vega had a magnitude of $m = 0$ irrespective of how it was measured. The problem with this approach is that Vega is actually a fairly complicated source and

doesn't have a good measurements across all the wavebands. Recently, astronomers have begun to move to a new system called the AB magnitude system where there is an absolute flux scale introduced.

$$m_{\mathrm{AB}} = -2.5 \log_{10} \left(\frac{f_\nu}{3631 \text{ Jy}} \right). \tag{1.8}$$

Finally, these magnitudes are often called *apparent* magnitudes to distinguish them from the absolute magnitudes discussed next.

Absolute Magnitude is the equivalent to the concept of luminosity in the magnitude system. The absolute magnitude of an object is defined as the magnitude that would be observed for a source if it were located at a distance of 10 parsecs (abbreviated pc) from the observer, where $1 \text{ pc} = 3.09 \times 10^{16}$ m. Parsecs will be motivated and defined in section 1.3, but for now, it is just a useful astronomical distance unit. To understand the effects of distance, we need to apply the inverse square law to the flux of a source (f) to find out the equivalent flux for a source observed at 10 pc, denoted f_{10}.

$$\begin{aligned} f_{10} &= \frac{L}{4\pi (10 \text{ pc})^2} \\ f &= \frac{L}{4\pi d^2} \end{aligned} \tag{1.9}$$

Defining the magnitude corresponding to f_{10} as the absolute magnitude M, we have

$$m - M = -2.5 \log_{10} \left(\frac{f}{f_{10}} \right) = -2.5 \log_{10} \left(\frac{(10 \text{ pc})^2}{d^2} \right) = 5 \log_{10} \left(\frac{d}{10 \text{ pc}} \right) \tag{1.10}$$

where the last equality follows by the rules of logarithms. We can solve this equation for M for our definition of absolute magnitude:

$$M = m - 5 \log_{10} \left(\frac{d}{10 \text{ pc}} \right) = m - 5 \log_{10} (d_{\mathrm{pc}}) + 5 \tag{1.11}$$

where d_{pc} is the distance expressed in units of pc.

The expression $m - M$ in equation 1.10 is called the *distance modulus* because it only depends on the distance to the source. We usually use the distance modulus as shorthand for distance to the source and many measurements of distance in astronomy naturally yield magnitude units and are well expressed as distance moduli.

Spectroscopy – The next quantification of lights we need to introduce the measure of the flux density. Measuring the flux density as a function of wavelength / frequency produces a *spectrum* of a source. Figure 1.1 shows a spectrum of Vega from the optical and NUV part

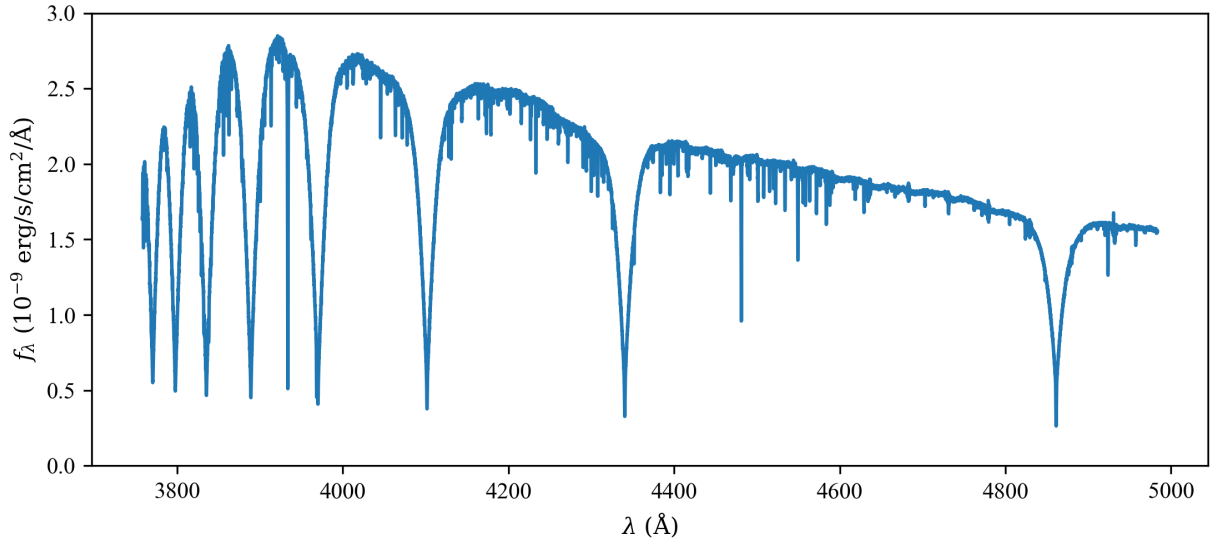


Figure 1.1: Spectrum of Vega taken by the European Southern Observatory's Very Large Telescope.

of the spectrum. Recall that Vega is the zero point of several versions of the optical magnitude system. The big dips in the spectrum are the spectral "lines" of hydrogen in the Balmer series with the rightmost feature being Hβ corresponding to the $n = 4 \to 2$ transition of hydrogen at $\lambda = 4861$ Å. This is a weird unit, and we should pay attention these units so we can interpret some figures from astronomy.

The units in Figure 1.1 use the "centimetre-gram-second" (cgs) set of units combined with the units of Ångströms (abbreviated Å), which aren't really part of any standard unit system. Astronomers love cgs for reasons that I don't really understand. The table at the right summarizes some of the key non-trivial cgs units equivalencies in SI units. As a word of warning, cgs-based electromagnetism is especially weird because the units of charge aren't equivalent to Coulombs.

Unit	SI equivalent
1 erg	10^{-7} J
1 Gauss	10^{-4} T
1 Å	10^{-10} m = 0.1 nm

This spectrum clearly contains a huge amount of information, so you may wonder why astronomers don't use spectroscopy exclusively in studying space. The answer lies in the colour-blind nature of CCD detectors. A spectrum like the above feature uses optical elements like diffraction gratings and lenses to take the light from a single object (e.g., Vega) and "disperse" it spatially, spreading it out in one or two dimensions where the location that the light travels depends on its wavelength. Thus, the spatial dimensions on the detector are used to keep track of the wavelength of the light rather than location from which the light originated, as you would in an image. Hence, spectroscopy can usually only be carried for one or a few objects at at time, requiring a relatively large amount of time to

collect data on an object, but the information you receive has a lot of detail. Even so, we would really like a shortcut.

Key Points

- We quantify the amount of light using several different measures including:
  - Luminosity – power emitted by a source.
  - Flux – power received per unit area.
  - Flux density – power received per unit area per unit wavelength or frequency.
  - Specific luminosity – Power emitted per unit frequency or wavelength.
  - Magnitudes – A logarithmic relative brightness scale that astronomers use to the brightness of two sources.
- Flux and luminosity are related by the inverse square law (Equation 1.3.
- The flux density can be expressed per unit wavelength (f_λ) or per unit frequency (f_ν). These are related by Equation 1.6.
- Equation 1.7 defines the relationship between magnitudes and the fluxes (or flux densities over equivalent bands) of sources.
- The distance modulus (Equation 1.10) incorporates the inverse square law into the magnitude system and is just a measure of distance.
- Spectra of sources plot the flux density of sources as a function of wavelength or frequency. Spectral features arise from quantum transitions in atoms and molecules.

1.2.2 *Filters, SEDs and Colours*

Filters provide the spectroscopic shortcut we are looking for since three lines ago. Since CCD cameras can only detect the intensity of light, the use of *filters* makes a detector colour sensitive by only allowing certain colours of light to fall on the detector. A filter is just an optical element that transmits light on a certain range of wavelengths and blocks the transmission of all light outside of that limited range. When an observer inserts a filter in the optical path of the

light, the detector system becomes colour sensitive.

Figure 1.3 shows the *transmission* of an ideal filter, which illustrates the fraction of light that the filter transmits as a function of wavelength. This ideal filter curve can be expressed as a function $T(\lambda)$, which would have the form:

$$T(\lambda) = \begin{cases} 1; & \lambda_1 < \lambda < \lambda_2 \\ 0; & \text{otherise} \end{cases}, \tag{1.12}$$

transmitting light on the wavelength range $\lambda_1 < \lambda < \lambda_2$.

To model the effect of filters on detector systems, we imagine a detector like a CCD which detects light with a wide range of different wavelengths. If that detector perfectly responded to all the light from a source then we could relate the flux observed by a telescope (f_{obs}) to the flux density of the object:

$$f_{\text{obs}} = \int_0^\infty f_\lambda(\lambda)\, d\lambda, \tag{1.13}$$

where the bounds of integration are just the mathematical representation of the "responds to all light" assumption in our ideal detector. If we then put the ideal filter (Equation 1.12) in the optical path of a detector and measure the flux received from an object, we could relate this power to the flux density of the source over a narrow range

$$f_{\text{obs}} = \int_0^\infty T(\lambda) f_\lambda(\lambda)\, d\lambda. \tag{1.14}$$

Since the filter transmission curve sets the integrand to zero anywhere outside the range of $\lambda_1 < \lambda < \lambda_2$, we can change bounds to this equivalent integral.

$$f_{\text{obs}} = \int_{\lambda_1}^{\lambda_2} T(\lambda) f_\lambda(\lambda)\, d\lambda = \int_{\lambda_1}^{\lambda_2} f_\lambda(\lambda)\, d\lambda. \tag{1.15}$$

The latter equality follows because $T(\lambda) = 1$ in this wavelength range. This integral is the essence of why filters are useful: they turn the broad sensitivity of a CCD detector into one with a limited range of sensitivity. Furthermore, you can switch the filters (relatively cheap) and keep the same CCD camera (expensive).

Over time, astronomy has developed several different "standard" filters. Figure 1.4 shows the transmission curves for some of the most common filters in astronomy. Each filter is given a specific name corresponding to its transmission curve. For example, the Johnson-Cousins B filter transmits light near 420 nm in the blue, and is often called the B "band"[8]. Comparing these curves to the ideal filter (Figure 1.3) shows the limitations of our manufacturing technologies. The earlier filter systems, like the Johnson-Cousins filters, shows substantial overlap between adjacent bands and few filters get close to a

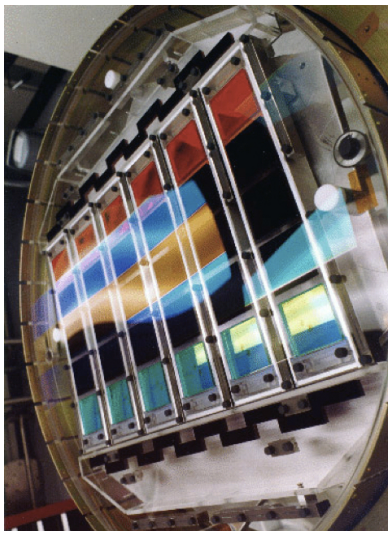


Figure 1.2: The camera system used for the Sloan Digital Sky Survey, a major survey of galaxy evolution. The system consisted of four rows of CCD cameras each with five columns corresponding to the different filters used in the survey. The colours of the filters can be seen in the image.

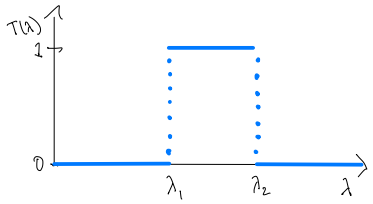


Figure 1.3: An idealized filter transmission curve.

[8] This can be agonizng because there are also letter names for frequency bands in the radio part of the spectrum.

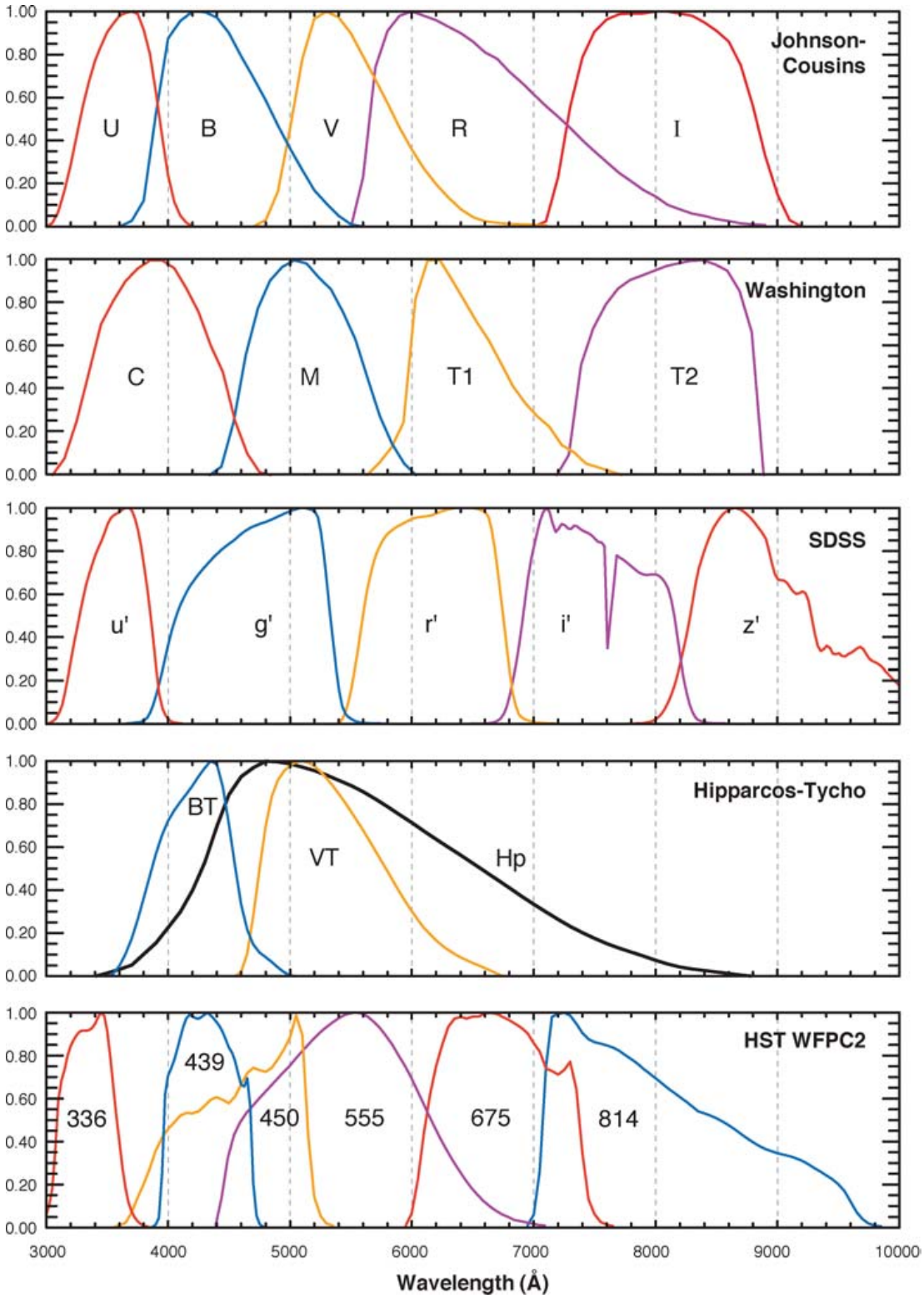


Figure 1.4: Commonly used filter sets in optical astronomy plotted as transmission (from 0 to 1) as a function of wavelength from Bessell [2005]. The curves show the myriad of different bands used in astronomy and their names. The different rows show the various "systems" of filters.

transmission of 100%. Using the actual transmission curves in Equation 1.14 leads to sensitivity over a wider range of the spectrum than the ideal curves.

SPECTRAL ENERGY DISTRIBUTIONS – The use of filters and CCDs is incredibly useful in the optical, UV and NIR and similar methods exist across the full electromagnetic spectrum. In studying galaxies, it is extremely useful to examine how the light from a galaxy changes from the radio up to the X-ray. This necessarily requires observing with several different telescopes for the different wavebands. When these measurements are aggregated into a single place, the resulting data are called a *spectral energy distribution*, abbreviated SED.

Figure 1.5 shows an example SED of the radio-bright galaxy helpfully named 3C138 based on its number in the Third Cambridge Catalogue of radio sources. This figure aggregates flux density data from the radio where it well studied, infrared and optical observations as well as some data from X-ray missions. The bright radio flux is the signature of the accretion of matter onto the central supermassive black hole in the galaxy. SEDs typically represent all the flux density over the full galaxy being studied, but for nearby systems, they may only refer to the flux density from a part of a galaxy like its nucleus. If you look at a graph like the one at the right and see data at two different y-values for the same x value, usually the difference is because the SED is showing data from the whole galaxy as well as a part of the galaxy. We will use SEDs later in the book as we try to understand how the total light of the galaxy reflects the physical processes happening inside the system.

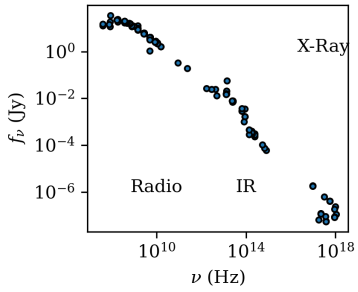


Figure 1.5: SED for the radio-bright galaxy 3C138 using data aggregated by the NASA Extragalactic Database.

COLOUR INDICES – The final connection to make is to link up the filters and measurements of flux to the magnitude system used for optical astronomers. This combination leads to the creation of a *colour index*. Specifically, a colour index[9] is the difference between apparent magnitudes when the flux is measured in two different filters. One of the classic colour indices in the study of galaxies is the difference between the magnitudes measured in the Sloan Digital Sky Survey (SDSS) g' band and the r' band. This difference is denoted with the shorthand $g' - r'$.

[9] sometimes just called a "colour"

$$g' - r' = -2.5 \log_{10} \left(\frac{f_{\text{obs},g'}}{f_{\text{obs},r'}} \right) = -2.5 \log_{10} \left(\frac{\int_0^\infty T_{g'}(\lambda) f_\lambda(\lambda)\, d\lambda}{\int_0^\infty T_{r'}(\lambda) f_\lambda(\lambda)\, d\lambda} \right). \tag{1.16}$$

This form of integral is called a colour index because it can numerically represent the actual colour of an optical object with a numerical scale. It is worth reasoning through how the math for this works out.

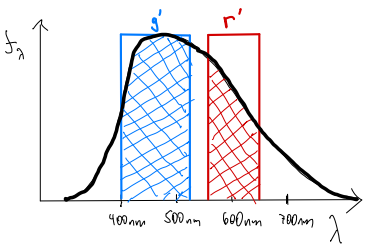


Figure 1.6: Example spectrum (black curve) with the wavelength ranges for the SDSS g' and r' filters indicated.

Figure 1.6 shows an example spectrum with the wavelength ranges for the g' and r' bands indicated. The use of filters essentially measures the flux in each of these wavelength ranges, which are indicated by the filled in portions of the spectrum, blue corresponding to the section of the spectrum that contributes to the flux in the g' band and red indicating the portion fo the spectrum that contributes to the flux measurement in the r' band. Very approximately, we can estimate that the flux that would be measured in the g' band would be about $1.5\times$ larger than the flux in the r' band (i.e., $f_{g'} = 1.5 f_{r'}$). In this case,

$$g' - r' = -2.5 \log_{10} \left(\frac{1.5 f_{r'}}{f_{r'}} \right) = -0.44. \tag{1.17}$$

If, a different spectrum showed $f_{g'} < f_{r'}$ then the fraction inside the log would be less than 1, and the log would be negative, leading to a positive value for the colour index after multiplying by -2.5. I find it useful to remember that small (including negative) values of a colour means that the source is brighter in the colour band on the left of the difference (g' in the above example). Colour indices that are positive indicate that the source is relatively brighter in the band that is on the right of the difference (i.e., the r' as above). Thus a colour index is a quick way of summarize the shape of a spectrum. Colour indices are complementary to SEDs and are frequently used in the properties of stars and stellar populations since stars emit a lot of light in the optical.

Key Points

- Filters and bands help measure the amount of light in small sections of the EM spectrum.
- The standard names of filters define bands and refer to standard colours in the optical spectrum. The most common filter sets are the Johnson-Cousins ($UBVRI$) and Sloan ($ugriz$) where the sequences of letters run from short to long wavelength (Figure 1.4).
- Spectral Energy Distributions (SED) plot the amount of light across several different wavebands.
- An example of a colour indices is given in Equation 1.16. Colour indices are given on the magnitude scales and indicate the ratio of light in different bands. In general in a $X - Y$ colour, more negative values indicate the light in band X is bright relative to Y. More positive values indicate Y is brighter than X.

1.2.3 *Blackbody Radiation*

To give context to all these measures of fluxes, spectra and SEDs, astronomers use a *blackbody* spectrum for interpreting the physical processes at work in an object. A blackbody is a theoretical object that acts as a perfect absorber, which then re-emits its radiation with a characteristic spectral profile. Blackbody radiation is sometimes also called *thermal radiation*. Deriving this profile was a triumph of early quantum mechanics leading to the introduction of the now-ubiquitous Planck constant. Here, we use it for its physical properties. A nearly blackbody spectrum develops for any object that is opaque. The flux density for a spherical blackbody of radius R viewed from a distance d away is:

$$f_{\nu,\mathrm{bb}} = \frac{\pi R^2}{d^2} B_\nu(\nu, T) = \frac{\pi R^2}{d^2} \frac{2h\nu^3}{c^2} \frac{1}{\exp\left(\frac{h\nu}{kT}\right) - 1}. \tag{1.18}$$

Here, $B_\nu(\nu, T)$ is variable used for the reference spectrum and the $\pi R^2/d^2$ term converts the reference spectrum into a flux density. In the above equation, k is the Boltzmann constant ($k = 1.38 \times 10^{-23}$ J/K), h is Planck's constant ($h = 6.63 \times 10^{-34}$ J s), ν is the frequency of the radiation and T is the temperature of the source. This is the expression for a blackbody flux density per unit frequency (i.e., has units of W/m^2/Hz) and the expression of power per unit wavelength is different because of Equation 1.6:

$$f_{\lambda,\mathrm{bb}} = \frac{\pi R^2}{d^2} B_\lambda(\lambda, T) = \frac{\pi R^2}{d^2} \frac{2hc^2}{\lambda^5} \frac{1}{\exp\left(\frac{hc}{\lambda kT}\right) - 1}. \tag{1.19}$$

We won't dive into all the minutiae of the blackbody spectrum here, but there are a few useful properties to be reminded of. First, blackbody spectra are functions of two parameters: the temperature of the source and the wavelength / frequency. All else being equal a hotter blackbody will be brighter at all wavelengths / frequencies than a cooler one. Blackbody spectra have a single maximum at a wavelength (λ_{max}) or frequency (ν_{max}) that is set by the temperature of the source as described by the Wien Law:

$$\lambda_{\mathrm{max}} = \frac{0.002898 \text{ m K}}{T} \tag{1.20}$$

$$\nu_{\mathrm{max}} = (5.879 \times 10^{10} \text{ Hz K}^{-1}) T \tag{1.21}$$

If we integrate the spectra in either Equation 1.18 or 1.19, we also arrive at another useful expression, namely the Stefan-Boltzmann law that gives the total luminosity of a source with radius R and temperature T:

$$L = 4\pi R^2 \sigma_{\mathrm{SB}} T^4 \tag{1.22}$$

where $\sigma_{SB} = 5.67 \times 10^{-8}$ W m^{-2} K^{-4}. Finally applying the inverse square law to this luminosity equation gives the total flux of the blackbody source a distance d away:

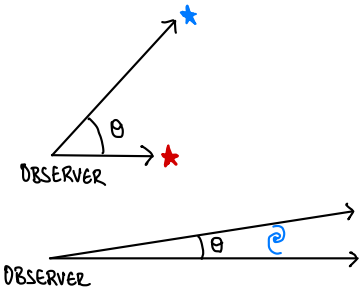
$$f = \frac{R^2 \sigma_{SB} T^4}{d^2}. \tag{1.23}$$

Key Points

- Blackbody radiation, also known as thermal radiation, is characterized by a specific spectral shape (Equation1.19 or 1.18) and is a function of temperature and distance from the source.
- A blackbody spectrum has a single maximum defined by the Wien law (Equation 1.21) and an integral over wavelength / frequency given by the Stefan-Boltzmann law (Equation 1.22).

1.3 *Describing Light: Direction*

Where the previous section laid out how much light was coming from an astronomical source, we'll also need to know a bit about where those sources are on the sky. Again, there are too many astronomical conventions for describing coordinate systems, and the details are a deep pit of tedious details we really want to avoid. However, we will need to be aware of a few conventions so we can look at actual astronomical data and understand what the coordinate axes mean.



Measuring in Angles – Astronomers measure separations on the sky in units of angle. A difference in angle is the angle between two rays (in the geometry sense), starting at the observer and projecting out, one to intersect each point.

The figure at the right illustrates the angle θ between the two rays, demonstrated in two different cases, but both illustrate the angle between two points in the sky. In the top case, the rays can go to different objects and different distances (here to a red star and a blue star). In the bottom case, the rays can go to either side of an extended object, like the poorly drawn galaxy. In the latter case, we say that the galaxy "subtends" an angle θ on the sky. In astronomy, we usually measure angles in degrees and follow a sexagesimal (base-60) convention of dividing 1° into 60 arcminutes (written 60′) and then dividing each arcminute into 60 arcseconds (i.e., $1' = 60''$). Thus, $1^\circ = 60' = 3600''$.

1.3.1 *Coordinate Systems*

Equatorial Coordinates – We establish coordinate systems so it is easier to measure the angles between two points on the sky. To do this, we have established the concept of the celestial sphere, a hypothetical spherical shell concentric with the Earth, and we chart out the locations of all objects on the sky onto this celestial sphere. We can then divide the sphere up into latitude- and longitude-like coordinate systems, where one example is sketched out in Figure 1.7.

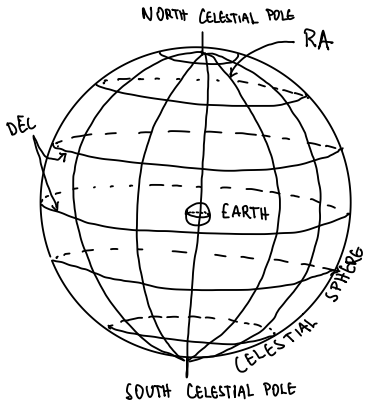


Figure 1.7: Sketch of the Equatorial Coordinate system.

This system is called *equatorial coordinates*, and it charts out the positions on a sky as if it were a sphere, linked to the orientation of the Earth. The longitude-like coordinate is the *right ascension* (denoted RA or α) and the latitude-like coordinate is the *declination* (Dec or δ). Like latitude and longitude, the RA and Dec are not measured in linear distance but in angles or equivalent. The declination is the easiest place to start since this is like latitude. Similar to lines of latitude on the Earth, there is an equal unit of arc between each line of declination, whereas lines of right ascension converge at the poles. In contrast, right ascension is just crazy, but all for good reasons. RA is also an angular coordinate but it is measured in units of time (?!). This apparently insane convention is because of the Earth's rotation relative to the stars. If you are actually steering a telescope, the sky is rotating at a rate where 24 hours corresponds to 360° of rotation, so this convention becomes helpful. Thus, 24 h = 360°.[10]

[10] Furthermore, this time is actually measured in units of *sidereal* time rather than solar time, so the 24 hours in this circle correspond to 23 hours and 56 minutes of our civil time. Again, really useful if you want to actually steer a telescope, but not actually important right now.

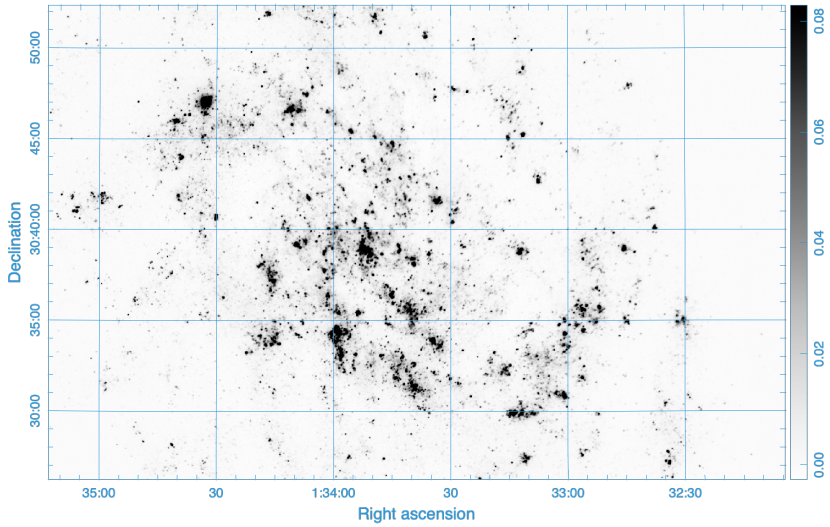


Figure 1.8: Example astronomical image showing a nearby galaxy (M33) in FUV light from the GALEX mission. The axes show the equatorial coordinates for the image.

The equatorial coordinates come in handy for interpreting the axes of images like the one shown in Figure 1.8. This figure shows the

map of flux density in the FUV from a nearby galaxy. Astronomical convention is to usually express declination in terms of degrees-minutes-seconds as the axes in Figure 1.8 are labelled. For example, the coordinate "30:40:00" is 30 degrees, 40 arcminutes, 0 arcseconds, sometimes written as $30^\circ\,40'\,00''$. The right ascension units of the x-axis in Figure 1.8 are measured in units of time and have coordinates like "01:34:00," which is read 1 hour, 34 minutes, 0 seconds and is written $1^\mathrm{h}\,34^\mathrm{m}\,00^\mathrm{s}$. Here, 1 hour = 60 minutes = 3600 seconds (i.e., not "arcseconds" just "seconds"). RA can also be expressed in units degrees where $1^\mathrm{h} = 15^\circ$, and this is usually just written as a decimal. This is just a change of unit and not a change of coordinate system. Note that the x-axis is increasing in the "wrong" direction: to the left instead of to the right as we'd normally expect. This is because our coordinate system is being viewed from the inside of the celestial sphere rather than the outside like the surface of the Earth. Finally, as a longitude-like coordinate, lines of constant right ascension converge as you approach the pole of the coordinate systems. Hence, an interval of RA is not a constant unit of angle like it is for Dec. Given all this craziness, you need to remember one thing: **if you use an axis to measure the angle subtended by a source, reference your measurement to the declination axis not the right ascension.**

Galactic Coordinates – The Galactic coordinate system is the other main coordinate system we will use, since it is particularly useful for determining the locations of objects in our own Galaxy. Galactic coordinates have the coordinates axes "galactic latitude" and "galactic longitude" and follow the same basic rules of latitude and longitude as on the surface of the Earth. Galactic longitude usually uses the variable ℓ and galactic latitude uses the variable b. The origin of the galactic coordinates ($\ell = 0^\circ, b = 0^\circ$) points toward the Galactic[11] centre so that disk of the galaxy (i.e., the Milky Way that we see on the sky at night), falls along the galactic equator $b = 0^\circ$. The Sun's orbit around the centre of the galaxy is (approximately) in the direction $\ell = 90^\circ, b = 0^\circ$. Figure 1.9 shows a top down schematic of the Milky Way galaxy with the galactic longitude system indicated, showing that $\ell = 0^\circ$ does indeed point toward the Galactic centre. The figure to the right shows a schematic of the galactic longitude and latitude systems when viewed from the Earth. Again, the longitude-like coordinate ℓ increases "backwards", i.e., to the left, because we are inside the celestial sphere.

[11] When capitalized, "Galaxy" and "Galactic" refer to our Milky Way Galaxy.

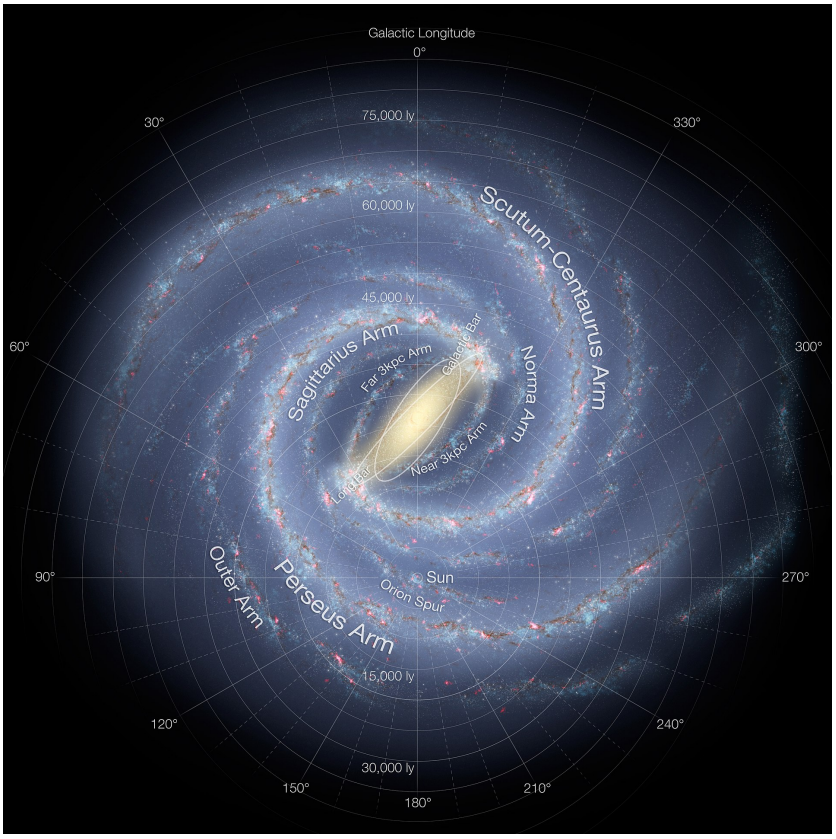
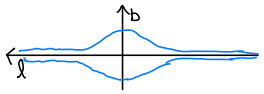


Figure 1.9: Top down visualization of our Milky Way galaxy with galactic longitude coordinates indicated. Image credit: NASA/JPL-Caltech/ESO/R. Hurt.

The Galactic coordinate system is tilted with respect to the Equatorial coordinates. They both represent positions on the Celestial sphere but the North Galactic Pole and the North Celestial pole are not located at the same place. There are some fancy coordinate transforms that can translate coordinates in one system to coordinates in the other. Again, this is a bit beyond the scope of what we need, but these transformations and the details of coordinate systems that we are blithely sweeping under the rug are essential parts of actually conducting observations. The figure at right illustrates the same image as shown before but with lines from Galactic coordinates instead of equatorial.



1.3.2 *Motion on the Sky*

Because of the vast distances to other stars and galaxies, they initially appear to be at fixed positions on the celestial sphere. This makes it meaningful to give the coordinates of a star. For example, I can specify that the star Vega is located at $\alpha = 18^{\mathrm{h}}36^{\mathrm{m}}56.33635^{\mathrm{s}}, \delta = +38^{\circ}47'01.2802''$ which tells us where the star is located with respect other objects in the equatorial coordinate system. Those coordinates should be fixed to be useful. However, everything in space is moving with respect to all the other objects, meaning that these fixed coordinates will not remain fixed for long. While the speeds at which objects are moving are large, the distances that they must travel are also vast so that the motion of the skies plays out on timescales that are long compared to human lives. Nonetheless, we can see some motions of the stars and interpret what they mean.

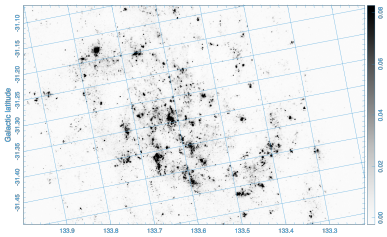


Figure 1.10: As per Figure 1.8 but with Galactic instead of equatorial coordinates.

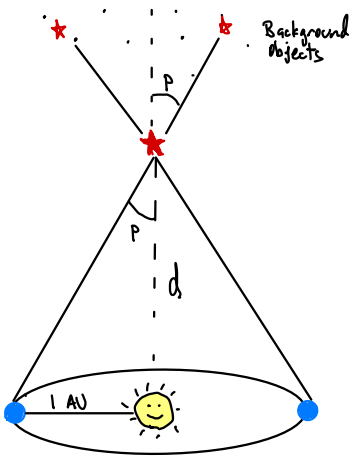


Figure 1.11: Sketch of parallax.

Parallax – Relatively nearby objects appear to move with respect to the celestial coordinate systems because of *parallax*. The parallax of an object is the apparent motion with respect to the coordinate system due to the Earth's motion around the Sun. The celestial coordinate systems are now established with respect to a set of highly luminous sources that are found a substantial fraction of the way across the visible Universe. These distant reference points give us a essentially stationary frame of reference against which we can see nearby objects move because of this geometric effect. The effect is illustrated in Figure 1.11, where the red star appears projected at different locations over the course of the year. We can measure the angle between the two apparent positions with respect to the coordinate system and we define *half* that angle as the parallax angle, p.

If we make careful measurements, we can determine the parallax angle p, which allows us to use simple geometry to measure the

distance to the red star. This is exciting! As we will see, measuring distances in astronomy is challenging but essential to understand the nature of astronomical sources. Parallax distances are the highest quality distance measurements that we can make and are a cornerstone of astrophysics. Parallax works using the right triangle made with legs of the distance to the star d and the separation between the Earth and the Sun. The Earth-Sun separation is well measured and its mean value is defined as 1 *astronomical unit* or 1 AU $= 1.49 \times 10^{11}$ m. Thus,

$$\tan p = \frac{1\text{ AU}}{d} \Longrightarrow d = \frac{1\text{ AU}}{\tan p} \approx \frac{1\text{ AU}}{p}, \tag{1.24}$$

where the final approximation holds using the small-angle approximation for the tangent function, namely that $\tan p \approx p$ for $p \ll 1$ **and p is measured in radians!**

The nearest stars have parallax of $\sim 1''$ making arcseconds a useful reference unit. Since 1 rad $= 206,265''$, we dimensionlize the parallax formula by measuring parallaxes in arcseconds.

$$d = \frac{206,265\text{ AU}}{(p/1'')} \equiv \frac{1\text{ pc}}{(p/1'')}. \tag{1.25}$$

Here, we have defined 206,265 AU as 1 parsec (abbreviated pc) so that, 1 pc $= 3.09 \times 10^{16}$ m. In words, an object that is 1 pc away will have a parallax angle of $1''$. Also, note the inverse relationship: more distant objects have smaller parallax angles. An object that is 2 pc away has $p = 0.5''$.

One final subtlety: parallax causes a star's position to trace out an elliptical path on the celestial sphere. The semimajor axis of this ellipse is oriented in the RA direction and has a size of p and the semiminor axis of this ellipse has size of $p \sin\beta$ where β is the ecliptic latitude, basically the angle between the Earth-Sun orbital plane and the star. The exact angle definition isn't important, only that a star traces an elliptical path with size that depends on where it is in the sky.

PROPER MOTION – The other motion of stars with respect to the celestial coordinate system is called *proper motion* and arises because the stars are really moving with respect to each other. The motion of the stars with respect to the Sun leads to this proper motion and is called such to distinguish it from the parallax motion. Proper motions are typically measured in scales of arcseconds per year with Barnard's Star being a nearby star renown for having a large proper motion of $|\boldsymbol{\mu}| = 10.3''\ \text{yr}^{-1}$, where $\boldsymbol{\mu}$ is the standard variable for the proper motion, represented as a vector since it has two components along the axes of the coordinate system.

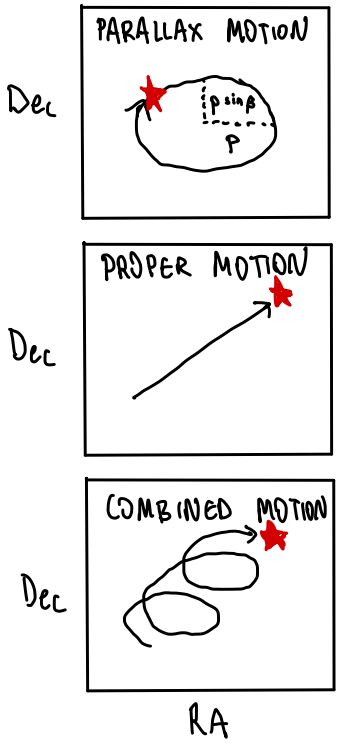


Figure 1.12: The panels illustrate the apparent motion of a star due to parallax (top), proper motion (middle) and combination of parallax and proper motion (bottom).

Parallax and proper motion can be of comparable sizes and the combined action of these two motions makes stars trace out helical paths on the sky. By fitting these helices, we can infer both the proper motion and the parallax. Figure 1.12 illustrates how these two motions combine to give the motion of a star with respect to the celestial coordinate systems.

Proper motion only manifests if stars are moving perpendicular to an observer's line of sight here on the Earth. If a star is moving on a line toward or away from Earth, it would manifest no proper motion but would show a significant Doppler shift.

DOPPLER EFFECT – The Doppler effect is the change in the observed wavelength (frequency) of light when a source of light is moving with respect to an observer. Since some of the light from objects will be emitted at known wavelengths (e.g., the spectral lines of atoms), we can measure the apparent wavelength of the light and use this to infer the component of the velocity of the source in the radial direction, i.e., along the line of sight. In terms of wavelengths,

$$\frac{v_r}{c} = \frac{\lambda_{\rm obs} - \lambda_{\rm rest}}{\lambda_{\rm rest}}, \tag{1.26}$$

where v_r is the radial velocity, c is the speed of light, $\lambda_{\rm obs}$ is the observed shifted wavelength, and $\lambda_{\rm rest}$ is the wavelength that would be observed if the source were at rest. In terms of rest ($\nu_{\rm rest}$) and observed ($\nu_{\rm obs}$) frequencies, we have:

$$\frac{v_r}{c} = \frac{\nu_{\rm rest} - \nu_{\rm obs}}{\nu_{\rm rest}}. \tag{1.27}$$

The careful reader will note that these two expressions aren't actually equivalent, i.e., you cannot substitute $\nu = c/\lambda$ into the first Doppler equation and derive the second. There will be a small-order that leads to a small error of magnitude $(v_r/c)^2$, which for most speeds that we will care about will be insignificant. In the cases where $v_r \lesssim c$, then we need the a relativistic treatment. It will only be important for a few applications in this class, but the full relativistic treatment of the Doppler effect for radial motion gives:

$$v_r = c\frac{\nu_{\rm rest}^2 - \nu_{\rm obs}^2}{\nu_{\rm rest}^2 + \nu_{\rm obs}^2}. \tag{1.28}$$

Note that even this expression is incomplete as it ignores the effects of time dilation from transverse motion, which we will again call 'beyond the scope' and blithely forge ahead.

The final thing to pay attention to in the Doppler formula is the sign convention. This acts in the sense of a standard spherical-polar coordinate system where a positive velocity indicates that the size of

the radius vector is getting larger, i.e., the distance between the object and the observer is increasing and a negative sign indicates that it is decreasing.

OBJECTS IN 6D – With all the information given above, we now have the capacity to measure the positions of objects in a six-dimensional phase space: three spatial dimensions and the three components of the velocity vector. The three spatial dimensions from the coordinates on the sky and the distance to an object, usually measured with its parallax. The coordinates on the sky are the angular part of a spherical polar coordinate system and the distance forms the radial vector. We will make use of this coordinate system primarily in the form of Galactic coordinates, where we can transform into a Cartesian coordinate system with the spherical polar transformation, switching the trigonometric function on the polar angle because b is defined to be zero at the equator rather than at the pole as it usually is in math problems.

$$x_{\rm gal} = d\cos b\cos\ell \tag{1.29}$$
$$y_{\rm gal} = d\cos b\sin\ell \tag{1.30}$$
$$z_{\rm gal} = d\sin b. \tag{1.31}$$

With this, we can map out the three dimensional space distribution for stars where we measure parallax.

We can also find all the components of a velocity vector, where Figure 1.13 shows the decomposition of the velocity vector relative to the line of sight. By measuring a proper motion vector, we can determine the proper motion in the direction perpendicular to the line of sight. We can transform that to true velocity by using the small angle formula and writing $v_\perp = d|\boldsymbol{\mu}|$ but this formula only works when we express the proper motion in units of an angular speed, i.e., radians per second. Scaling this to the units we typically use (arcseconds per year), we arrive at the following scaled equation:

$$\left(\frac{v_\perp}{\text{km s}^{-1}}\right) = 4.74\left(\frac{d}{1\text{ pc}}\right)\left(\frac{|\boldsymbol{\mu}|}{1\,''/\text{yr}}\right). \tag{1.32}$$

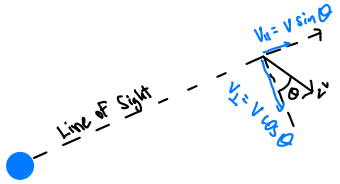


Figure 1.13: Measuring components of a velocity vector relative to the line of sight (dashed).

The parallel component of the velocity $v_\parallel$ in Figure 1.13 is measured by the Doppler formula and naturally gives measurements in linear velocity units rather than angular units. This gives us the third component of the velocity vector which then needs to be transformed into the same Cartesian coordinate system as for the spatial dimensions if we want to combine these values. Without belabouring

a bunch of coordinate transforms, we arrive at:

$$\begin{aligned} v_{x,\mathrm{gal}} &= -\mu_\ell d \cos b \sin \ell - \mu_b d \cos \ell \sin b + v_r \cos \ell \cos b \\ v_{y,\mathrm{gal}} &= \mu_\ell d \cos b \cos \ell - \mu_b d \sin \ell \sin b + v_r \sin \ell \cos b \\ v_{z,\mathrm{gal}} &= \mu_b d \cos b + v_r \sin b \end{aligned}$$

The important point in the above formula set isn't the full form of all those equations but instead it is the observation that we can measure every component in them if we know: (1) the coordinates of an object, (2) its parallax, (3) its proper motion, and (4) its radial velocity leading to a full six-dimensional representation of where objects are.

Projected Distances – For extended objects, we can use angle subtended on the sky, the distance to the object, and some trigonometry to measure the size of the object when measured "on the plane of the sky"[12] In nearly all cases, we use a helpful approximation that the angle is small, again meaning that $\theta \ll 1$ rad. In this case, the projected size of the object is just $s = d\theta$ where θ is the angle subtended expressed in radians, d is the distance to the object and s is the physical size of the object.

[12] This "plane" is the tangent plane to the celestial sphere at the location of the centre of the object.

Key Points

- We measure position the celestial sphere using spherical polar coordinates describing the location of an object on a celestial sphere. The two most common coordinate systems are the Equatorial system (RA/Dec) and the Galactic system (galactic longitude and latitude).
- These systems are spherical systems and the longitude-like lines converge toward the poles of the system. Thus, intervals of a fixed longitude angle change in size with latitude and it is important to use the latitude-like coordinate to measure the angular size of the object.
- Parallax is a primary method to measure distance to objects (Equation 1.24) and manifests as the apparent change of an object's position on the celestial sphere because of the Earth's motion around the Sun 1.11.
- The motion of objects with respect to the Sun causes objects to show a *proper motion*, which refers to continuous changes over time in the object's position on the celestial sphere. Proper motions, when combined with parallax, measure the true space velocity of objects perpendicular to the line of sight.

- The Doppler shift measures the motion of objects parallel to the line of sight (Equations 1.26 and 1.27).
- We measure the projected sizes or distances between objects on the celestial sphere using the small angle formula: $s = d\theta$ for θ in radians. Even though this is just one paragraph in the book, this is a very important technique in solving problems.

1.4 *Measuring Light: Telescope Detector Systems*

The preceding content all represents a framework for representing observations of astronomical objects and how to interpret these observations in terms of physics to deduce the intrinsic properties of object (e.g., the luminosity and the projected size). However, these observations are all made using telescope systems, and the limitations of those systems shape what is measurable. If you make observations, these limitations will shape the scientific information that you can infer, and it is critically important to think about what objects are not in your data because they would never be detectable given a telescope. A truism (cliché?) for astronomical observations is that "absence of evidence is not evidence of absence."

Angular Resolution – The optical structure of the telescope (meaning the mirrors and lenses that focus the light in a telescope) impose a limit on the size of the sources that you can detect. This limit is the *angular resolution* of the telescope, which is the minimum angular separation that two sources can have and still be distinguished as individual sources. This limit is set by the size of the primary optical element that focuses the light coming into the telescope. Physically this limit arises from wave interference: light for a point source is focused at the centre of an image where waves hit all over the primary mirror come together to interfere constructively. At small shifts from the centre of the source in the image, these waves are still nearly perfectly in phase so the image will still be bright with light from the source. Instead, for the waves to cancel out, there needs to be a shift of half a wavelength in path difference between light from one part of the mirror compared to another part. This interference will arise first for parts of the mirror that are located on opposite sides of the primary since their path lengths will show the largest differences. The formula for the angular resolution is given by the Rayleigh criterion, which is the mathematical formulation of the

above wall of text. This criterion is

$$\theta_{\text{res}} \approx \frac{\lambda}{D} \tag{1.33}$$

where θ_{res} is the angular resolution, λ is the wavelength of the observations, and D is size of the optical system used to collect the light. For a telescope with a mirror, the size D is the diameter of the primary mirror, but for something like a radio interferometer, D is the farthest separation between the antennas in the array (see Figure 1.14). Note that this equation is just an approximation. There is sometimes a coefficient in front of the λ/D term that depends on the shape of the primary mirror. For circular mirrors, this coefficient is 1.22.

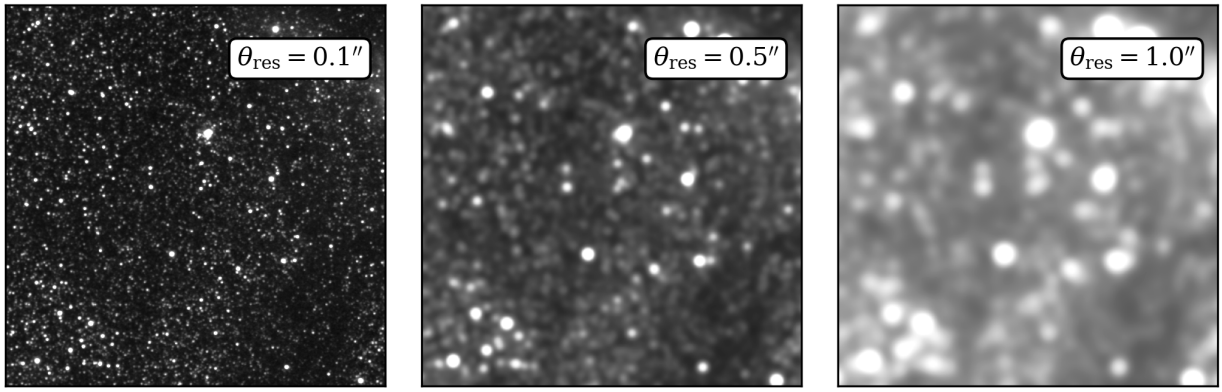


Figure 1.15: Example of an image with three different angular resolutions.

In general, lower values of θ_{res} are better. Figure 1.15 shows stars in a nearby galaxy (M33 again) image with the Hubble Space Telescope in the F435W filter (this is very nearly the same as the one labelled 439 the bottom panel of Figure 1.4). The panel on the left shows the "native" resolution of the image for HST which is $0.1''$. The middle and right panels show the same field at $0.5''$ resolution and $1.0''$ resolution. A high quality site on the ground could achieve the $0.5''$ and typical ground-based observatory conditions are usually about $1.0''$ resolution. Note the practical effects of angular resolution on what we can recover from the images: pairs of nearby stars blend into single blurs, in particular faint stars cannot be distinguished near bright stars. The background stars blend together to form a background emission which places limits on the sensitivity of the image. Only stars that are bright with respect to this background can be distinguished as individual objects.

As implied above, the resolution of ground based optical telescopes is limited to $\gtrsim 0.5''$ because of turbulence in the atmosphere even though the Rayleigh criterion would imply a much better angular resolution (i.e., $\theta_{\rm res} \ll 0.5''$) These turbulent motions stir up the air, compressing it slightly and distorting the light ray. Modern telescopes often use *adaptive optics* systems to compensate for the blurring from the atmosphere, pushing the angular resolution closer to the Rayleigh limit for the telescope. Radio telescopes do not suffer nearly as much from atmospheric turbulence but are limited instead by the long wavelengths of the light which makes their angular resolution intrinsically poorer than the shorter-wavelength UV/optical/IR. Based on this logic, X-ray and gamma ray telescopes should have even better resolutions, but these facilities run into a different problems. At these high energies, there are only a few photons, which behave much more like particles than waves. High energy photons have a nasty habit of just blasting through optical elements like mirrors and lenses unless they have a very high angle of incidence. This inability to focus effectively means that X-ray telescopes also have a resolution that is $\lesssim 1''$.

SENSITIVITY – The sensitivity of a telescope is the other major factor that makes telescope observations non-ideal. In general the sensitivity of the telescope derives from the size of its primary reflector and from the instrument to carry out the photometry or spectroscopy. In general, the instruments used on telescopes are of comparable quality to each other, but the size of the primary can vary substantially. The telescope's *light gathering power* a function of the area of its primary mirror since this is what catches the flux from sources and leads to a total amount of detected power. Larger telescopes mean more power is detected. A given type instrument (e.g., radio receiver) can detect a fixed amount of power, so larger telescopes can detect fainter sources because they capture more power from the same source. The full set of sensitivity calculations is affected by a myriad of details including the atmospheric conditions, the quality of the optics of a telescope, and the type of instrument being used. However, all else being equal, bigger telescopes are more sensitive.

Detecting signals from astrophysical sources is a combination of the properties of the instrument that's being used along with the amount of time spent gathering light from the source. Ultimately, astronomers want a system that has a good base sensitivity (large collecting area, low noise detectors). This is because, for most systems, the sensitivity of a telescope only improve proportional to the square root of the time invested. Collecting light for $4\times$ longer observations only improves sensitivity by a factor of $\sqrt{4} = 2$. To observe to sig-

nificantly deeper levels ($10\times \rightarrow 100\times$) thus requires a substantially larger investment of telescope time ($10^2 \rightarrow 10^4$).

Key Points

- The key parameters of a telescope are wavelength coverage, angular resolution (Equation 1.33), and sensitivity.

2
Stars

The theory of stellar evolution is the centrepiece of modern astrophysics. Compared to the rest of the field, we understand stars extremely well when compared to galaxies, planets, and the Universe as a whole. Stars are sufficiently well understood and well behaved (meaning their initial conditions translate to a known set of outcomes) that we can use their properties as a measuring device to understand the remainder of systems. In this chapter, we will first provide a minimal introduction of stars, treating them as black boxes. In the next Chapter, we will then discuss how groups of stars (populations) behave in aggregate since stars are the major source of light in galaxies. We have two courses (ASTRO 320 and ASTRO 465) dedicated to delivering a satisfying understanding of stellar structure and evolution. Here, we will steal the best parts of those courses that are needed to understand galaxy evolution.

2.1 A Primer on Stars and Stellar Evolution

This section is meant to give a plausible set of physical explanations describing the driving processes in how stars evolve, but it cannot be complete. Unfortunately, given the direction we are headed, we will have to accept some of these things as true and defer their clear demonstration to other classes.

To frame this discussion, note that most stars are found on the *main sequence*, which is defined to be the long phase of a star's life (about 90% of the stellar lifetime) where it is fusing hydrogen into helium in the core. Before and after the main sequence, the star will evolve and be powered by different energy sources operating in different parts of the star. Our Sun is on the main sequence and we use this stage of life as the "standard" state for describing stars.

2.1.1 *The Essential Properties of Stars*

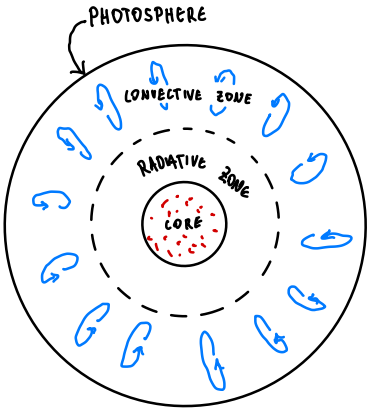


Figure 2.1: Simplified structure of a $M \sim 1\ M_\odot$ star like our Sun.

STELLAR ANATOMY – Stars are (nearly) spherical, self-gravitating bodies of gas that generate energy through nuclear fusion at some point in their lives. Stars range in mass from $0.08 < M/M_\odot \lesssim 300$. Figure 2.1 illustrates the basic structure of a star like the Sun with $M \sim 1\ M_\odot$. The nuclear fusion occurs in the star's *core* but this region is surrounded by an *envelope* of material that is not actively fusing material. The envelope consists of a *radiative zone* where energy is transported outward through radiative diffusion. Outside of that layer, there is a *convective zone* where large convective cycles transport energy from the base of the convective layer to the outer parts of the star. Only the outer surface of the star, called the *photosphere* is visible to the outside. The photosphere is the boundary where radiation flows from the interior out into space. This anatomy can vary for different stars. For example, stars with $M > 2\ M_\odot$ have convective cores and radiative outer layers. Stars with $M < 0.25\ M_\odot$ are convective throughout their interiors.

The star formation process also forms objects with mass $M < 0.08 M_\odot$, which never ignite nuclear fusion of hydrogen in their cores. These objects are called *brown dwarfs*. While cooler than stars, brown dwarfs still undergo deuterium fusion, converting rare isotopes $^2\mathrm{H} + {}^3\mathrm{He} \rightarrow {}^4\mathrm{He} + p^+ + \text{Energy}$. These rare isotope give the brown dwarfs some power, making them luminous in the infrared. However, they cannot ignite "standard" hydrogen fusion. Deuterium fusion happens in objects with $M > 0.013\ M_\odot$, which is only $13\times$ the mass of Jupiter. Objects with $M < 0.013\ M_\odot$ are just planets. It is not yet known whether the star formation process is capable of forming planets independent of stars or brown dwarfs.

The upper limit for stellar masses appears to be set by the maximum mass stable objects that can form. Objects forming with masses higher than this approximate limit undergo instabilities and fragment into smaller objects or self-destruct before arriving at a stable configuration. Early in the Universe, the maximum mass for stars is likely higher because there are fewer heavy elements, which means that material is less affected by opacity from these heavy elements.

THE FUNDAMENTAL "THEOREM" OF STELLAR EVOLUTION – The triumph of stellar evolution is to take physical inputs to make a robust prediction of how a given star will evolve. The important initial properties are, in order of the magnitude of influence on the star:

1. The initial stellar mass
2. The initial chemical composition of the star (its "metallicity")

3. Whether or not a star is in a binary or multiple star system (more next Chapter)

4. How fast the star is initially rotating

By knowing these inputs, in particular the initial mass and metallicity, it is possible to predict the long term internal structure, observable properties (radius, surface temperature, luminosity, and mass), and final product of a star throughout its lifetime. The final products of stars are called stellar remnants and include *white dwarfs*, *neutron stars*, and *black holes*.

The mass is a straight forward measurement but the *metallicity* of a star is more subtle. In a convention that makes chemists die a little inside, astronomers refer to every element except hydrogen and helium as a "metal." But chemistry is a reality and thus metallicity comes with a set of conventions in astronomy to account for different chemical variations between the elements. As an introduction, we follow standard stellar astrophysics by considering the chemical composition of a star in terms of its mass fraction, using the variables X, Y, and Z to indicate the mass fractions of hydrogen, helium, and everything else respectively. In this definition, $X + Y + Z \equiv 1$ and have typical values for sun-like stars as $X = 0.72$, $Y = 0.26$, $Z = 0.02$. Metallicity is a deeper subject in the context of galaxy evolution so we return to an expanded discussion of chemical composition in Section 2.9.

The physics of a star is primary set by (1) stars are nearly spherical bodies of matter supported against their self-gravity by pressure, and (2) that stars are in approximate steady state. The spherical shape of stars indicates that their structure is dominated by the balance between self-gravity and pressure because these forces are isotropic, i.e., spherically symmetric. The pressure can be provided from one of three sources (1) ordinary gas pressure, (2) radiation pressure, or (3) *degeneracy pressure*. The "nearly" in front of spherical indicates that forces like magnetism or stellar rotation are not significant factors in the force balance for a star though they can be present.

The self-gravity of a star is the dominant physics for its life. From a highly abstract view, stellar evolution is just what happens as matter resists gravitational collapse. For a star of mass M and radius R *something* must provide a source of pressure with magnitude

$$P_{\text{required}} \sim \frac{GM^2}{R^4}, \tag{2.1}$$

or else the star will collapse rapidly. The collapse time is 30 minutes for the case of our Sun. Stars are thus said to be in *hydrostatic equilib-*

rium, indicating a balance between the pressure (gradient) in a star and its self gravity.

We can estimate the required pressure to support a star by considering a very simple model for a spherical star of mass M and radius R as shown in Figure 2.2. We consider the star as two hemispheres of mass $M/2$ exerting pressure on a midplane surface of contact with $A = \pi R^2$ (shaded in the figure). Each hemisphere exerts a gravitational force of magnitude $F_g = G(M/2)^2/R^2$ on the surface where we have taken the separation between the centres of mass for the two hemispheres as $\approx R$ (actually $3R/4$ if you love calculus but that $\approx$ conceals a myriad of sins). The pressure on the midplane is then $2F_g/A$ or

$$P \approx 2\left(\frac{GM^2}{4R^2}\right)\frac{1}{\pi R^2} \sim \frac{GM^2}{R^4}. \tag{2.2}$$

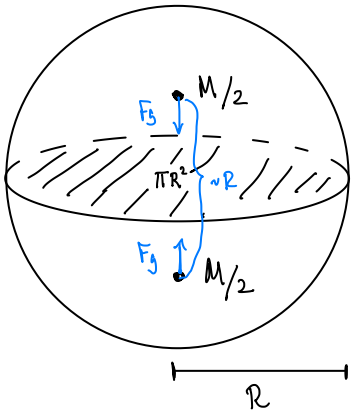


Figure 2.2: Schematic figure for estimating the pressure in the centre of a star.

Pressures – Ordinary stars, like the Sun, have their source of pressure from gas pressure that follows the perfect gas law. In this class, we will use a gas law of the form:

$$P_{\rm gas} = nkT, \tag{2.3}$$

where P is the pressure, n is the volume density of particles in the gas, k is Boltzmann's constant and T is the gas temperature. This is called the *equation of state* for an ordinary gas and holds whether the gas is chemically neutral or ionized into a plasma. This is equivalent to all the other gas laws you've seen but this is the most convenient for our class.

Since radiation can carry momentum, it can provide a force and thus a pressure. In the interiors of high mass stars, this means that radiation pressure becomes significant and the equation of state for radiation is:

$$P_{\rm rad} = \frac{4}{3}\frac{\sigma_{\rm SB}T^4}{c} \tag{2.4}$$

Here $\sigma_{\rm SB}$ is the Stefan-Boltzmann constant, c is the speed of light and T is the temperature.

While we won't use it much in this class, for completeness, the other important equation of state for stars is the equation of state for electron degeneracy pressure:

$$P_{\rm deg} = K_1\left(\frac{\rho}{\rm kg\ m^{-3}}\right)^{5/3} \tag{2.5}$$

where $K_1 = 3 \times 10^6$ Pa for ρ is the mass density measured in SI units. Degeneracy presure arises when the density of matter gets so high that forcing it to smaller volumes and higher densities would cause the material to violate the Pauli exclusion principle for fermions.

Hence, the fermions can only be compressed to higher density if they have higher momentum, leading to an increase in pressure. This spooky quantum mechanical pressure is important for the remnants at the end of a stars life and in some stages of stellar evolution. There is a parallel pressure that arises for neutrons called "neutron degeneracy pressure" which has a similar form, but different constant.

ENERGY SOURCES – For most stars, the pressure is provided by gas pressure with some possible contribution from radiation pressure in high mass stars. Both of these equations of state depend on the temperature of the gas. This temperature is significantly higher than the environment. For example, if we equate the required pressure with the gas pressure, we can derive the average internal temperature of star, taking the number density of particles as $M/(R^3 m_{\rm H})$ where $m_{\rm H}$ is the hydrogen mass:

$$\begin{aligned} \frac{GM^2}{R^4} &\sim \frac{M}{R^3 m_H} kT \\ T &\sim \frac{GMm_H}{kR} \\ T &\sim 2 \times 10^7 \text{ K} \end{aligned} \tag{2.6}$$

where we have substituted in solar values for the last estimate: $M = 1\ M_\odot = 1.99 \times 10^{30}$ kg and $R = 1\ R_\odot = 6.96 \times 10^8$ m. This value is about a factor of 5 larger than the actual mean temperature of the Sun because of the coarse assumptions we have made (we assumed the gas is neutral and not ionized and the simple estimate of the pressure). Even with this factor-of-5 correction, this is a big value, but the key point is that it is much larger than the mean temperature of empty space: $T = 2.73$ K. This means that stars are out of thermal equilibrium with their environments and are radiating away a lot of power. Since they are in steady state, the star must be producing energy to balance the power loss from the surface of a star.

We don't see 10^7 K temperatures from stars because we only see the surface layers of the stars. The surface temperature of the Sun is $T_{\rm eff} = 5777$ K. This surface temperature is often called the *effective temperature* since it is the temperature that you plug into the Stefan-Boltzmann equation (Equation 1.22) to get the right relationship between luminosity and radius. The outer layers of a star act as insulating layers much like a cozy jacket keeps you from radiating away too much warmth into the Edmonton winter.

During most stages of a star's life, the required energy is provided through nuclear fusion. Fusion happens when light atomic nuclei fuse together to form larger nuclei. The mass of the combined nucleus is smaller than the sum of the reactants, and the mass difference

is converted to energy following $E = \Delta mc^2$. We will adopt it as given but the key physics is how the strong nuclear force binds together atomic nuclei.

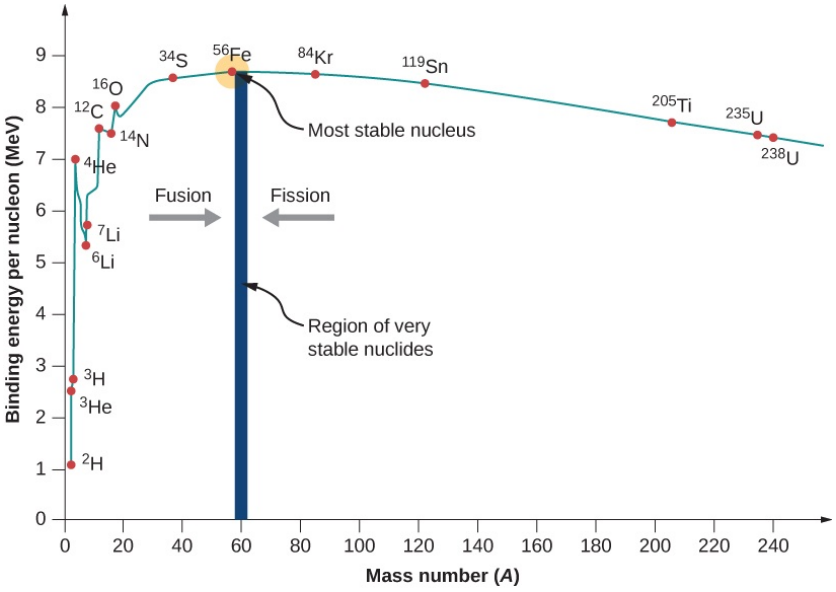


Figure 2.3: The binding energy per nucleon in an atomic nucleus as a function of atomic mass number A. From Physics LibreTexts under a CC-SA-NC-3.0 license.

The key result of the nuclear physics is summarized in Figure 2.3. This figure shows the binding energy per nucleon (i.e., protons or neutrons) as a function of atomic mass number (the total number of protons and neutrons in a nucleus). This formulation of the nuclear energy is particularly useful since it allows us to make large-scale estimates of the energy produced through fusion. For example, the binding energy per nucleon of deuterium, i.e., ^{2}H consisting of a proton bound to a neutron is about 1 MeV per nucleon. Thus, the energy produced by fusing 1 kg of protons with 1 kg of neutrons is the total number of nucleons involved in the reaction ($\mathcal{N}$) multiplied by the binding energy per nucleon. We calculate an approximate value of $\mathcal{N}$ by assuming that the nucleon mass is equal to a hydrogen mass, so

$$E = (2\text{ kg})\left(\frac{1\text{ nucleon}}{m_{\text{H}}}\right)(1\text{ MeV/nucleon}) = 1.2\times10^{27}\text{ MeV} = 1.9\times10^{14}\text{ J}. \tag{2.7}$$

For context, this is about 10 seconds of global human energy consumption.

Examining Figure 2.3 shows that fusion liberates energy up to $A = 56$ corresponding to the stable ^{56}Fe nucleus. After that, fusion is endothermic and the fission of nuclei is exothermic. Fission is our current energy source and one of the major goals of modern physics

is to provide stable nuclear fusion power. So far, we've learned how to engage in *unstable* nuclear fusion generation in nuclear weapons. The Figure also shows that most of the binding energy is released in the step from $4 \times {}^{1}\mathrm{H} \rightarrow {}^{4}\mathrm{He}$ (0 to 7 MeV per nucleon) and that later generations of exothermic fusion to, e.g., ${}^{12}\mathrm{C}$, ${}^{16}\mathrm{O}$, and ${}^{20}\mathrm{Ne}$ only release a relatively small amount of energy. This later observation is equivalent to the statement that the H→He fusion step in a stars life – the main sequence – is the longest phase since it provides the most efficient form of fusion energy generation.

Following our derivation for the average internal temperature of a star (Equation 2.6), we see that the temperature is inversely proportional to the radius, so that as an object gets smaller, it will also get hotter. Recall also that, in the absence of any source of pressure support, the system will collapse quickly. However, if there is no source of energy keeping the interior of the star warm, the star will cool off. This process is more gradual, with a characteristic time (called the Kelvin-Helmholtz timescale) being $t_{\mathrm{cool}} = E_T/L$ where E_T is the thermal energy of the star and L is the star's luminosity. This timescale is approximately 30 Myr for a star like the Sun.

These two relations tell us what happens when a star runs out of nuclear fuel. The core cools off on t_{cool}. It thus contracts and heats up and then one of two things will happen:

1. The star will reach the threshold temperatures and densities require to ignite another stage of nuclear fusion. For example, the temperatures could reach the $T = 10^8$ K temperatures needed to start the fusion of He into C under the 3α fusion process.[1]

2. Alternatively, the core could become supported by electron degeneracy pressure. This source of pressure will provided the necessary support to keep the star in a hydrostatic equilibrium without nuclear fusion and the collapse will halt and the star will lose energy.

[1] It seems like two ${}^{4}\mathrm{He}$ particles should fuse to form ${}^{8}\mathrm{Be}$ but this is actually endothermic reaction (see Figure 2.3), so He needs to fuse directly to ${}^{12}\mathrm{C}$.

Key Points

- Stars are defined by something that, at some point in its life, fused hydrogen into helium.
- Stars are spherical and described in two main parts: the core where nuclear fusion is taking place and the envelope, which is not producing energy but energy is flowing through to escape into space at their surface.
- Stars have initial mass ranges from $0.08 < M/M_{\odot} \lesssim 300$.
- Most stars have hydrogen mass fractions of $X \sim 0.72$ and helium mass fractions of $X \sim 0.26$. The metallicity of a star

means everything by hydrogen and helium and can range from $0 \lesssim Z \lesssim 0.02$. Thus sun has $Z_\odot = 0.02$.

- Stars are in hydrostatic equilibrium where the force of self-gravity is balanced by a pressure gradients inside the star. In main sequence stars, most of this pressure is provided by gas pressure (Equation 2.3), but radiation pressure (Equation 2.4) and degeneracy pressure (Equation 2.5) can also be important at some stages of a star's life.
- Stars produce energy by nuclear fusion of light elements into heavy elements. The binding energy per nucleon (Figure 2.3) shows how much energy can be liberated per unit mass for each stage of nuclear "burning". Most of a star's energy from fusion comes from H to He fusion.
- Stars evolve when they run out of the different stages of nuclear fusion.

2.2 *Observed Stellar Properties*

Combining the above physics in detail develops the full models of stellar astrophysics, which is amazing but we will reduce to a bit of a "black box" where we just focus on the net properties of stars as they influence their environments.

Mass-Luminosity Relationship – The most important relationship we obtain from modelling stellar interiors is the mass-luminosity relationship, which gives a fitting function to represent the output luminosity (L) of a star given an input mass (M), while ignoring small variations due to stellar metallicity.

Figure 2.4 shows the result of modelling the structure and evolution of stars using the MESA program, which you would use in ASTRO 465. Very approximately, stars fit the relationship

$$L = 1\ L_\odot \left(\frac{M}{M_\odot}\right)^{3.5}. \tag{2.8}$$

shown in the Figure as the "simple" relationship. This is a coarse approximation useful for quick scaling estimates. In this class, we will sometimes use a more refined "composite" version that is approximately valid for stars on the main sequence, reflecting the different

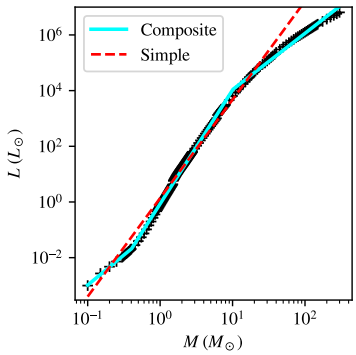


Figure 2.4: Mass-Luminosity scaling for zero-age main sequence stars derived from the MIST models of Dotter [2016].

transitions in dominant physics in stars.

$$\frac{L}{L_\odot} = \begin{cases} 0.16(M/M_\odot)^{2.1} & ; \quad M/M_\odot < 0.4 \\ 1.00(M/M_\odot)^{4.1} & ; \quad 0.4 < M/M_\odot < 10 \\ 126.(M/M_\odot)^{2.0} & ; \quad 10 < M/M_\odot \end{cases} \tag{2.9}$$

Both these relationships get across the basic point: more massive stars are *much* more luminous. Using the naive scaling relationship that $L \propto M^{3.5}$, we would estimate that a 10 $M_\odot$ star has a luminosity of $L \sim 3200\ L_\odot$ and the composite relationship would give $L \sim 12,600\ L_\odot$.

Given this mass-luminosity scaling, and the properties of nuclear fusion discussed in the previous section, we can make an estimate of how long stars spend on the main sequence. The Sun's core contains about $f_c = 13\%$ of the total mass and the mass fraction $X = 0.72$ of that indicates how much hydrogen is available for nuclear fusion. Each hydrogen proton is a single nucleon with mass $\approx m_\mathrm{H}$ and, in fusing to helium would yield 7 MeV of energy based on Figure 2.3. Thus, the total nuclear energy available in the Sun during its main sequence lifetime, $E_{\mathrm{nuc},\odot}$,

$$\begin{aligned} E_{\mathrm{nuc},\odot} &= \frac{f_c X M_\odot}{m_H}\left(\frac{7\ \mathrm{MeV}}{\mathrm{nucleon}}\right) \\ &= \frac{0.13 \cdot 0.72 \cdot 1.99 \times 10^{30}\ \mathrm{kg}}{1.67 \times 10^{-27}\ \mathrm{kg}}\left(\frac{7\ \mathrm{MeV}}{\mathrm{nucleon}}\right) \\ &= 7.8 \times 10^{56}\ \mathrm{MeV} = 1.3 \times 10^{44}\ \mathrm{J}. \end{aligned}$$

This nuclear energy reservoir is being consumed at a rate given by the stellar luminosity, so the time to run out of fuel is $E_{\mathrm{nuc},\odot}/L_\odot$, which we call the main sequence lifetime, $\tau_{\mathrm{MS},\odot}$. Substituting in constants gives $\tau_{\mathrm{MS}} = 1.0 \times 10^{10}$ yr $= 10$ Gyr.

Using this as a reference scale, we can refer again to the mass-luminosity relationship and determine the main sequence lifetimes for stars of different mass. Using the very coarse relationship (Equation 2.8), we can estimate the main sequence lifetime for a star of mass M to be:

$$\tau_{\mathrm{MS}} = \tau_{\mathrm{MS},\odot}\left(\frac{M}{M_\odot}\right)\left(\frac{L}{L_\odot}\right)^{-1} = 10^{10}\ \mathrm{yr}\left(\frac{M}{M_\odot}\right)^{-2.5} \tag{2.10}$$

The upshot of this equation is that high mass stars have very short main sequence lives and evolve relatively quickly. Given Equation 2.9 and the composite scaling, a $M = 100\ M_\odot$ star has a lifetime of 1 Myr. This ends up being a little shorter than the full model which takes into account the full physics and mass loss. The shortest stellar lifetime is about 3 Myr. These are all really long time periods, but for context, 2 Myr is about the time that humans have been on Earth

and is comparable to a massive star's lifetime. In contrast, Earth has been around for 4.5 Gyr with life on it for the last 3.5-4.0 Gyr, which is comparable to a medium mass star's lifetime.

It is worth noting that the lifetime for stars with $M < 0.9\ M_\odot$ are longer than the age of the Universe: 14 Gyr. This means that any star formed with this mass will not have evolved off the main sequence.

SPECTRAL TYPES – The other major observable properties of stars is their spectrum. For example, Figure 1.1 shows the spectrum of the star Vega. Without observing the entire spectrum, we see several notable features in the stellar spectrum. These are spectral lines formed in the atmosphere of the star above the photosphere. Inside the star, the radiation field is almost perfectly thermal, i.e., it follows a blackbody spectrum and passes out through the photosphere as a near blackbody. Above the photosphere, there is cooler gas in the stellar atmosphere which creates absorption lines in the stellar spectrum. Stars with common properties (i.e., mass, luminosity, radius, composition etc.) have similar spectra. This provides an efficient way to measure the properties of individual stars well without surveying their emission across the entire electromagnetic spectrum.

Indeed, before astronomy developed good tools for precisely measuring light, astronomers could readily measure the relative strength of absorption lines in the spectrum of stars. Hence, starlight is usually categorized by the *spectral type*, based on the strength of absorption lines. Table 2.1 below shows the typical spectral lines seen in the stars. The weird lettering sequence in this table is an odd reflection of how observations take place: astrophysicists didn't originally understand stellar spectra so they just sorted them into categories. It was only later that a physically motivated theory was developed that could reorganize and consolidate the original classification (A,B,C...) into the temperature based sequence we see here. Each spectral type is further subdivided into a linear numeric scale, e.g, the G class of stars is divided into G0, G1, G2 $\cdots$ G9 in order of decreasing surface temperature.

Figure 2.5 shows the actual stellar spectra for stars collected in the Pickles [1998] library of standard stars. These spectra illustrate how the shapes of the stellar spectra and the relative strengths of the absorption lines change with the spectral type of stars. Stellar spectroscopy is a particularly powerful technique for measuring the properties of stars. Using a relatively narrow portion of the spectrum, an observer can carefully fit the lines present in the stellar atmosphere to determine a wealth of properties about a star.

Spectroscopic fitting is used to measure the (1) stellar effective temperature, (2) the abundances of elements in the atmosphere, and

Type	T_{eff} (10^3 K)	Strongest Line Features
O	>28	He II
B	10-28	He I, Weak H I
A	7.5-10	Strong H I
F	6-7.5	Ca II, Medium H I
G	5-6	Ca II, Weak H I, metals like Fe I, Cr I
K	3.5-5	Fe I, Mn I, Si I, molecules like CN, CO
M	2-3.5	TiO, CN, CO
Brown Dwarfs		
L	1.3-2	Metal hydrides (FeH, CrH, MgH, CaH), OH
T	0.7-1.3	CH_4
Y	<0.7	NH_3

Table 2.1: Spectral types of stars. The table uses spectroscopic notation for some species where the Roman numeral indicates the ionization stage of the atom: HI=H^0, HII=H^+.

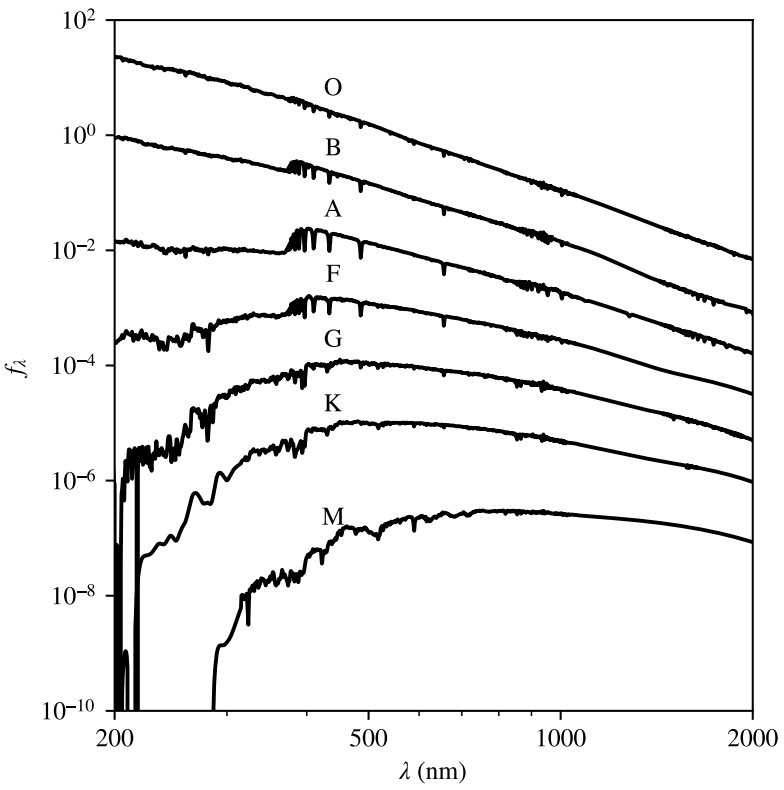


Figure 2.5: Sample stellar spectra from the work of Pickles [1998] downloaded from the ESO archive.

(3) the line width of the spectral lines. While the effective temperature of the star is formally a property of the blackbody radiation field at the photosphere of the star, this cannot be directly measured in the presence of the stellar absorption lines. Instead, the relative strengths of different species (i.e., types) of atomic ions indicates the temperature of the gas in the atmosphere of the star. The relative strengths of the spectral lines from different elements can be measured to determine the the abundances of those elements. This yields a fine-grained study of the different elemental abundances, moving beyond just a simple Z for metallicity (see Section 2.9). Finally, the line width of the spectral lines gives information about the conditions on the surface of the star.

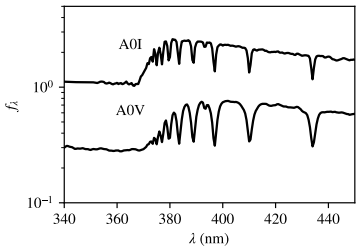


Figure 2.6: Stellar luminosity classes are determined by the width of the spectral lines. The underlying shape of the continuum is set by the blackbody spectrum and the spectral lines are from the atmosphere in the star. The relative strengths of the spectral lines are correlated with the shape of the continuum and thus the effective temperature of the star.

In Figure 2.6, we compare two different stellar spectra with the same effective temperature and similar abundances. These stars are type A0, so the strong lines you can see in the spectrum are the Balmer transitions of the hydrogen atom. The reddest line at 438 nm is the $n = 2 \rightarrow 5$ transition of the line and the bluer lines are increasing values of the upper transition. In this Figure, the top spectrum has significantly narrower lines than the bottom spectrum. Without going through a full, formal demonstration, the line width turns out to be a measure of the pressure in the stellar atmosphere. In high pressure environments, the line is broader. Thus, the top spectrum has a relatively lower pressure, which translates into having a lower surface gravity, g, which in turn implies a large radius for the star. A large radius at the same temperature implies a large luminosity ($L \propto R^2T^4$). Hence, these spectra are said to come from different *luminosity classes* measured by the line width of the stars. The luminosity classes are given Roman numerals to distinguish them as well as descriptive terms:

Class	Description
I	Supergiants
II & III	Giants
IV	Subgiants
V	Dwarfs or main sequence
VI	Subdwarfs

Table 2.2: Stellar luminosity classes

In Figure 2.6, the A0I star is a supergiant and the A0V is a main sequence or dwarf star. Our Sun is fully described in this system as a G2V star, or more casually, a "G dwarf."

THE HERTZSPRUNG-RUSSELL DIAGRAM – The primary view of observational stellar properties that we use in astrophysics is called the *Hertzsprung-Russell* (HR) diagram. This diagram has two different views. When the axes are luminosity and effective temperature, this

is called the "theorist's HR diagram" since these properties are readily calculated from stellar models. If the axes are a colour index and (absolute) magnitude, this is called a "colour-magnitude diagram" or an "observer's HR diagram." The same basic structures are visible in both diagrams.

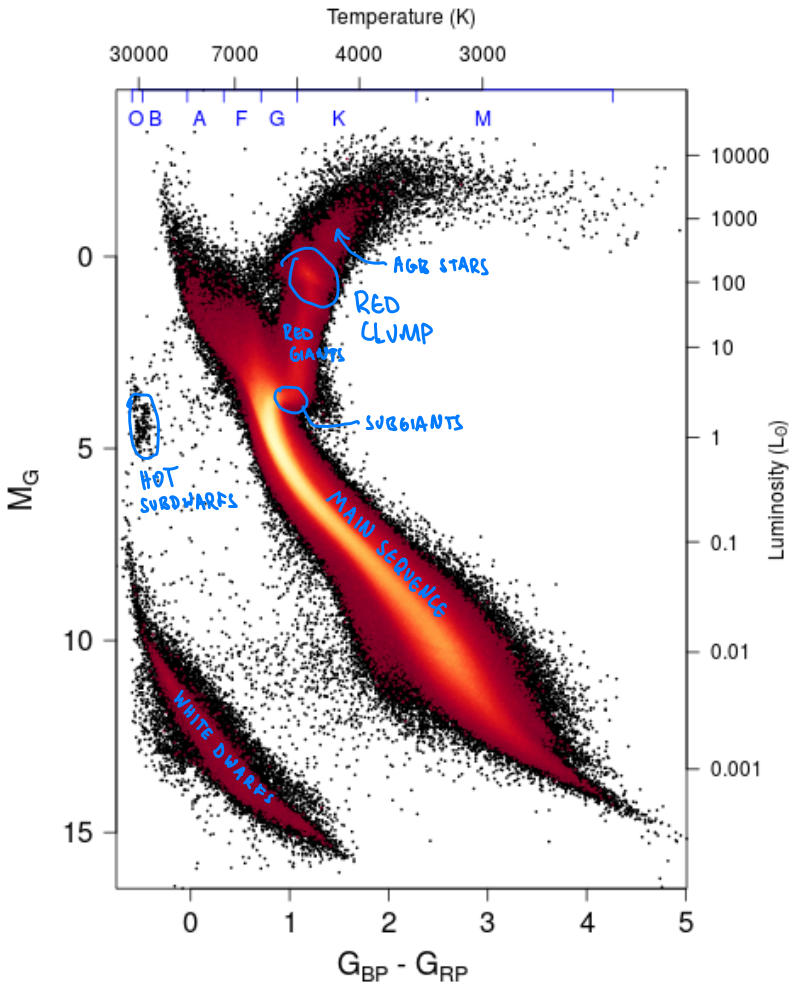


Figure 2.7: Hertzsprung-Russell diagram from Gaia Collaboration et al. [2018]. The red-white colour scale indicates the number of stars in that part of the HR diagram. Blue labels indicate the names of common stellar populations.

Figure 2.7 shows the Hertzsprung-Russell diagram observed by the *Gaia* satellite, which displays absolute magnitude in *Gaia* data (M_G) vs. the colour index in the Gaia blue and red filters ($G_{BP} - G_{RP}$). The secondary axes show *approximate* representations of the effective temperature and the luminosity of the stars corresponding to the different regions of the diagram. This diagram is made by selecting all 4.2 million of the nearby stars in the Milky Way with high quality parallax measurements whose colours are not significantly affected by the presence of interstellar dust.

The basic orientation of the HR diagram is always the same: more luminous stars are on the $+y$ axis and less luminous stars are toward $-y$. Hotter stars are on the $-x$ axis and cooler stars are on the $+x$. Note that this means that one of the axes is an HR diagram is always "backwards" by convention. Here, the absolute magnitude is increasing toward larger numerical values toward the bottom of the y axis, meaning that the luminosity of the star is increasing toward the top of the figure. Furthermore, from the Stefan-Boltzmann law, $R = \sqrt{L/(4\pi\sigma_{\rm SB}T^4)}$ which means that stars with large radii are in the upper right corner of the diagram and stars with small radii are in the lower left.

Even though the HR diagram only plots the observed (surface) properties of the stars, its main utility comes because the different populations of stars correspond to distinct physical stages in the interior. We can infer a star's evolutionary state based on where it is in the HR diagram.

We will make ample use of the HR diagram throughout this course. The structures and density of stars that you see in Figure 2.7 end up telling us a great deal about the history of the local Milky Way. However this story requires understanding telescope observations (Chapter 1), the evolution and properties of individual stars (this Chapter), and how a population of stars evolves as an ensemble (Chapter 3).

Key Points

- Most stars are found on the main sequence, which is the stage where stars are fusing H to He in their cores. Such stars follow a mass-luminosity relationship (e.g., Equation 2.8) which implies that the main sequence lifetimes of stars vary (Equation 2.10) with high mass stars being much shorter lived than high mass stars.
- Compared to low-mass stars, high mass stars also have high surface temperatures, bluer colours, and significantly higher luminosities.
- Stars are categorized observationally by their spectral type, designated by a letter sequence (Table 2.1) and a luminosity class (roman numeral). The spectral type is set by which spectral absorption lines are present in a star's spectrum.
- The Hertzsprung-Russell diagram (HR diagram) is our primary means of classifying stars and stellar populations plotting the temperature or colour index on the x-axis and

absolute magnitude or luminosity on the y-axis. No matter how these plots are made, one of the axis is "backwards" in the standard way that we plot this.

- A star occupies different parts of the HR diagram in different parts of its life.

2.3 *Stellar Evolution Summary*

This section runs through the processes that govern stellar evolution. These are presented for completeness and connection to any previous explorations of stellar astrophysics that you have encountered.

EVOLUTION ON THE MAIN SEQUENCE – Stars on the main sequence are, by definition, fusing (1) hydrogen into helium (2) in their cores through two different reaction chains that lead to the same outcome. Low mass stars are fusing primarily through the *pp chain* and medium and high mass stars are fusing primarily through the *CNO cycle*. Either way, the net effect of this fusion is to change the chemical composition in the core of the star from H into He. Or, in terms of mass fractions, X decreases toward 0, Y increases toward $1 - Z$. Practically, this means that the mean particle mass of the gas in increasing. The net effect of fusion is to convert

$$4p^+ + 2e^- \rightarrow {}^4\mathrm{He} + 2\bar{\nu}_e + E \tag{2.11}$$

where $\bar{\nu}_e$ is an electron antineutrino and E is the energy released in a fusion reaction (26 MeV for this reaction). The neutrinos flow out of the sun and the energy goes into the thermal energy of the gas. However, counting carefully, there are 6 particles on the left hand side and only 1 particle remaining on the right after the neutrinos leave the core of the star. This means that the pressure support from the particles will become smaller if the temperature remains the same simply because there are fewer particles. However, the total mass of the reactants doesn't change too much: the mass-energy conversion is only about 0.7% of the rest mass. With the reduction of pressure support, the star will contract and heat up until it comes into equilibrium again. The core temperature will be higher, leading to a rate of fusion reactions (since this depends on temperature) and a higher luminosity.

The tl;dr on this whole physics is that stars evolve on the main sequence: they increase in luminosity over the course of their lifetime. For example, the Sun formed with $L = 0.7L_\odot$ and will increase to about $L = 1.4L_\odot$ before it evolves off the main sequence.

2.3.1 *Low-mass Stellar Evolution*

Here we mean stars with masses $0.8 < M/M_\odot < 2$. We base our discussion of stellar evolution on these stars, and then describe the changes with mass in terms of this basic evolutionary scheme. See below for some figures that show how these stars are moving the HR diagram.

THE SUBGIANT PHASE – These stars have radiative cores and radiative-plus-convective envelopes. For our purposes, the *core* means the part of the star that is undergoing nuclear fusion and the *envelope* is the inert outer layers that are not participating in the fusion process. The radiative nature of the stellar cores means that the material is static and a given parcel of gas remains at the same radius in the star. Since the temperature and density profiles in a star are highest in the centre ($r = 0$) and decreasing with radius, this means that the nuclear energy generation rate ($\epsilon \propto \rho T^4$) will be strongly centrally peaked and it will remain so over the life of the star. In terms of energy generation, this means that matter at the centre of the star is will fuse its hydrogen into helium at a faster rate, depleting the hydrogen content and arriving at a depleted state first ($X = 0, Y = 0.98, Z = 0.02$). In this gas, fusion will shut off, but the gas will be hot and surrounded by a region of the core that is still fusing hydrogen. This gas will also be isothermal since it is not producing any fusion within it. Instead it will be held at a constant temperature set by the energy generating region around it.

As the star continues to fuse, this inert, isothermal core will get larger and the fusing region move to progressively larger radii. In this time, it transitions from being a core-burning star into a shell-burning fusion source. Shell-burning stars obey something called the *mirror principle*, which is an observation gleaned from the models that when a shell contracts, the envelope of the star will expand and vice versa.

There is a maximum mass for the isothermal helium core before it collapses under its own self-gravity. This is called the Schönberg-Chandrasekhar limit and is about 10% of the star's mass. At this point, the isothermal core contracts and heats up on the cooling timescale (also known as the Kelvin-Helmholtz timescale). During this contraction, the core, and shell-burning source shrink, meaning that the envelope expands. More of the mass will become convective. The contraction is eventually halted by the onset of degeneracy pressure.

THE RED GIANT BRANCH – At this point in the evolution, the star is

at the base of the red giant branch. It consists of a degenerate He core with a shell burning source on top of it. As the shell source continues fusing the hydrogen into helium, the He ash builds up in the degenerate core, adding mass to it. As the mass of the core increases, its radius decreases, because of the mass-radius relationship for degenerate objects, the radius of the core gets smaller. Following the *mirror principle* the shell burning source (and the degenerate core) will heat up and the envelope will expand. The red giant will increase its luminosity at roughly constant surface temperature ($T_e \approx 3500$ K). Thus, it will appear to move "up" the red giant branch.

This evolution occurs on a nuclear timescale and will take nearly 10^9 years for a $1M_\odot$ star. During this phase, the core gets smaller and hotter, approaching the threshold temperature for 3α fusion, 10^8 K. All stars reach this temperature when the core mass is about 0.4 $M_\odot$. Once He fusion ignites, it provides an energy source to its environment, which heats it up. Under normal conditions, this local heating would produce an increase in pressure, which would cause the gas to expand and reduce the fusion generation rate. This thermostatic mechanism keeps fusion in our Sun stable. However, in a degenerate gas, the equation of state has no sensitivity to temperature and the perfect gas component of the equation of state is much smaller than the degenerate component. (Note that the combination of the two equations of state isn't as simple as just adding the pressures, but it will suffice for our discussion.) This lack of a pressure sensitivity means that there is no thermostatic process that will reduce the rate of fusion generation. Instead, the local heating will prompt the nuclear fusion under the 3α process to increase much more ($\epsilon_{3\alpha} \propto T_8^{40}$). This increase in temperature leads to a runaway fusion process and our truism: *Nuclear burning in degenerate matter is unstable.* This is just a nice way of saying a big He thermonuclear explosion goes off in the centre of the star. In their quiet, understated way, astronomers call this the *Helium Flash*. The amount of energy is not enough to destroy (unbind) the star. Note that, the helium flash could do so for lower mass stars, but no such stars have ever evolved off the main sequence. The flash provides enough thermal energy input to raise the temperature throughout the core and to settle into a perfect gas equation of state (and pressure support) again. The star then stabilizes on the Helium burning sequence.

The Helium Burning Sequence – These stars are undergoing 3α fusion of He into C in their core. They have a H burning shell source (using the CNO) cycle and radiative envelopes. Since all the stars evolve from the ignition of a 0.4 $M_\odot$ core, all these stars have a roughly constant luminosity of $L \approx 50L_\odot$. Their surface temperatures

can vary depending on whether there is mass lost during the movement onto the helium burning sequence (HB) or the metallicity of the star. This means that, in a Hertzsprung-Russell diagram, they will all lie on a horizontal line, leading to the other name for such stars: the Horizontal Branch. Most solar metallicity stars end up on the red side of the branch where they are called the *Red Clump*.

THE ASYMPTOTIC GIANT BRANCH – HB stars will evolve over another 10^9 years on the main sequence and build up a C/O core that contracts and becomes supported by degeneracy pressure. At this point, they move onto the *Asymptotic Giant Branch*. These stars have a convective envelope, a degenerate C/O core and two nested shells of H and He burning. Like the RGB, these stars deposit material on the C/O core, causing it to get smaller and hotter. If these stars reached the internal temperatures of $T = 10^9$ K where ^{12}C fusion could ignite, they would proceed to another phase of fusion. However, the shell fusion in these stars gets sufficiently intense that it begins to force off the outer layers of the stars. This process happens in pulsations: the shell source builds up thermal energy, pushes off the other layers which reduces the pressure on the shell source and then its luminosity. The envelope of the star settles back onto the shell, building up energy again, which then pushes out. This is the thermostatic mechanism at work, but the pulses are much larger. These pulses throw off the outer layers of the star, which are observed as *planetary nebulae* and are some of the most amazing pictures we have of stuff in space. `http://hubblesite.org/gallery/album/nebula/planetary/)` has some cool Hubble pictures.

After the thermal pulses on the Asymptotic Giant Branch throw off of the whole outer layer into a planetary nebula, the remnant is a hot C/O core, supported by degeneracy pressure. We call this object a white dwarf and they show up in the lower left-hand corner of the HR diagram.

2.3.2 *Medium mass Stellar Evolution*

For stars with masses $2 < M/M_\odot < 9$, their evolution is similar to low mass stars, with the exception that there isn't the build up of an isothermal core in the subgiant phase. This is due mostly to the dominance of the CNO cycle during main sequence burning. Stars with CNO cycle burning tend to have convective cores because of their centrally peaked energy generation rate. This means that, instead of building up an inert region in the centre of the star, the fusion at the core will be fed by convection of material from higher up in the star. This means that the fusion will deplete the material in the core

all at once rather than building up an isothermal core steadily. These stars will undergo core contraction and move toward and up the red giant branch, but their cores will not become strongly degenerate as the low mass stars do. They thus do not undergo a helium flash and instead have a smooth ignition of helium on to the helium burning sequence on a cooling (Kelvin-Helmholtz) timescale. Such stars are briefly on the Red Giant Branch based on their envelope properties but they do not have the same structure as for a low mass star below.

From this point on, their evolution follows that of a low mass star: helium burning sequence followed by thermal pulsations and the formation of a C/O white dwarf.

2.3.3 High Mass Stellar Evolution

High mass stars are notable because they do not experience a degenerate core during their evolution. Every stage of nuclear burning ends, contracts, heats up and ignites the next heaviest elements of nuclear burning. Of note, such stars, after helium fusion, will be able to ignite $^{12}C + {}^{12}C \rightarrow {}^{24}Mg$ and other heavy element fusion processes that occur above $T = 10^9$ K. These fusion stages all occur at roughly constant luminosity, but the stars undergo significant internal restructuring. For example, when their core contracts and their envelopes expand, their surface temperatures must get smaller meaning they move to the right in the HR diagram. Similarly, when the core ignites and expands again, the envelope will shrink and the star will move to the left.

Once stars become massive enough to fuse past carbon, they are also sufficiently massive to fuse up to the peak of the binding-energy-per-nucleon curve at ^{56}Fe. They do so but ultimately build up an iron core that cannot be fused exothermically. The stars undergo a core collapse and subsequent supernova explosion. This means that the ability to ignite C fusion represents the boundary between stars that will end their lives as white dwarfs ($M < 9M_\odot$) and those that explode in a supernova, leaving behind neutron stars and black holes ($M > 9M_\odot$).

Key Points

- Stars slightly increase luminosity and temperature on the main sequence because their cores need to heat up to maintain pressure support. The pressure support reduces because fusion reduces the number of particles in the star.
- Low mass stars ($M < 2\ M_\odot$) and medium mass stars ($2 < M/M_\odot < 9$) evolve into red giants once their cores ex-

haust H fusion. H fusion occurs in a shell around the core which causes the envelope to swell up and cool. After the red giant phases, stars can ignite He to C fusion in the core and occupy the red clump (solar metallicity) or the horizontal branch (low metallicity) when the fuse He to C in their core. They then go through a red-giant like phase called the Asymptotic Giant Branch and then lose their envelopes (planetary nebulae) and become white dwarfs.

- High mass stars follow similar physics as low mass stars but can also get hot enough to ignite carbon fusion allowing them to fuse into magnesium and heavier elements up to iron. After exhausting Fe fusion, the star undergoes a core-collapse supernova explosion. Though, low mass stars will undergo different outcomes (See section 2.7).

2.4 *Stellar Mass Loss*

A major subtlety in stellar evolution is that entire picture we have been painting so far uses the "initial" mass of the star as a parameter for how the star evolves. This measurement means the mass of the star when it starts hydrogen fusion in its core on the *zero-age main sequence* (ZAMS). All stars are losing mass from their outer layers in the form of stellar winds (see Ch. 1 for a mention of our solar wind). For moderate mass stars like the Sun, the effect of this mass loss is relatively weak ($\dot{M} \sim 10^{-14}\ M_\odot/\mathrm{yr}$). However, for high mass stars with extremely strong radiation fields, the radiation pressure boosts the stellar mass loss causing these stars to lose a significant amount of mass over the course of their main sequence lifetimes. These losses are generically called stellar winds and have two effects that we note in the context of galaxy evolution. First, the winds can inject energy and momentum into their surroundings and shape the gas in their environment, playing a vital role in self-regulation of galaxies. Second, these winds reduce the mass of the star, leading to differences in the star's evolution. Wind strength (mass loss rate) is dependent on the metallicity of the star scaling roughly as $\propto \sqrt{Z}$ where Z is the mass fraction of metals. Stars with higher metallicity have stronger winds because the material will have a higher opacity, making radiation pressure driving more effective.

Figure 2.8 shows the fraction of mass lost as a function of the initial stellar mass at two different metallcities: high (solar metallicity, $Z_\odot$) and low ($10^{-3} Z_\odot$) where the latter corresponds to the metallcity expected for stars formed early in the history of a galaxy. Note that

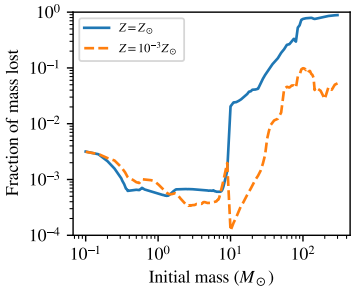


Figure 2.8: Fraction of mass lost from stars of different masses at different metallicities.

the solar metallicity, high-mass stars can lose almost their entire mass due to winds before they run out of hydrogen to fuse. At low metallcities, the mass loss is less significant, leading different outcomes at the end of a star's life.

Key Points

- High mass stars lose a significant fraction of their mass during their lives prior to their deaths. High metallicity stars lose more mass than low metallicity stars.

2.5 *Star Structure*

Figure 2.9 shows the cutaways of the evolution of a low mass star at different stages of its evolution. It's worth noting that the figures are very much *not to scale*, and the transitions between the stages are gradual. The diagrams are just to remind you of relative ordering of the components of the stars.

2.6 *Supernovae*

We now turn to laying out the different types of supernova explosions. Formally, a *supernova* (SN) is an explosion of a star with a characteristic energy injection into surrounding galaxy of $E_{\rm SN} \sim 10^{44}$ J. This is an observationally defined convention and fits into an explosion spectrum, where the weakest explosion is a *nova* with ($E_{\rm nova} \sim 10^{38}$ J), a kilonova (10^{42} J) and even a hypernova ($E > 10^{45}$ J). These are all from different physical mechanisms and here we focus on supernova because they are energetic and relatively common. Thus, the impact of supernova can significantly shape galaxy evolution.

There are two main physical mechanisms for supernova, referred to as *thermonuclear supernova* and *core collapse supernova* sometimes refereed to based on their observational classification scheme of Type I and Type II respectively[2].

Core collapse supernova occur when the core of a massive star can no longer generate thermal support through fusion, and it collapses until the collapse is halted at very high densities where the material forms into a neutron star. Most of the mass-energy from a supernova actually leaves the system in the form of neutrinos ($E \sim 10^{46}$ J) and the energy injection of 10^{44} J is only a small fraction of that. The energy is mostly a kinetic energy blast wave with 1% of the energy injection coming out in the photons that we observe. Formation of a neutron star, even if only temporarily, is required for most supernova explosions.

[2] Again, the type I/II classification is a reflection of how observations take place: a mysterious phenomenon has two different broad groups of objects so these just get sorted into two types: I and II. Later refinements start to break these down Type Ia, Ib, Ic etc. The neat story that you are getting is the product of a long process of thinking about how to explain the observations.

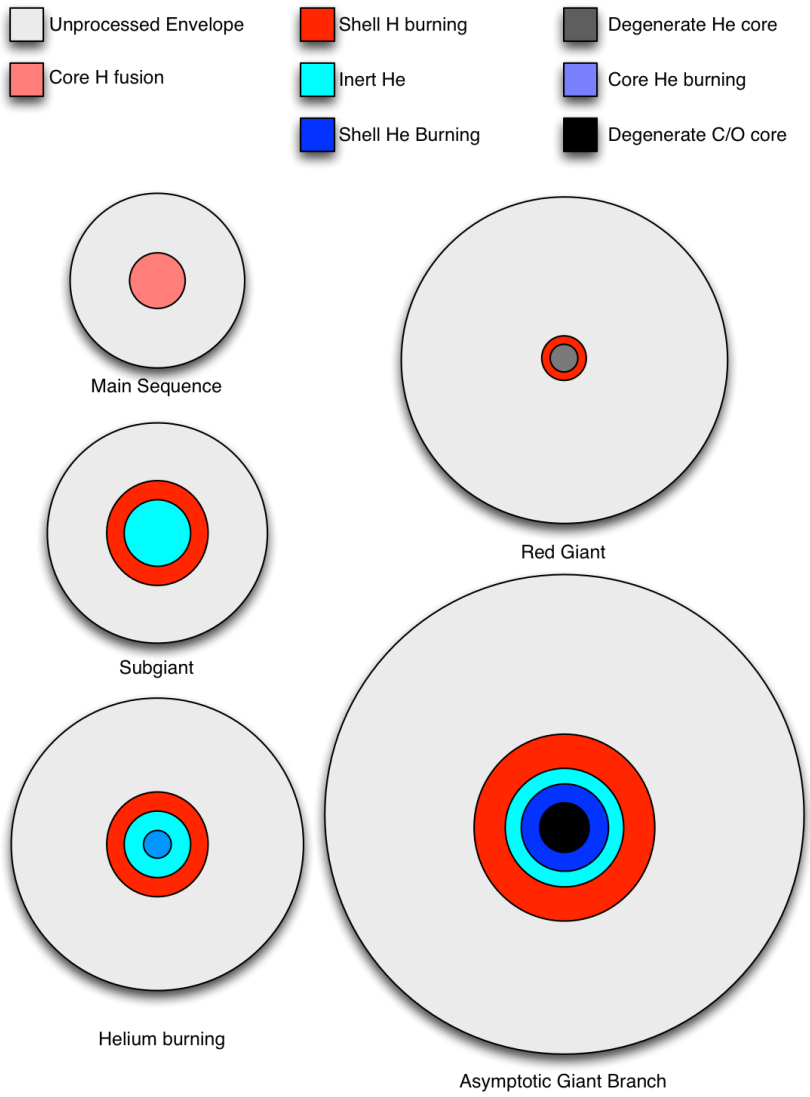


Figure 2.9: Schematics of stellar interiors during different parts of their life.

Thermonuclear supernova are thought to occur when two white dwarfs with a typical mass of 0.6 to 0.8 $M_\odot$ in a close binary system spiral in and merge through their interactions under binary evolution. These white dwarfs consist of C/O ions with their electrons in a degenerate gas. Under these conditions, the C/O begins unstable nuclear burning to generate a lot of Si, Fe, Ni releasing their nuclear energy as a supernova explosions. It is coincidental that these have the same energy scale as a core collapse supernova since the mechanisms are vastly different.

Finally, there is a hypothesized rare supernova explosion called a *pair instability supernova* explosion that occurs in very low metallicity systems. The mechanisms for this explosion is that during collapse the high energy photons spontaneously pair produce into a population of electron/position pairs which temporarily act as a sink for radiation pressure, prompting further collapse, more pair production, etc., leading to an unstable collapse and subsequent explosion. Pair instability SN explosions are only thought to occur in very low metallicity gas and were thus relevant in the early universe. Now, stellar nucleosynthesis has created an abundance of metals making these SN rare.

Key Points

- There are many kinds of supernova explosions with the main divisions being *core collapse* and *thermonuclear* supernova. These are different physical mechanisms but both explode with $\approx 10^{44}$ J of energy.
- Thermonuclear supernova come from two white dwarfs merge, triggering runaway fusion of their C/O into Si/Fe/Ni.

2.7 *Remnants*

Stellar Remnants refers to the objects that are left over after a star completes its evolution. Stars have four main outcomes from stellar evolution:

1. White Dwarfs – These are formed by low and medium mass stars ($M < 9\ M_\odot$) and consist of the former core of the star. The pressure support is provided by electron degeneracy pressure, leading to objects with a characteristic size of $R \sim 10^7$ m, which is about Earth sized. The characteristic mass of most white dwarfs is about $M \sim 0.6 - 0.8\ M_\odot$. White dwarfs are hot with the residual heat from the nuclear burning period of the stars lifetime, and they

have a surface temperature of $T_{\rm eff} \gtrsim 10^4$ K. White dwarfs evolve by cooling off. Because their pressure support does not depend on temperature, the star stays the same size and evolves along the $L \propto T^4$ line in a HR diagram.

2. Neutron Stars – Neutron stars are formed in the supernova explosions of high mass stars with $9 < M/M_\odot < 25$. These objects have pressure support from neutron degeneracy pressure, which leads to objects with a typical mass of $M = 1.4\ {\rm M}_\odot$ though neutron star masses can reach up to $M \sim 2.2\ M_\odot$ before the neutron degeneracy cannot sustain pressure support. Neutron star radii are $R_{\rm NS} \sim 12$ km, implying that their densities approach that of atomic nuclei. At these densities, both gravitation, degeneracy pressure, and particle interactions with the strong nuclear force are significant effects, making neutron stars an excellent tool to probe fundamental physics.

3. Black Holes – If a remnant is more massive than the maximum mass threshold that can be sustained by pressure support – between 2 and 3.2 $M_\odot$ depending on the details of the system – the system will collapse into a black hole. Non-rotating black holes have radii $R_{\rm BH} = 2GM/c^2$ with no upper limit on this mass. In galactic astrophysics, black holes can originate as the endpoint of stellar evolution leading to masses $3 < M_{\rm BH}/M_\odot < 100$. There are also *supermassive black holes* found in the centres of galaxies with masses $M > 10^5\ M_\odot$ and astrophysicists are also searching for *intermediate mass black holes* between these two regimes ($10^3 - 10^4\ M_\odot$).

4. No remnant – Stars that undergo pair instability supernova explosions (see previous section) will leave behind no remnant.

Figure 2.10, from Smith [2014] (see also Heger et al. [2003]), shows the different cases of how stars end their lives based on the initial mass and the metallicity of the original star. This is a complicated diagram that associates the different parts of the parameter space with the observed phenomena. At roughly solar metallicity and below, stars between $9 < M/M_\odot < 25$ undergo core-collapse supernova and leave behind a neutron star remnant.

For $M > 25\ M_\odot$, there can be a supernova explosion and fallback material (that doesn't get blasted away and instead falls back onto the neutron star) leading to collapse to a black hole or, for lower metallcities, the star can collapse directly to form a black hole without a supernova ("Direct black hole" in the Figure). Figure 2.11 shows a star in the nearby galaxy NGC 6946 disappearing between observations of the same field at different times [Adams et al., 2017]. This

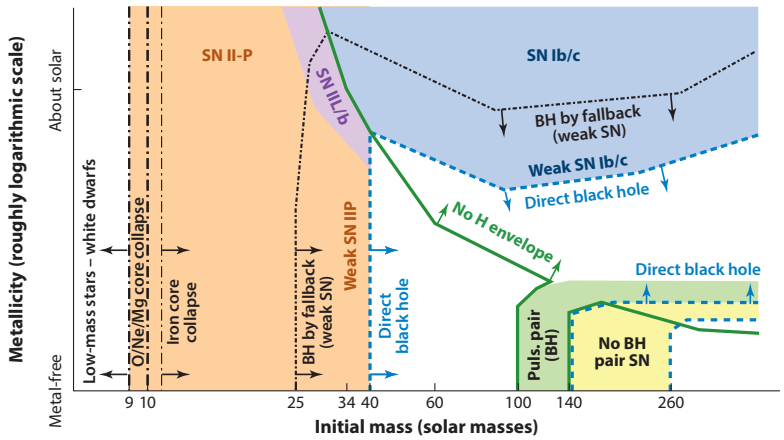


Figure 2.10: Different outcomes of high mass stellar evolution from Smith [2014], including the different types of core-collapse explosions.

event is likely the direct collapse of a massive star to form a black hole.

The metallicity dependence in this figure arises mostly from the mass loss effects described in Section 2.4. At high metallicities, a high mass star can lose a significant fraction of its material through winds. Having this loss of material reduces the amount of pressure support that nuclear fusion needs to provide. This leads to these very high mass stars ($M > 100\ M_\odot$) following the evolutionary pathway of a lower-mass star and, for example, ending their lives in a relatively normal supernova explosion and formation of a neutron star.

Practically, we will work with defined boundaries at solar metallicity for the remnants produced by a star give its initial mass as described in Table 2.3.

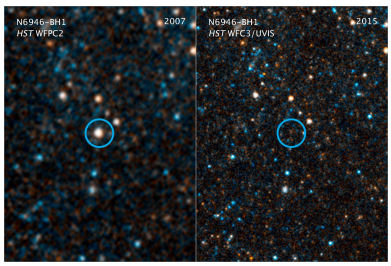


Figure 2.11: A massive star in NGC 6946 disappearing in the eight year interval between images, likely collapsing to form a black hole without a supernova explosion.

Mass range ($M_\odot$)	Outcome
< 0.9	No time to evolve off main sequence
$0.9 < M < 9$	Planetary nebula, C/O white dwarf
$9 < M < 25$	Core collapse supernova, neutron star
$M > 25$	Core-collapse supernova, black hole

Table 2.3: Outcomes of stellar evolution.

Key Points

- Table 2.3 summarizes the outcomes of different mass ranges of stars, depend on the initial mass and metallicity of the star.

2.8 Evolution in the HR Diagram

The above discussion focuses on the behaviour of the interior of the stars with some attention to the envelope. We can only see the properties of the stars from the outside, and thus it is only the observable parts of the stars (Section 2.2) that are detected in the telescope.

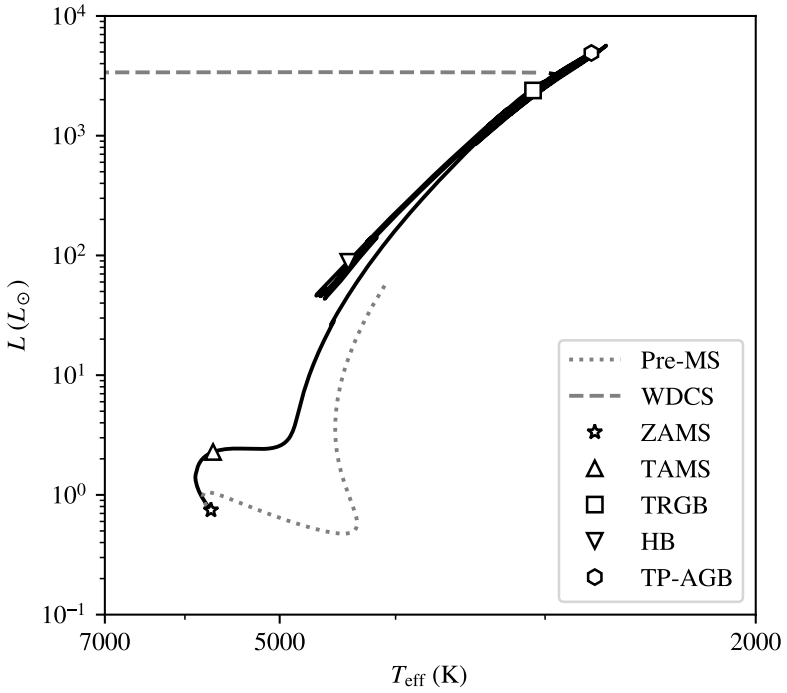


Figure 2.12: Evolutionary path of a 1 $M_\odot$ star with $Z = Z_\odot$ in the HR diagram. The solid black line indicates the path during the stellar phase of life. The dotted line shows the pre-main sequence evolution during the star formation process. The dashed line shows the white dwarf cooling sequence as the outer layers are shed and the star transitions to the lower-left portion of the HR diagram.

In Figure 2.12, we show the evolutionary track of a 1 $M_\odot$ star with solar metallicity in the HR diagram. We often say a star "moves" in the HR diagram but this practically means that the luminosity, temperature, and radius of the star is changing. It is not moving through space. For example, as the star moves later it its life, its luminosity tends to increase ($+y$ direction) and the surface temperature decreases ($+x$ direction), which means that the radius increases and the star appears brighter and redder. These general properties give this stage the name 'red giant.' The different line styles break the evolutionary track into its pre-main sequence stage (grey dotted line), main sequence and post main sequence (solid black) and the transition to the white dwarf stage (grey dashed). The markers indicate various stages of the star's life: ZAMS - zero-age main sequence; TAMS - terminal age main sequence; TRGB - tip of the red giant branch; HB - helium (core) burning; TP-AGB - Thermal

pulsation asymptotic giant branch. Our Sun is currently located at $T_{\rm eff} = 5777$ K and $L = 1\ L_\odot$.

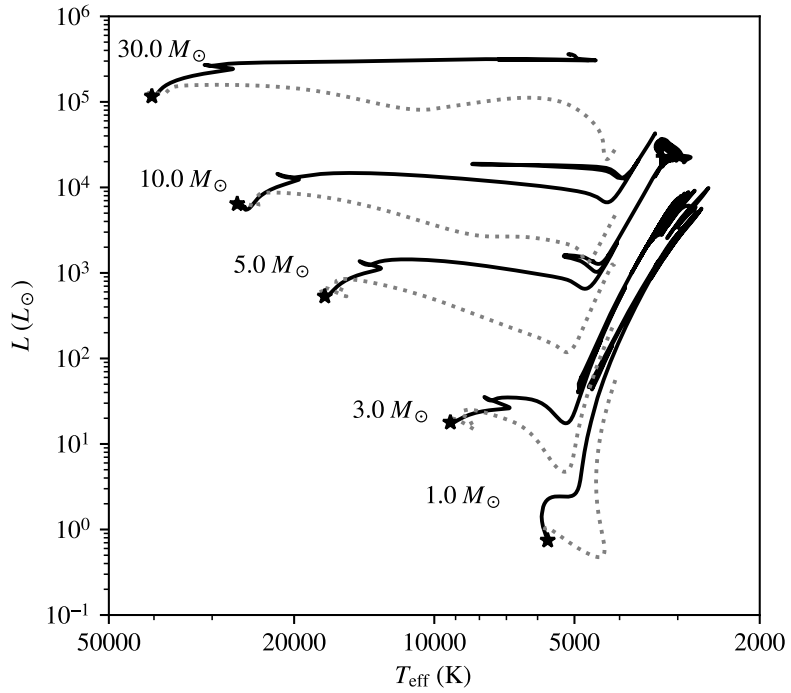


Figure 2.13: Evolutionary tracks for different mass stars with $Z = Z_\odot$

Figure 2.13 show the evolutionary tracks for stars with solar metallicity but different initial masses. These follow much of the same general structure as the 1 $M_\odot$ case with moving into the RGB/AGB portion of the diagram later in their life. In general, high mass stars move horizontally in the HR diagram through their life, which reflects that their energy transport is dominated by radiative diffusion. Lower mass stars tend to move vertically, which reflects the convective nature of their outer envelopes.

For a single metallicity, the tracks for different masses don't cross significantly and so we can translate a position in the HR diagram into an understanding of the interiors and the evolutionary state of stars. Note how the changing interior structure of the star is indeed reflected in the envelope.

Our neat picture of the HR diagram becomes more complicated once we introduce the effects of metallicity. The metal content of a star plays a significant role in where stars are found in the HR diagram, primarily because stars with higher metallicity have material with a higher opacity. This opacity arises because metals can bind electrons at high temperatures providing a target for the photons in

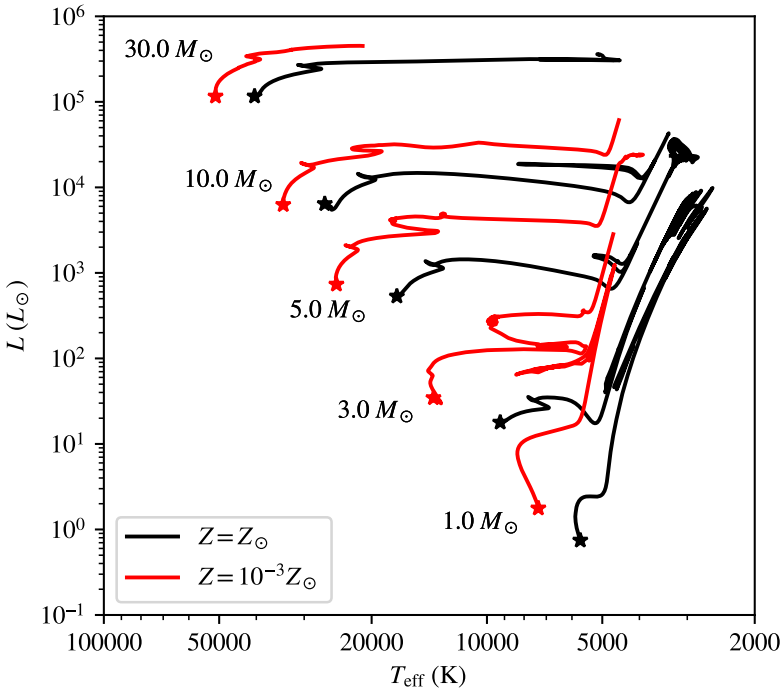


Figure 2.14: Evolutionary tracks for stars with different masses with $Z = Z_\odot$ (black) and $Z = 10^{-3} Z_\odot$ (red).

the stellar interior to hit and ionize, slowing down the rate at which energy is diffusing out of the core. Figure 2.14 compares the evolutionary tracks for stars with solar and subsolar metallicity. In general, stars with low metallicity compared to the sun are hotter and brighter, which reflects that their interiors have lower opacities and their radiation generated by fusion has an easier time diffusing out to the photosphere. Without detailed information from stellar spectra, it can be difficult to distinguish the position of a star with a lower mass and lower metallicity from a star with higher mass and higher metallicity. The two tracks can overlap in parts of the HR diagram.

Key Points

- As stars evolve they "move" in the HR diagram, which really means their outer layers change to show a different luminosity and surface temperature. These changes respond to the evolution in the interior of the star.
- The path a star traces out in the HR diagram is called its evolutionary track and depends on the stars mass and metallicity (Figures 2.12 and 2.13).

- Low metallicity evolutionary tracks are displaced to higher temperatures and luminosities compared to their solar metallicity counterparts because these stars have lower opacities, allowing radiation to escape their interiors readily (Figure 2.14).

2.9 *Metallicity and Chemical Composition*

As a coda to this Chapter, we return to the topic of the chemical composition of the Universe and the abundances of elements. In the study of galaxy evolution, chemical abundances turn out to be an important tool. This importance comes because the abundances of elements are relatively easily measured using spectroscopic methods in both stars and in the gas that makes up the interstellar and intergalactic medium. In addition, during the process of stellar evolution, the elements in the surface layers of stars remain unchanged through the main sequence stage of stellar evolution[3].

[3] The exception to this are low mass stars with $M \lesssim 0.25\ M_\odot$ and are convective through their entire mass. This mixes the products of nuclear fusion into the outer layers where they can be observed.

Like so many things in astronomy, there are three different conventions for expressing the metallicity/chemical composition of material which are used in the different subfields of the discipline. As discussed previously in this chapter, the study of stellar evolution keeps track of the *mass fraction* of different elements. This is important because something like helium has a mass $\sim$ 4 times of that hydrogen. Thus, a typical star is 75% hydrogen by mass but 90% of the atoms in a star are hydrogen.

For context, let's first consider what the actual abundances of elements look like in our solar system and where those elements originated. Figure 2.15 shows the relative abundances of elements on a logarithmic scale in terms of the **number** of atoms rather than the mass fraction. To calculate the number of atoms such as the relative number of iron (Fe) atoms to hydrogen, we calculate the differences in logs: $10.4(\text{for hydrogen}) - 5.9(\text{for iron}) = 4.5$. Thus there are $10^{4.5} \approx 32,000$ hydrogen nuclei in the solar system for every iron nucleus.

The chemical abundances of the species show a characteristic sawtooth pattern with elements that have a number of nucleons divisible by four being more abundant than the other species. This overabundance arises because these divisible-by-four nuclei are more stable under the strong nuclear force than the other elements, leading to it being easier to form these elements. This can be seen, in part, by comparing Figure 2.15 with the binding energy per nucleon curve in 2.3 shows that elements that are at local maxima of the binding energy-per-nucleon curve also have high abundances. It also shows

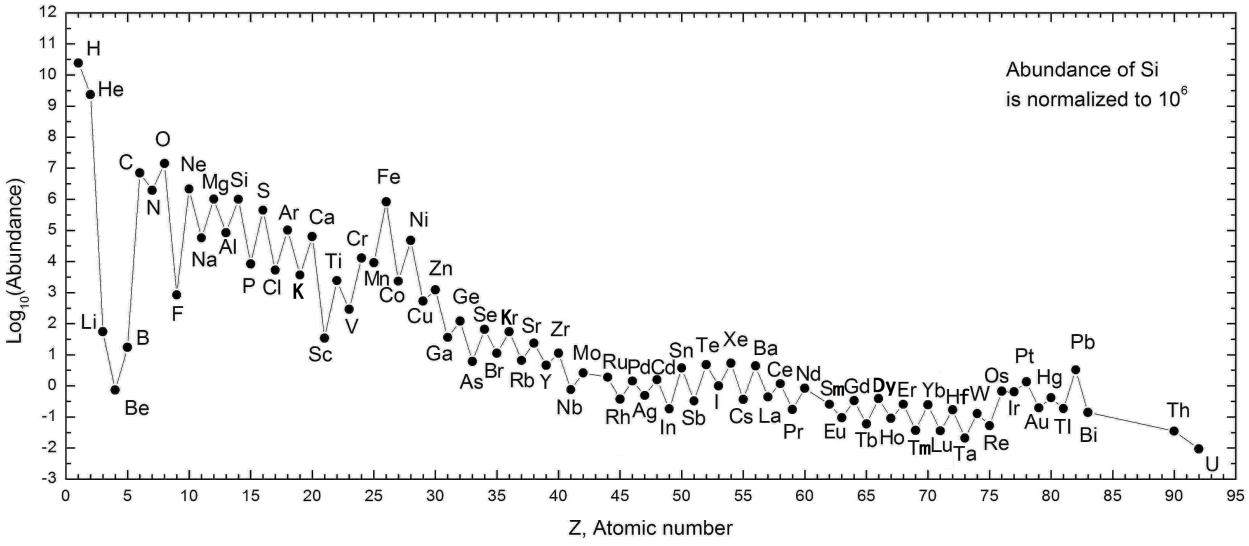


Figure 2.15: Relative abundances of the elements as measured in logarithmic units. From Wikipedia user 28Bytes, reproduced under the CC-SA-3.0 license.

that elements beyond the iron peak are quite rare.

- Hydrogen, helium, lithium – These abundances are set by nuclear reactions in the early universe, a process called *Big Bang Nucleosynthesis*. This sets the primordial hydrogen-to-helium ratio at $X = 0.75, Y = 0.25, Z \sim 0$.
- "Metals" lighter than iron (oxygen, nitrogen, carbon, neon) – These are formed through nuclear fusion processes in stars.
- Iron-peak elements (iron, nickel) – These are formed from thermonuclear supernovae, when two white dwarf stars merge together.
- Elements heavier than the iron peak (e.g., gold, uranium) – The come primarily from the merger of binary neutron stars, which create the neutron-rich conditions required to build up heavily nuclei before radioactive decay breaks those nuclei apart.

Because of these varied origin processes, studying the detailed chemical elements of stars and gas gives insight to the physical processes that have influenced that material. Hence it is useful to consider the middle ground between tracking just hydrogen, helium and metals and following the abundances of every element.

Element Groups – It sometimes suffices to consider the abundances of two different element groups associated with the different processes of stellar evolution. The two main metal groups are the

iron-peak elements (i.e., Fe, Ni) from thermonuclear supernovae explosions and the 'α' process elements which are the species formed from the combination of α particles, which is a common name for helium nuclei. One of the most common representation of tracking element groups is as a logarithmic ratio compared to the values found in the Sun. For example, a common indicator is $[\alpha/\mathrm{Fe}]$, which compares which compares the abundances of α-process elements to that of Fe peak elements.

$$[\alpha/\mathrm{Fe}] = \log_{10}\left(\frac{N_\alpha}{N_{\mathrm{Fe}}}\right) - \log_{10}\left(\frac{N_{\alpha,\odot}}{N_{\mathrm{Fe},\odot}}\right) \tag{2.12}$$

Here the variable N indicates the number particles of the species, this ratio is formed by *number* and is not tracking *by mass*. In this equation, the α-process elements are those elements built up through fusion processes that add α particles ($^4_2$He) to nuclei (e.g., $^{12}_{6}$C, $^{16}_{8}$O, $^{20}_{10}$Ne, $^{24}_{12}$Mg, etc.). Fe refers to iron-peak elements, i.e., the elements found near the binding-energy per nucleon peak at ^{56}Fe. As an aside, the reason we care about this distinction is that these two families of elements separate enrichment caused by nuclear processing in high mass stars and subsequent core-collapse supernovae, which tend to form disproportionately more α elements vs. Type Ia supernovae, which tend to form elements near the iron peak. Core-collapse supernovae come from massive stars and thus require recent star formation. Type Ia supernovae are delayed with respect to the star formation history by as much as gigayears. Stars with high values of $[\alpha/\mathrm{Fe}]$ are thought to have formed relatively early in the star formation out of near-primordial gas. Stars formed later sees the gas form decreases over time, reflecting the Fe abundance rising up to "normal" levels. Because of this traction we have, this diagnostic can indicate the star formation history for a given star.

METALLICITY NOTATIONS – To briefly summarize, we have three main ways of expressing metallicity. First, we have presented the mass fraction of metals expressed with the variable Z and fixed on the range from 0 to 1 by the convention that $X + Y + Z = 1$. Next, we used the "bracket notation" which is a comparison of the number ratio of a species to the same ratio measured for the Sun (Equation 2.12).

The final common convention in astronomy is also a "by number" convention, which is just the $\log_{10}$ of the number of atoms of a given atom compared to hydrogen. Because the abundances of all elements are less than hydrogen, this logarithm will have negative values and astronomers typically add an offset of 12 to the measure. For example, the oxygen abundance can be expressed as $12 + \log_{10}(\mathrm{O/H})$.

This convention is particularly common for measurement of the metallicity in the gas of the interstellar medium.

Key Points

- Stellar metallicity Z is a coarse measure and the specific groups of elements (CNO, iron-peak, trans-iron) have different origins. By following the relative abundances of these groups, we can trace the enrichment history of a star or stellar population.
- We use three different metallicity notations: the Z for stars, bracket notation (e.g., Equation 2.12) and the $12 + \log_{10}$ notation where the former measures mass and the latter two count the numbers of atoms.

3
Stellar Populations

Equipped with the theory of stellar evolution, we are now equipped to understand the emergent light from groups of stars and then use that information to understand the structure of galaxies. We aim to understand this in two cases: resolved and unresolved *stellar populations*. Here, the distinction means whether you are measuring the light from individual stars / stellar systems (resolved) or grouping the stars all together to measure an integrated spectrum of light. A great example of the resolved stellar population in the Solar Neighbourhood is shown in the HR diagram in Figure 2.7.

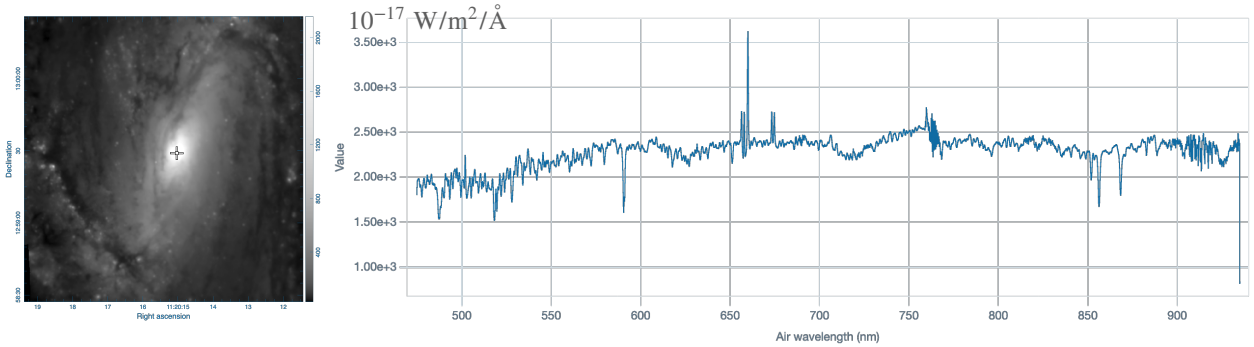


Figure 3.1: Sample Spectrum from the centre of NGC 3627 taken with the MUSE instrument on the Very Large Telescope [Emsellem et al., 2021]. The left image shows the SDSS *r* band equivalent image of the galaxy and the right panel shows the spectrum taken at the indicated point.

We haven't explored unresolved stellar populations in great detail yet but Figure 3.1 shows the spectrum of a region in the nearby galaxy NGC 3627 extracted at the point highlighted with the + sign. The spectrum shows a huge wealth of detail from the spectral lines of many different types of stars. The spectrum also shows some bright emission lines that are associated with nebulae in the galaxy. We will explore these in more detail in our study of the interstellar medium. To understand this spectrum, we need to know the types of stars that give off the light, taking into account the relative numbers and luminosities of those stars.

The study of stellar populations aims to understand both these diagrams in terms of stellar evolution combined with important factors in galaxy evolution. Relevant for galaxy evolution the life of a single star can be described by its its initial mass, metallicity, and age. We also care whether that star is in a binary or multiple star system. This tells us what we need to know about galaxies to predict the behaviours of their light. In short, we need to know:

- the total mass of stars formed as a function of time. This is called the *star formation history*. We usually use our understanding of stellar populations to infer the star formation history, giving us insight into galaxy evolution. Thus, this is an *output* of our study here through a process of *population synthesis*.
- the probability density function (PDF) of the masses of those stars (and any variations in that PDF). This is called the *initial mass function*. The IMF is observed to be fairly constant in the cases where we can measure it.
- We need to know the metallicity of those stars as they form. This doesn't have a special name but it is described through the study of *enrichment*.
- We need to know whether those stars are formed in a binary or multiple system. This is described by the *companion frequency*.

Here, we will focus on the IMF, only briefly noting how metallicity and companion frequency affect our conclusions. Our main goal is to understand how stars will evolve. Our first goal will be to understand the evolution of a *simple stellar population*, namely a group of stars formed at the same time with the same metallicity but different masses. From there, we will explore how sets of simple stellar populations combine to give us the stars we see.

FRAMING THOUGHTS – In our study of stellar populations, we must keep in mind a few things about the stars we are studying. Specifically, we must remember that high mass stars have disproportionately high luminosities and very short lifetimes in a cosmic sense. Low mass stars are common and long lived but contribute relatively little light to the ensemble of stars. However, we shall see that most of the mass of stars in the galaxy is in these nearly invisible low-mass stars.

We also need to remember that our observations are fundamentally shaped by observational effects, in particular telescope sensitivity and resolution. For a telescope with a fixed flux sensitivity, we are able to detect brighter things out to a farther distance. This is

just a reframing of the inverse square law of light: $f = L/(4\pi d^2)$. For a fixed f set by sensitivity, objects with a larger L can be see to larger distances d. We refer to the distance out to which a given telescope survey can detect objects of a given luminosity as the distance to which the survey is *complete*. All surveys will be complete over a larger volume for high L objects compared to low L. Our surveys of high mass stars are complete over a huge chunk of the Milky Way but our surveys of faint brown dwarfs are only complete to distances of ~ 20 pc, a fraction of the Galaxy's thickness.

3.1 *The Initial Mass Function*

Stars from the cold component of the ISM, specifically, they form in $T \sim 10$ K clouds where the chemical state of hydrogen is molecular so most of the material is in the form of H_2 and He atoms. In such clouds, the gas pressure forces are sufficiently small that they are unable to oppose the gravitational self-attraction of the gas, allowing it to collapse into stars. The material in the cold interstellar medium are shaped by a variety of physical effects, in order of importance, these are turbulence in fluid flows, gravitation, and then magnetic fields. The modern theory of star formation holds that turbulence drives a set of high density fluctuations, which then undergo gravitational collapse to form a distribution of stellar masses. The characteristic scale for star formation is of order the Jeans mass, which is the characteristic scale of gravitational fragmentation in a gas cloud. To estimate this scale, we first consider the *free-fall collapse time* for a uniform, spherically symmetric cloud of gas of mass density ρ and radius R such that the total mass of the cloud is $M_c = \frac{4\pi}{3} R^3 \rho$. In the absence of pressure support, this gas cloud will collapse under the force of gravity going from a radius R to a radius of o in time:

$$t_{\rm ff} = \left(\frac{3\pi}{32 G\rho}\right)^{1/2} \tag{3.1}$$

$$= 3.4 \text{ Myr} \left(\frac{n_{\rm H_2}}{10^8 \text{ m}^{-3}}\right)^{-1/2}. \tag{3.2}$$

Here, we have scaled ρ with the number density of hydrogen molecules, anticipating a truism of the star formation process, namely that star formation occurs in cold clouds of molecular hydrogen gas. These clouds also contain helium in atomic form (noble gas and all that).

As briefly stated, this collapse time is relevant when there is no pressure support. Under most interstellar conditions, this is not true and pressure support will stop gravitational collapse of the gas cloud, halting the process of star formation. We can derive a quick estimate of the mass and length scales involved in gravitational collapse given

this basic setup, though a more formal proof gives a more accurate answer.

For star formation to proceed, the gas must collapse faster than the time it takes for a pressure wave to traverse the object. This pressure wave crossing time is set by the sound speed in the gas cloud, c_s. If we set these two timescale equal to each other, we get the size of the cloud (λ_J) at which gravitational collapse can occur. We call this scale the Jeans[1] length.

[1] No apostrophe; this is named after a scientist whose last name was "Jeans."

$$t_{\rm ff} = t_c \tag{3.3}$$

$$\left(\frac{3\pi}{32G\rho}\right)^{1/2} = \frac{\lambda_J}{c_s} \tag{3.4}$$

$$\lambda_J = \left(\frac{3\pi c_s^2}{32G\rho}\right)^{1/2} \tag{3.5}$$

$$= \left(\frac{3\pi kT}{32Gm\rho}\right)^{1/2} \tag{3.6}$$

The free-fall timescale is an essential timescale in astrophysics and is fundamental timescale for the star formation process.

To find out the mass enclosed in this length scale we set this scale equal to the diameter of a spherical cloud of mass density ρ to derive the Jeans mass.

$$M_J = \frac{4\pi}{3}\rho\left(\frac{\lambda_J}{2}\right) \tag{3.7}$$

$$= \frac{\pi}{6}\rho\left(\frac{3\pi c_s^2}{32G\rho}\right)^{3/2} \tag{3.8}$$

$$= \frac{\pi}{6}\left(\frac{3\pi}{32}\right)^{3/2}\frac{c_s^3}{G^{3/2}\rho^{1/2}} \tag{3.9}$$

The scalings here are correct, but the prefactor is inaccurate. A more careful derivation gives

$$M_J = \frac{\pi^{3/2}}{8}\frac{c_s^3}{G^{3/2}\rho^{1/2}} \tag{3.10}$$

$$= 6.9\ M_\odot\left(\frac{T}{10\ \mathrm{K}}\right)^{3/2}\left(\frac{n_{\mathrm{H}_2}}{10^8\ \mathrm{m}^{-3}}\right)^{-1/2}. \tag{3.11}$$

where c_s is the speed of sound in the gas, ρ is its mass density and we have again scaled the result to typical conditions in a star forming molecular cloud. In this scaling, we have also used isothermal sound speed in a gas, which is

$$c_s = \sqrt{\frac{kT}{m}} \tag{3.12}$$

where m is the typical mass of a particle in the gas. For a gas cloud of *molecular* hydrogen, this value is $\sim 2.4 m_{\rm H}$ because all of the hydrogen

is H_2 and the helium is atomic and has a mass of $\sim 4m_H$. This leads to a typical scale[2] of $c_s = 0.2$ km/s $\sqrt{T/\mathrm{K}}$.

[2] Fun fact: this is quite close to the speed of sound in air, but the conditions are warmer ($T = 300$ K) and the gas particles are heavier with a mix of oxygen and nitrogen molecules such that $m \approx 30m_H$.

The gas motions in molecular clouds have typical speeds of > 2 km/s, which means the gas motions are supersonic, which generates turbulence and breaks up the collapsing clouds smaller. The characteristic mass of fragmentation turns out to be smaller than the Jeans mass by a factor of the sonic Mach number for turbulent flows, which is ~ 10 leading to a typical mass of a star on order $M = 0.5\ M_\odot$, which is the average mass of a star in most stellar systems.

This scaling of the Jeans mass with temperature through the sound speed also indicates why star formation preferentially occurs in cold clouds of gas. In warmer regions (say $T = 10^4$ K), the size and mass scales would be larger, requiring these larger objects to collapse to form stars. The pressure forces and other effects like turbulence and tidal force end up disrupting this larger scale collapse efficiently preventing the formation of massive objects.

To summarize, stellar mass objects fragment from clouds of molecular gas, collapsing under their own gravity to form stars on the timescale of a few Myr. There are a few other important amendments to this summary. First, star formation is highly inefficient. Only a small fraction of a molecular cloud ends up turned into stars. Most of the mass is dissipated from the injection of energy and momentum from the radiation of newly formed stars as well as protostellar jets. This is frequently described in terms of the efficiency per free-fall time. One can parameterize the star formation rates of gas clouds $\dot{M}_\star$ in different systems as $\dot{M}_\star = \epsilon_{\mathrm{ff}} M_{\mathrm{cloud}} / t_{\mathrm{ff}}$ where t_{ff} is the free-fall time for the gas cloud, M_{cloud} is the initial cloud mass and ϵ_{ff} is an efficiency parameter that describes all the glorious physics of the star formation process with $\epsilon_{\mathrm{ff}} \sim 1\%$ in most systems.

3.1.1 *The Initial Mass Function*

A rough characteristic scale for stars to form is just governed by the Jeans mass, but the ensemble of fluctuations leads to a range of collapse timescales and masses. This leads to distribution of stellar masses being produced by the process of star formation called the initial mass function (IMF). Empirically, our estimates of the IMF reveal that it appear to be relatively uniform in the regions where it can be measured in detail. The process of determining the IMF is extremely detail oriented. Broadly speaking, this works by counting the number stars at different luminosities and ages and using our knowledge of stellar evolution to map these backwards to the mass distributions at the time of formation.

There are a suite of observational data and a set of empirically determined distribution functions. The original work was completed and named after Salpeter, leading to the simplest form of the IMF:

$$\frac{dN}{d\mathcal{M}} \propto \mathcal{M}^{-2.35} \tag{3.13}$$

Here, the mass is expressed in units of the stellar mass relative to a solar mass: $\mathcal{M} \equiv M/M_\odot$. The IMF formally runs over the mass range of stars but substellar mass objects (brown dwarfs) appear to form part of a continuous mass distribution below the hydrogen burning limit $\mathcal{M} \approx 0.08$. The upper mass limit of stars is poorly determined and the IMF is these regimes is also unknown. Regardless, all IMFs predict them to be relatively rare, so we tend to run the limits up to $\mathcal{M} \sim 200$ but not worry too much about these results.

The Salpeter IMF is bottom-heavy relative to observations, specifically it predicts too many low mass stars relatively to the observed values. To address the observed dearth of lower mass stars, two other empirical forms have been used to represent the observational data, named after their original proponents. These are the Chabrier and Kroupa IMFs where the Charbrier IMF is

$$\frac{dN}{d\mathcal{M}} \propto \begin{cases} \exp\left[-\frac{(\log\mathcal{M}-\log 0.22)^2}{2\times 0.57^2}\right], & \mathcal{M} < 1 \\ \exp\left[-\frac{(-\log 0.22)^2}{2\times 0.57^2}\right]\mathcal{M}^{-2.35}, & \mathcal{M} \geq 1 \end{cases}, \tag{3.14}$$

and the Kroupa IMF is

$$\frac{dN}{d\mathcal{M}} \propto \begin{cases} \left(\frac{\mathcal{M}}{\mathcal{M}_0}\right)^{\alpha_0}, & \mathcal{M}_0 < \mathcal{M} < \mathcal{M}_1 \\ \left(\frac{\mathcal{M}_1}{\mathcal{M}_0}\right)^{\alpha_0}\left(\frac{\mathcal{M}}{\mathcal{M}_1}\right)^{\alpha_1}, & \mathcal{M}_1 < \mathcal{M} < \mathcal{M}_2 \\ \left[\prod_{i=1}^{n}\left(\frac{\mathcal{M}_i}{\mathcal{M}_{i-1}}\right)^{\alpha_{i-1}}\right]\left(\frac{\mathcal{M}}{\mathcal{M}_n}\right)^{\alpha_n}, & \mathcal{M}_n < \mathcal{M} < \mathcal{M}_{n+1} \end{cases}, \tag{3.15}$$

with

$$\begin{aligned} \alpha_0 &= -0.3 \pm 0.7, \quad \mathcal{M}_0 = 0.01 \\ \alpha_1 &= -1.3 \pm 0.5, \quad \mathcal{M}_1 = 0.08 \\ \alpha_2 &= -2.3 \pm 0.3, \quad \mathcal{M}_2 = 0.5 \\ \alpha_3 &= -2.3 \pm 0.7, \quad \mathcal{M}_3 = 1, \ \mathcal{M}_4 \to \infty \end{aligned} . \tag{3.16}$$

Both of these IMFs have a break at lower masses to become shallower relative to the Salpeter IMFs. Note, however, that all IMFs predict that the slope is a power law in mass with index of about $\mathcal{M}^{-2.3}$.

Figure 3.2 shows the variation in the shape of the IMF for these three different functional forms. At the high mass end, they all converge to a similar shape, but the low-mass ends show a turnover to smaller values. The exact shapes of the IMF, particularly in the substellar regime are an ongoing research problem.

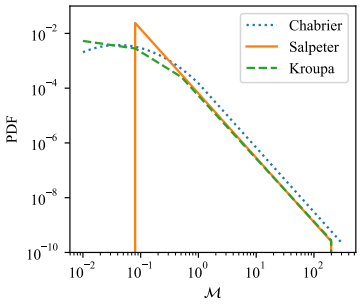


Figure 3.2: Comparison of IMF shapes.

The variation in the low-mass end means that the emergent light does not vary significantly for a given population of stars dependent

on the IMF, since this is dominated by the emission of the luminous high mass stars. However, the total stellar mass required to produce that light does vary significantly. To illustrate this, Figure 3.3 shows the emerged spectrum of light from a stellar population synthesis model [Conroy and Gunn, 2010] that is designed to reproduce unresolved stellar populations. The spectrum is given at emergent flux density per unit mass of stars initially formed in the population. The spectrum is shown after a time of $\tau = 10$ Gyr, which is comparable to the main sequence lifetime of the Sun. The Salpeter IMF is significantly less luminous per unit mass of stars formed, reflecting the bottom-heavy nature of its initial mass function.

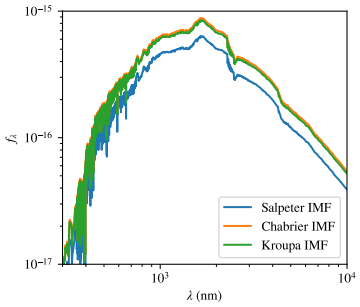


Figure 3.3: Variation in the emergent spectral energy distribution per solar mass of stars formed predicted by assuming different stellar IMFs. The plots are shown after $\tau = 10$ Gyr of evolution for the stellar population While the Kroupa and Chabrier IMFs are similar the Salpeter IMF is significant less luminous per unit mass.

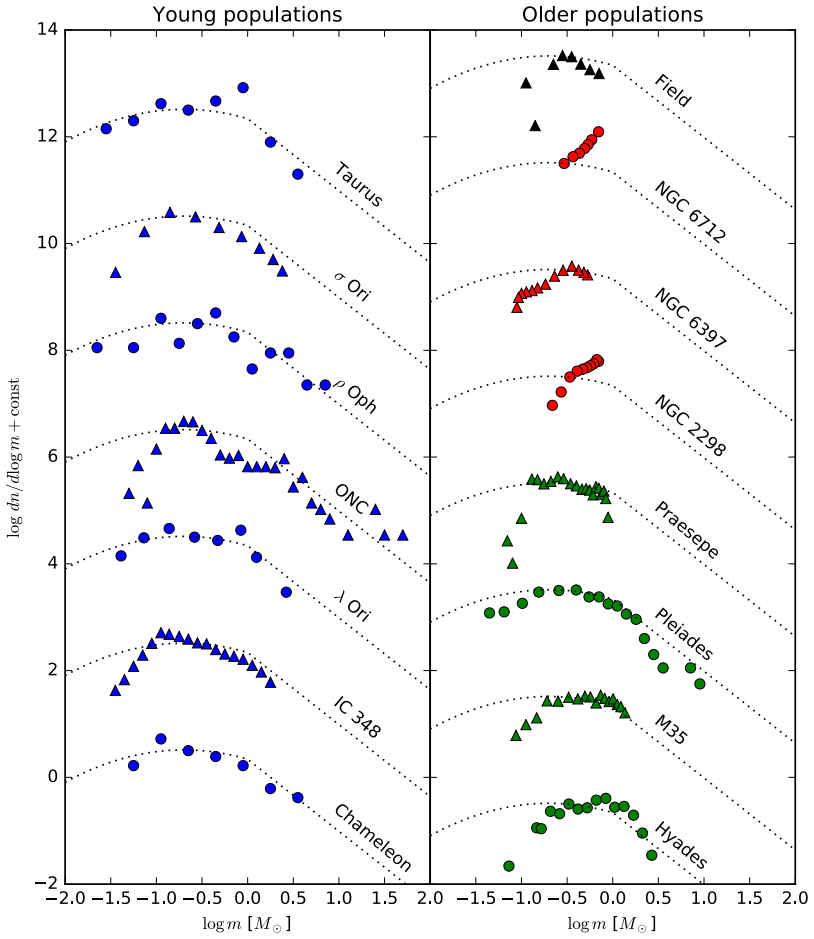


Figure 3.4: The initial mass function from several nearby clusters. The data are compiled from Bastian et al. [2010] and the figure uis from Krumholz [2015].

Local observations suggest that the Kroupa or Chabrier IMFs are good representatives of the field star population in the disk of the Milky Way. Figure 3.4 show the mass functions in several young and

older clusters in the solar neighbourhood. The distribution functions of star mass are all compared to a single mass function background, which is similar to the Chabrier IMF. Despite the vastly different conditions and mass scales of these clusters, the masses that are formed all approximately follow the same IMF.

Modern star formation theory holds that the Jeans mass should scale like the temperature and density of the medium that is collapsing to form stars. Moreover different molecular clouds show variations in the Mach number and magnetic field strength, even in the disk of the Milky Way. Given most theories that predict the initial mass function depend on these parameters, it is actually surprising that the IMF is relatively constant in both time and in space for the population of stars in the disk. Broadly speaking, if the IMF did vary as much as the parameters that govern the Milky Way molecular cloud population, we should have seen the IMF show significant and large changes. Why the IMF remains so constant is an ongoing theoretical puzzle in the study of star formation.

As such, it should not surprise us when we see significant variations in the initial mass function. There are significant lines of evidence that argue for variations in the initial mass function in different environments. However, every piece of evidence is not without its debate. Some studies find top-heavy initial mass functions, with an overabundance of massive stars. These are typically found in young massive clusters of stars, such as the recent results argued for in the cluster 30 Doradus [Schneider et al., 2018] Such a top-heavy IMF would produced comparatively more radiative luminosity and mechanical feedback into a galaxy in the form of supernova explosions and stellar winds. This fits a nice picture since 30 Doradus is in the Large Magellanic Cloud which is relatively poor in metals, so the ISM should cool less effectively, leading to a larger Jeans mass for collapse. However, there are ample concerns that such surveys of the high mass stars incorrectly account for the fraction of multiple stars. Unresolved binary stars would make a pair of stars appear as a single brighter star. Erroneously counting binary systems as single high mass stars would lead to an overestimate the number of high mass stars and thus incorrectly report a top-heavy tail. Even if individual clusters do show evidence for a top-heavy IMF, it is not clear that this holds for all clusters so that the average over the population of star forming events may still yield the observed IMFs.

Some systems show evidence for a bottom-heavy IMF, or an excess of low mass stars. The standard evidence for this comes from analysis of the mass-to-light ratio of a system such as a globular cluster. When the mass can be inferred through other means (such as the virial theorem) and the light counted directly, we can make a direct

measurement of the mass to light ratio. This can be compared to the expected mass-to-light ratio expected for a system formed with different IMFs. If there is an excess of mass in the system relative to these models, one way of explaining this is to have an excess of low mass stars. These stars contribute more to the mass than they do to the light. Unfortunately, there are a myriad of other ways of increasing the mass-to-light ratio and the foundational measurements of dark matter's existence come from unexpected mass-to-light ratios. There are more compelling spectroscopic based measurements from van Dokkum and Conroy [2010] that show an excess of spectral features associated with main-sequence low-mass stars in elliptical galaxies compared to the light seen ordinary stellar populations.

The IMF is one of the basic ingredients for a galaxy and its variations are an important unknown that sets how galaxies evolve.

3.1.2 *Averaging over the IMF*

The IMF is fundamentally a probability density function, which means that we can use it to infer the properties of stellar populations in aggregate. It describes how a given total mass of stars can be related to the individual stars that form. A few examples can help illustrate the utility and meaning of these functional forms. Like all density functions, the IMF is meant to be integrated.

Number of Stars – We can use the IMF to find out the expected number of stars in a given mass range in an IMF. For example, if we want to know how many stars will produce supernova explosions after $M_\star = 10^4\ M_\odot$ of stars form, we can use the IMF. We will take the Salpeter IMF running from $0.1 < \mathcal{M} < 100$ for simplicity and consider stars that undergo supernova explosions as having $\mathcal{M} > 9$. The total value of $M_\star$ allows us to turn Equation 3.13 from a proportionality into an equality:

$$\frac{dN}{d\mathcal{M}} = c_\star \mathcal{M}^{-2.35}. \tag{3.17}$$

We can integrate up the total mass of stars in the system, counting the number of stars at a mass $\mathcal{M}$:

$$M_{\rm tot} = \int_{0.1}^{100} \mathcal{M} \frac{dN}{d\mathcal{M}} d\mathcal{M}. \tag{3.18}$$

This allows us find the coefficient $c_\star$ for $\mathcal{M}_{\rm tot}$:

$$\begin{aligned}
\mathcal{M}_{\rm tot} &= \int_{0.1}^{100} \mathcal{M} \cdot c_\star \mathcal{M}^{-2.35} d\mathcal{M} \\
&= c_\star \left(\int_{0.1}^{100} \mathcal{M}^{-1.35} d\mathcal{M} \right) \\
&= c_\star \left(\frac{1}{-0.35} \mathcal{M}^{-0.35} \Big|_{0.1}^{100} \right) \\
&= c_\star \left(\frac{1}{0.35} \mathcal{M}^{-0.35} \Big|_{100}^{0.1} \right) \\
&= \frac{c_\star}{0.35} \left(0.1^{-0.35} - 100^{-0.35} \right) \\
c_\star &= \frac{0.35 \cdot 10^4}{0.1^{-0.35} - 100^{-0.35}} \\
c_\star &= 1716. \qquad (3.19)
\end{aligned}$$

Now, given $c_\star$, we can just integrate the IMF over $9 < \mathcal{M} < 100$:

$$\begin{aligned}
N_{\rm SN} &= \int_9^{100} \frac{dN}{d\mathcal{M}} d\mathcal{M} \\
&= c_\star \int_9^{100} M^{-2.35} d\mathcal{M} \\
&= c_\star \left(\frac{1}{-1.35} \mathcal{M}^{-1.35} \Big|_9^{100} \right) \\
&= c_\star \left(\frac{1}{1.35} \mathcal{M}^{-1.35} \Big|_{100}^{9} \right) \\
&= \frac{1716}{1.35} \left(9^{-1.35} - 100^{-1.35} \right) \\
&= 62.9.
\end{aligned}$$

Thus, when $M = 10^4\ M_\odot$ worth of stars form, there are about 63 supernova explosions. The details of these calculations will change when using a Chabrier or Kroupa IMF, becoming more mathematically complex, but the IMF is incredibly powerful in its ability to make these measurements.

AVERAGING OVER OTHER PROPERTIES – The IMF can also be used to average over quantities other than mass. For example, the mass-luminosity relationship can be used to determine the total luminosity of the stellar system. If we use the simplest form of the relationship: $L = 1\ L_\odot (M/M_\odot)^{3.5}$, we can calculate the total luminosity of the $M = 10^4\ M_\odot$ worth of stars in the previous problem, assuming all

stars are on the main sequence.

$$\begin{aligned}
L_{\text{tot}} &= \int L\, dN \\
&= \int_{0.1}^{100} L \frac{dN}{d\mathcal{M}} d\mathcal{M} \\
&= 1\ L_\odot \int_{0.1}^{100} \mathcal{M}^{3.5} \frac{dN}{d\mathcal{M}} d\mathcal{M} \\
&= 1\ L_\odot \int_{0.1}^{100} \mathcal{M}^{3.5} c_\star \mathcal{M}^{-2.35} d\mathcal{M} \\
&= 1\ L_\odot c_\star \int_{0.1}^{100} \mathcal{M}^{3.5} \mathcal{M}^{-2.35} d\mathcal{M} \\
&= 1\ L_\odot c_\star \int_{0.1}^{100} \mathcal{M}^{1.15} d\mathcal{M} \\
&= 1\ L_\odot c_\star \frac{1}{2.15} \left(\mathcal{M}^{2.15} \Big|_{0.1}^{100} \right) \\
&= 1\ L_\odot (1716) \frac{1}{2.15} \left(100^{2.15} - 0.1^{2.15} \right) \\
&= 1.6 \times 10^7\ L_\odot.
\end{aligned}$$

That's a lot of luminosity! While the approach is sound, the estimate is an overestimate because the mass-luminosity relationship becomes significantly shallower above $\mathcal{M} > 10$ as per Equation 2.9. In this simplified example, the high mass stars contribute a large amount of luminosity. Repeating this process with the Kroupa IMF and the composite Mass-Luminosity relationship gives a value that is lower: $8.1 \times 10^6\ L_\odot$. This is entirely dominated by the high mass stars. Since a single $M = 50\ M_\odot$ star has a luminosity of $570\,000\ L_\odot$, it just takes a few stars to add up to a large luminosity.

Clusters don't stay this bright for long. The main sequence lifetimes of the high mass stars contributing light is short (3 Myr), so the cluster quickly fades away. If we instead consider the brightness of the cluster after all the $> 10\ M_\odot$ stars have died, we can change the upper bound of the previous integral to $\mathcal{M} = 10$ from 100 and the luminosity drops precipitously to $L_{\text{tot}} = 1.1 \times 10^5\ L_\odot$. Note that we do not change $c_\star$. That is set by the total number of stars that have formed.

3.1.3 Binary and Multiple Stars

An additional wrinkle is that stars are frequently found in binaries, and the presence of stellar companions will influence a star's evolutionary history. This in turn can affect a galaxy's synthesized light and possibly also how those stars shape the rest of material in the galaxy, notably the interstellar medium. In addition to the IMF, the star formation process must set the origins of binary systems. Out-

side of rich clusters, stellar collisions are infrequent, so most binaries are formed together. These include (1) the stellar multiplicity, which measures the number of stars found a bound system $\langle N_\star \rangle$, (2) the mass ratio of the stars q which is defined as M_2/M_1, and (3) the distribution of semi-major axes a or equivalently through orbital dynamics, the period P.[3] The multiplicity is also sometimes defined in term of the companion fraction, which is $C = \langle N_\star \rangle - 1$.

[3] These variables are common since the orbits are frequently described in terms of Kepler's third law, with $P^2 = 4\pi^2 a^3/[G(M_1 + M_2)]$.

These parameters are observed to be functions of the total mass in the system and also are thought to depend on the age of the system. The time evolution is interesting because some multiple star systems frequently form in unstable configurations and end up ejecting one or more members of the system. Stellar evolution effects can also change the mass ratios and even the orbital periods. The work of Moe and Di Stefano [2017] presents a meta-analysis of a suite of different observations of these parameters and leads to a few important conclusions for binary systems. For solar-type primary stars, 50% of the systems are binary or multiple star systems, however this fraction rises to 95% for the most massive stars. Of the non-single high-mass stars, 75% of these are in multiple star systems. The highest mass stars ($M > 16\ M_\odot$) are found in close binary systems far more often of the time compared to low mass stars (30% vs 3%). Finally, close binary stars tend to be "twins" having masses that are indistinguishable from each other. The fraction of twins in close binaries ($P \sim 1$ day) of solar-mass stars is 30% vs. 10% for high-mass stars.

The precise details of all these facts are interesting and tells us a great deal about how stars form from molecular clouds and the deposition of angular momentum in these systems. The same physics must also set the conditions for planet formation, though that's even farther out of scope for this class. Despite the general awesomeness of this information, these details should instill within us some skepticism of the details of population synthesis models and some determinations about the variation in the IMF. These results are solid to first order, but the making precision measurements on these topics remains an area of active research.

Key Points

- Stars form in cold clouds of molecular hydrogen (H_2). They form in molecular because these clouds have the smallest Jeans masses (Equation 3.11) and shortest free-fall time (Equation 3.6) of any phase of the interstellar medium.
- The star formation process produces a remarkably uniform distribution of initial stellar masses called the initial mass

function (IMF). While there are several different functional forms for the IMF (Equations 3.13, 3.15, 3.14), they all have similar functional forms (Figure 3.2).

- The general shape of the IMF is that star formation produces dramatically fewer high mass stars compared to low mass stars. However, the luminosities of these stars are so high that, while the exist, they dominate the light of stellar populations.
- The IMF is a probability density function for stellar masses and can be used to find the numbers, masses, and average properties of stellar populations.
- The IMF refers to all stars formed in the star formation process. Star formation produces binary and multiple star system where the relative masses of stars can be different from the IMF. For example, high mass stars are more likely to be in close binary systems, i.e., the binary has a smaller orbital radius.

3.2 *Simple Stellar Populations*

The IMF and its ability to track the average properties of the stellar population at birth gives some indication of how we could begin to consider stellar populations. Our next step will be to combine the knowledge of how individual stars evolve with the IMF. We begin this exploration in the context of *simple stellar populations* (SSP). An SSP is defined as a group of stars that formed at the same time from gas that is the same metallicity. Given our understanding of the star formation process, we assume that the molecular clouds hosting star formation are well mixed so that the metallicity distribution of the gas is homogeneous. We also note that the star and cluster formation process is relatively rapid compared to the main sequence lifetimes of most stars. As is true throughout their lives, high mass stars form more quickly than low mass stars (0.01 Myr at the high mass end but several Myr at the low mass end). Even so, since most stellar lifetimes are significantly longer than a Myr, formation is rapid and we can consider stars formed in a molecular cloud to be a simple stellar population. However, the timescales for cloud evolution are uncomfortably close to the lifetimes of high mass stars, so relationship of SSPs to the star formation process at very early times is tenuous.

The most commonly assumed SSP is that of a *stellar cluster*. Clusters form in molecular clouds as a group of stars in a larger, ap-

proximately[4] gravitationally bound cloud of gas with a bound mass significantly larger than the stars forming inside of it. The bound gas structure eventually dissipates through the star formation process and the resulting cluster may or may not be bound depending on the rate at which the progenitor molecular gas is pushed out of the star forming region by feedback. Clusters form a set of stars with a mass function that is close to the IMF (Figure 3.4), which then begin to evolve all together. Hence, we can examine the properties of this cluster in the HR diagram and colour-colour diagrams to understand it changes. We see snapshots of the stars all at the same time from their formation, but because the stars have different evolutionary timescales they are found at different parts of the diagram. For example at 200 Myr since birth, a 1 $M_\odot$ star will still be on the main sequence but a 5 $M_\odot$ star will be on the helium burning sequence. We can synthesize this behaviour across all mass using the study of isochrones.

[4] This cautionary "approximately" is inserted because it can be quite difficult to assess from observations where a cloud of gas is bound to itself. But, using simple assumptions, the cluster-forming clouds sure do look gravitationally bound.

3.2.1 *Isochrones*

An *isochrone* is the locus (set of points) in an HR diagram that shows where stars in a simple stellar population will be found. Isochrones (in terms of Latin roots: *iso*=same, *chron*=time) are complementary to stellar evolutionary tracks shown in, e.g., Figure 2.13. Evolutionary tracks show one star of a fixed mass at different times. Isochrones show stars of different masses at the same fixed time.

Figure 3.5 shows a set isochrones for a stellar population with solar metallicity (formally [Fe/H]=0.0). These run from very early in the star formation process with age $\tau = 1$ Myr such that $\log_{10}(\tau/\mathrm{yr}) = 6$ up to 10 Gyr, with the colours corresponding the age of the track. Isochrones look a lot like stellar evolutionary tracks, even though they are tracing out stars with different masses because the time that a star spends off the main sequence is relatively short compared to the time that it is on the main sequence. As such, the stars on the RGB/HB/AGB are all about the same mass and undergoing similar behaviour so it looks a little like an evolutionary track. Examining the patterns in the isochrones, we can see a few features. First, for the youngest isochrones, the lowest mass stars are not yet present on the main sequence. This reflects that high mass stars evolve more quickly than low mass stars through protostellar evolution. We can also see that stars do not populate the RGB until about 100 Myr (10^8 years).

The most visible feature on most isochrones is the rather sharp change between stars on the main sequence those evolving onto the RGB. The location of stars at their TAMS is indicated on Figure 3.5 with a circular marker. While the isochrones continue on

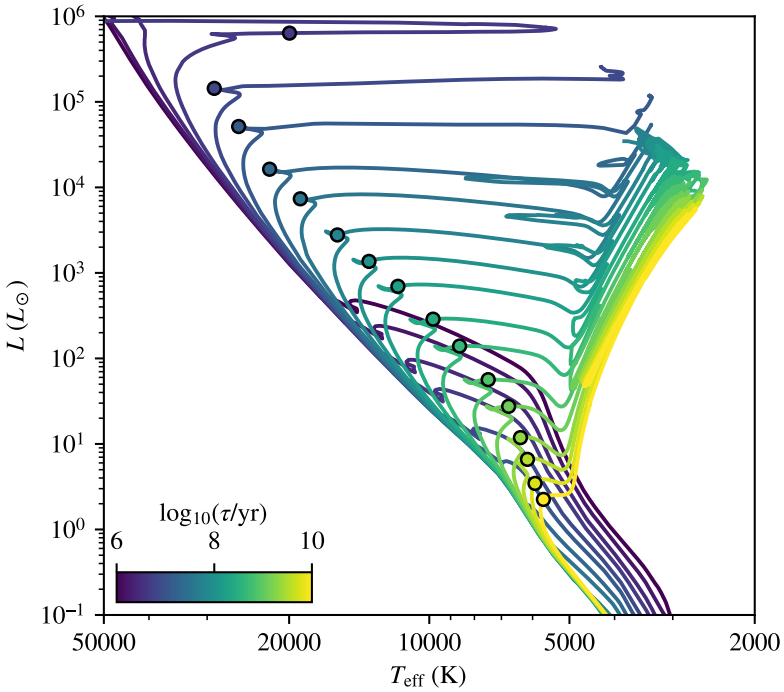


Figure 3.5: Isochrones for a solar metallicity population.

toward lower surface temperatures at this point, their rate of evolution through after this point is rapid, so these sections of the HR diagram tend to be sparsely populated. This feature is called the Main Sequence Turn Off (MSTO). This feature is particularly important in age-dating clusters: the age of the simple stellar population is equal to the main sequence lifetime of the most massive star on the main sequence. Stars more massive than the stars at the MSTO have already evolved into their later stages and stars less massive than the MSTO will have evolved more slowly and will still be on the main sequence.

Isochrones can also be mapped into observable spaces like the Gaia HR diagram, transforming luminosities and temperatures into Gaia colours and absolute magnitudes. This changes the shape of the isochrones but not the overall patterns. Figure 3.6 illustrates a subset of the isochrones from Figure 3.5 mapped into the colour bands of the Gaia satellite. The same basic structure is visible in these isochrones as is apparent in the Gaia colour-magnitude diagram shown in Figure 2.7.

Finally, like evolutionary tracks, isochrones show significant variations with the metallicity of the simple stellar population. In Figure 3.7, we show how isochrones in the colour-magnitude diagram of

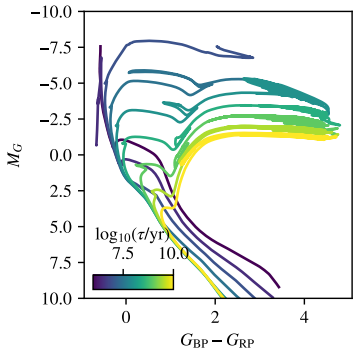


Figure 3.6: Isochrones as per Figure 3.5 except for Gaia observed filters.

the Gaia satellite would change as a function of metallicity. These isochrones show the positions of a 100 Myr stellar populations with metallicity of $[\mathrm{Fe/H}] = \{-4.0, -3.5, \cdots, 0.0, +0.5\}$. Recall that [Fe/H]=0 on this scale corresponds to solar metallicity. We see that populations with lower metallicity tend to look bluer and brighter (more negative M_G) compared to populations with higher metallicity. This reflects the same trend that we saw in the evolutionary tracks for stars, namely that individual stars with low metallicity were more luminous and had higher surface temperatures compared to solar metallicity stars.

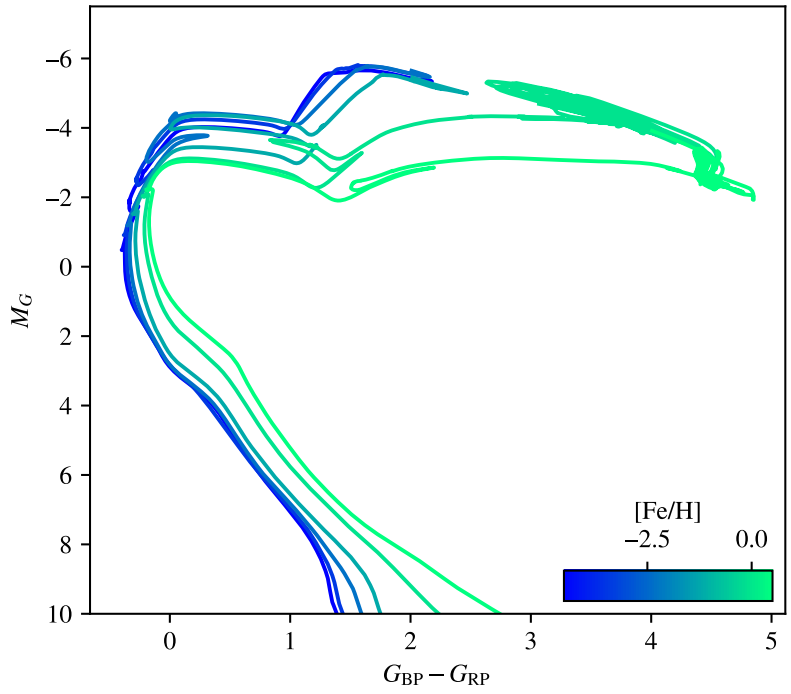


Figure 3.7: Variations of the Gaia colour-magnitude diagram with metallicity measured in [Fe/H] units.

3.2.2 *Isochrone fitting*

With the tools of isochrones, we can start to examine systems that we believe are simple stellar populations, like young clusters of stars. Again, data from the Gaia satellite is particularly high quality in identifying the properties of clusters. The method follows a few steps as outlined in Gaia Collaboration et al. [2018] which consist of finding stars that are in a common direction in the sky that share a similar proper motion. These stars formed together and are initially moving through the Galaxy together on a trajectory set by their natal molecular cloud. Eventually tidal forces from the rest of the matter

in the Galaxy will disrupt the clusters, but since they are young, their initial motions are still well aligned.

Figure 3.8 shows the result of the process for the cluster Praesepe. The colour-magnitude diagram shows a well defined main sequence, some evolved stars (RGB/HB/AGB) and some white dwarfs. The presence of binary stars is visible as a set of points displaced upward from the main sequence by up to $\Delta M_G = 0.753$, which is the magnitude offset for a stellar system that is at the same colour but twice as bright as a single main sequence star. The best fitting isochrone for this population is also overlaid on the CMD. Isolated stars on the main sequence and on the Red Giant branch follow the isochrone. Notably there are several stars on the main sequence up to the MSTO of the cluster, and then very few stars between that point and the RGB. The properties of the stars at the MSTO are one of the prime factors in setting the age of the cluster with $\log(\tau/\mathrm{yr}) = 8.85$ or $\tau = 710$ Myr. The value of $Z = 0.020$ is the solar value so that this would correspond to $[\mathrm{Fe/H}] = 0.0$ in our plots of isochrones with varying metallicities.

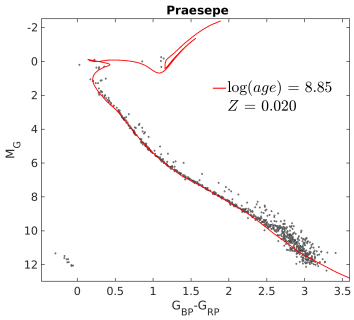


Figure 3.8: HR diagram for Praesepe from Gaia Collaboration et al. [2018] with isochrone fit.

3.2.3 *Unresolved Simple Stellar Populations*

We can also consider all the light from stars in an aggregate sense. Instead of splitting things up and calculating individual stars and their position in a colour-magnitude diagram, we can examine the colours of the stars added up across the population. Because of the strong scaling of luminosity with initial stellar mass, the light from a stellar population is usually dominated by the most massive stars on the main sequence as well as the stars on the RGB/TP-AGB. Figure 3.9 shows the evolution of the emergent SED at different times. The plot uses a strange y-axis, which is the luminosity-to-mass ratio Υ_λ as a function of wavelength as a function of time.

The plot encapsulates several key features about stellar SEDs over the course of time. First, younger stellar populations have vastly higher bolometric (integrated over wavelength) luminosities than do older stellar populations. For example, the bolometric luminosity at $\tau = 10$ Myr is 100 $L_\odot$ per solar mass but at $\tau = 10$ Gyr is only 0.2 $L_\odot$ per solar mass. Only the youngest phases of the stellar population emit significantly at $\lambda < 91.2$ nm, which corresponds to the ionization potential of atomic hydrogen. Photons with such short wavelengths are said to be in the Lyman continuum. They have short mean-free-paths in the ISM and typically ionize neutral gas in the immediate vicinity around these newly formed stars. After $\tau \approx 20$ Myr, the ionizing photon production of a stellar population drops precipitously, though the stellar population remains blue for significantly

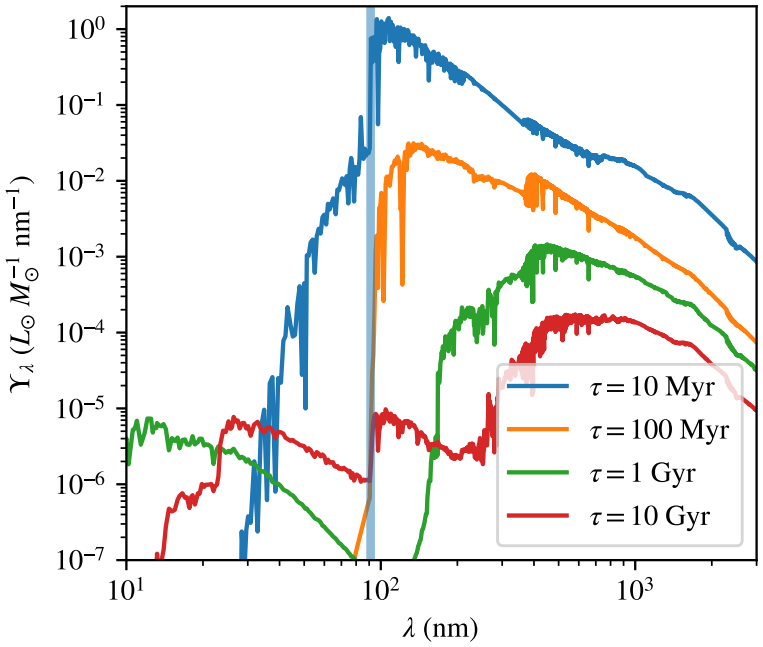


Figure 3.9: Variation of the emergent SED from a $Z = Z_\odot$ simple stellar population with age since burst. Light emitted at wavelength shorter than the shaded vertical line indicates the boundary of ionizing photons at 91.2 nm. The model has no dust absorption or nebular emission.

longer. This can be seen more clearly in Figure 3.10 as shown in the Figure below, which highlights the SDSS g and r filters in the spectrum, which are commonly used to describe galaxies in surveys. The changing flux ratio in these respective regions is reflected in the changing $g - r$ colour of the spectrum.

The overall SED 3.9 also shows several stellar absorption features clearly, which can also be seen in 3.10. Most of these stellar absorption features can be attributed to the absorption lines of hydrogen but also some metals in atmospheres of the stars. The Balmer series and Lyman series are particularly prominent. These series show progressively more closely spaced line limiting to a continuum feature at $\lambda = 364.4$ nm and the Lyman continuum reaches the aforementioned continuum limit at $\lambda = 91.2$ nm.

One deviation from the general trend of stellar populations aging to become "red-and-dead" is that the 10 Gyr stellar populations still emit significantly in the highest energy photons. These ultraviolet and x-ray portions of the spectrum come from the treatment of old stars aging off the asymptotic giant branch. Because there so many of these stars, this emission can be significant compared to the longer wavelength radiation.

While we don't have the details that we get from studying indi-

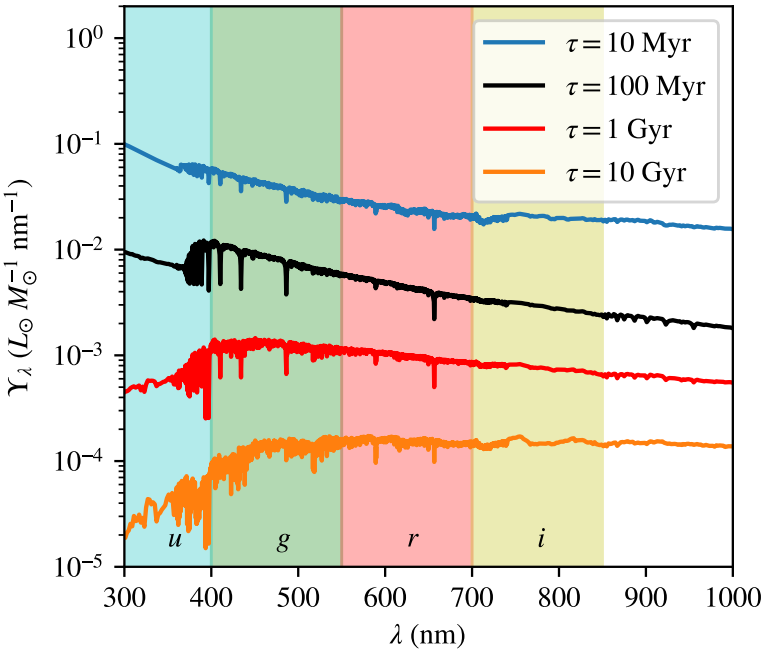


Figure 3.10: Stellar populations as per Figure 3.9 but only showing the optical portion of the spectrum. The nominal coverage of the SDSS bands are overlaid on the Figure. The evolution of the $g - r$ colour is clear in the figure.

vidual stars, the aggregate light of an unresolved stellar population still clearly reflects the evolutionary state of stellar population that generates it. By understanding stars and the IMF, we can infer a great deal about the age and metallicity of parts of galaxies from their integrated light.

Key Points

- A simple stellar population is a group of stars that formed together and thus all stars have the same age and metallicity but different masses.
- Simple stellar populations are found in the HR diagram along isochrones (Figure 3.5).
- As a population ages it becomes redder in colour and significantly less luminous (e.g., Figure 3.9).
- In an HR diagram, the main sequence turnoff indicates the age of a stellar population. The main sequence lifetime of the highest mass and luminosity stars found remaining on the main sequence is equal to the age of the stellar population.

- Isochrones also vary with metallicity with lower metallicity isochrones appearing bluer that high metallicity systems.
- Thus, fitting isochrones to stellar populations assumed to be simple (e.g., stellar clusters) will yield the metallicity and age of the population.

3.3 *Multiple Stellar Populations*

Outside of clusters, simple stellar populations don't exist. A galaxy is a superposition of different SSPs that have all been formed at different points in the past with different metallicities, and possibly even different IMFs. Morever, the stars in the galaxy's actual stellar population are not uniformly detectable: bright stars can be resolved at megaparsecs distances but a M8V star is barely detectable a few pc away. Finally, a galaxy is made of more than just the stars, and the intervening dust layers block out the light from background stars. Below, we discuss how real stellar populations are shaped by these different effects to ultimately produce the *field* star population, referring to the stars that are just part of a galaxy as a whole but not part of any individual cluster.

3.3.1 *Observational Effects*

The imperfect nature of telescopes means that the stellar population we observe is not the same as the true stellar population. The major effect here is that luminous stars are easier to detect at larger distances compared to fainter stars. This is because telescopes have a characteristic sensitivity at a roughly constant magnitude threshold or threshold flux density, below which we cannot detect object because they are too faint for the telescope to see. Observing a given part of the sky for longer will enable the detection of fainter sources (larger magnitudes), but this depth only reduces slowly with time: $m_{\rm limit} \propto 1.25 \log_{10} t$ where t is the "integration time" spent observing the source. Since this limiting magnitude for detecting a source is fixed, we can detect more luminous stars out to a larger distance, or equivalently, to a larger distance modulus. For a given limit, the distance modulus for detection of a source with absolute magnitude M is $m_{\rm limit} - M = 5 \log_{10}(d/10\ {\rm pc})$. For a star with a smaller M (i.e., a larger luminosity), the distance over which such a increases exponentially:

$$d_{\rm limit} = 10\ {\rm pc}\, {\rm dex}\left(\frac{m_{\rm limit} - M}{5}\right), \tag{3.20}$$

where dex is the decimal exponential operator[5]; i.e., $\mathrm{dex}\, y \equiv 10^y$.

This limiting magnitude or flux density sets the *survey volume* over which a survey is said to be *complete*. This volume is the three-dimensional volume within which a survey detects a fixed fraction of all the sources with a given (specific) luminosity. That fraction is usually set to be a reasonably high fraction like 90% or 99%. You will read expression like "the survey of AoV stars is 99% complete out to a distance of 3.2 kpc," meaning that 99% of all AoV within 3.2 kpc are identified within a spherical volume that has a radius of 3.2 kpc.

To make these arguments more concrete, consider the Gaia mission which will has a limiting magnitude in the main band of $G = 20$ (or better) over most of the sky. High-mass ZAMS stars between $40 < M/M_\odot < 50$ will have $M_G < -4.73$ (see Table 9.1) whereas low-mass stars with masses between $0.4 < M/M_\odot < 0.5$ will have $M_G < 9.63$. For a given field star population, the ratio q of the number of stars in the high mass bin compared to the low mass bin will be, taking a Salpeter IMF.

$$\begin{aligned} q &= \frac{\int_{40}^{50} c_\star \mathcal{M}^{-2.35} d\mathcal{M}}{\int_{0.4}^{0.5} c_\star \mathcal{M}^{-2.35} d\mathcal{M}} \\ &= 0.002. \end{aligned} \tag{3.21}$$

However, for these absolute magnitudes, distance limit for the high mass stars (Equation 3.20) gives $d_{\mathrm{limit,high}} = 883$ kpc but $d_{\mathrm{limit,low}} = 1.19$ kpc. The volume over which the high-mass stars could be detected is $(d_{\mathrm{limit,high}}/d_{\mathrm{limit,low}})^3 = 4.1 \times 10^8$ times larger than the volume over which the low-mass stars can be detected. If these stars were distributed evenly in this volume, then the Gaia survey would contain 800,000 times more stars with $40 < M/M_\odot < 50$ than stars with $0.4 < M/M_\odot < 0.5$. The larger luminosities of high mass stars more than make up for their relative rarity. Of course, looking at the Gaia HR diagram in Figure 2.7 shows that most of the stars on the main sequence aren't very high mass (O and B types) but rather stars like our Sun (G dwarfs). This arises from two effects: (1) high mass stars are short lived stars but the low-mass stars essentially last longer than the current age of the Universe. The main sequence lifetimes of the high mass stars stars are $10^5\times$ shorter than the low-mass interval so, assuming a constant star formation rate, q should be reduced by this factor to $q_{\mathrm{corr}} = 8$. The true value is even lower than this because our Galaxy is disk like with a radius of $R \sim 10$ kpc so the full volume over which we can see these stars is not completely filled with Galaxy.

This long example illustrates the main point here: the relative numbers of the actual stars we see in a survey are shaped both by the true underlying population of stars and the sensitivity of the

[5] We will sometimes use the shorthand that 1 dex is just an order of magnitude in an expression like "the bias is approximately 0.5 dex" meaning "half-an-order-of-magnitude" or $10^{0.5} \approx 3.16$

telescope with which the survey was carried out.

In unresolved populations, there is a related effect, namely that the light we observe from the population is dominated by the brightest stars in the population, weighted by the number of those stars present. The HR diagram may look like a simple, compact plot, because of the logarithmic nature of the magnitude system, but the luminosities of stars span orders of magnitude in brightness. These few stars at the "top" of the HR diagram set the colour of the light from a galaxy. For young galaxies, this is mostly the blue stars on the main sequence but once the population is > 100 Myr old, the population has a substantial amount of light from red giants as well.

3.3.2 *Dust*

The other main observational effect that shapes the properties of stars that we see is the presence of interstellar dust along the line of sight between us and the distant star. Overall, dust does three things to light, it (1) absorbs and (2) scatters relatively short wavelength light and it (3) re-emits the absorbed radiation as comparatively long-wavelength light, typically in the infrared. Below, we will give a quick introduction to what dust is, returning to a broader discussion of dust in the context of the interstellar medium (ISM; Chapter 4).

WHAT IS DUST? – Dust is a population of particles in the ISM. Dust is mostly comprised of metals (i.e., not hydrogen and helium). There are thought to be three main constituents of dust in the ISM driven by the need to fulfill the optical properties of the dust distribution. First, silicate and carbonaceous grains are solid pieces of matter. Silicates are basically rocks, consisting of Fe, Si, and O. Carbonaceous grains consist of C, O, and H in various compounds. Finally, we consider polycyclic aromatic hydrocarbons (PAHs) as a significant constituent of dust. Unlike the other two species, PAHs are really large molecules (10^2 to 10^3 hydrogen masses) rather than solid state objects, though they exist in the transition regime between solids and individual molecules.

Aromatic means that the molecule is flat and composed of carbon rings. Polycyclic means that there are more than one of those rings. Hydrocarbons means that the molecule is primarily made of H and C. That's all I know.

The big difference between these regimes are whether the objects are held together with molecular bonding mechanisms (ionic or covalent bonding) or weaker solid-state interactions like van der Waals forces (solids). The reasons we care about these grains are that the solid-state properties of these grains have associated normal modes for molecular bending and vibration that appear to match up with observations. For example, cross section for dust absorption appears to increase around $\lambda = 217.5$ nm, which is the resonant wavelength for a common bending mode in PAHs. There is an infrared feature

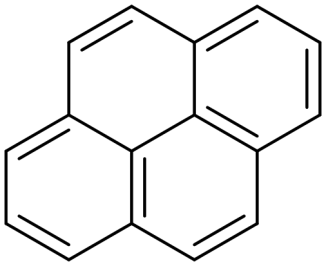


Figure 3.11: Pyrene is a polycyclic aromatic hydrocarbon. Following organic chemistry convention, the figure shows the bonds between carbon atoms and unaccounted for bonds are bonded to hydrogen atoms.

near $\lambda = 10$ μm associated with silicate normal modes.

The canonical distribution of dust grains follow a bottom-heavy size distribution for grains of characteristic radius r:

$$\frac{dN}{dr} \propto r^{-3.5} \tag{3.22}$$

meaning there are many more small grains than large grains. The power-law is bounded by grains with a size of $r_{\rm max} \sim 250$ nm and $r_{\rm min} \sim 5$ nm. Since the mass of a grain is proportional to its volume with only a small change due to volume density, most of the mass in the dust grains is concentrated in the largest grains. In contrast, the optical cross section is proportional to the area of the grain, i.e., r^2 so most of the absorption cross section is found in the small grains. The small grains do most of the extinction of the light.

This is called an MRN distribution of grain sizes based off the work of Mathis et al. [1977]

Observational Effects of Dust – Dust is thought to be well-mixed through the neutral gas found throughout the galaxy in the ISM. Since this gas is found throughout the galaxy, more distant stars have a larger quantity of dust between us and the star we are trying to observe. The dust along the line of sight will absorb or scatter photons from the star propagating toward us, reducing the amount of light we receive. Furthermore because of the dust grain size distribution is heavily tilted toward small grains (Equation 3.22), the optical properties preferentially scatters and absorbs short wavelength light more effectively than longer wavelength light.[6] The net effect of this absorption and scattering is that the light from background sources is absorbed or *extinguished*[7] and this effect is called *dust extinction*. Since the short wavelength (i.e., blue light) is preferentially lost through absorption and scattering relative to the long wavelength light (i.e., red light), dust is also said to *redden* the light. This term can be misleading since dust is "de-bluening," since dust neither adds red light nor does it convert blue light to red light.

[6] For the E&M aficionados, in the regime where $\lambda \sim r_{\rm dust}$, the scattering of light is through *Mie scattering*, has an absorption cross section (i.e., probability) $\propto \lambda^{-1}$ and then for $\lambda \gg r_{\rm dust}$, the cross section scales like $\sigma_\lambda \propto \lambda^{-2}$

[7] Some astronomers say the light is "extincted." Those astronomers are wrong. Come at me.

Figure 3.12 shows a gas and dust cloud that blocks out all the light from background stars in the optical, but if the same cloud is observed at longer wavelengths, the background starlight will past through the screen of dust extinction. Note that, even in the optical image on the left of this figure, that the stars that are at the periphery of the cloud are redder than the stars near the edge of the image where there is no foreground extinction. Since only the "reddest" light passes through the cloud, this illustrates both the extinction and reddening properties of light.

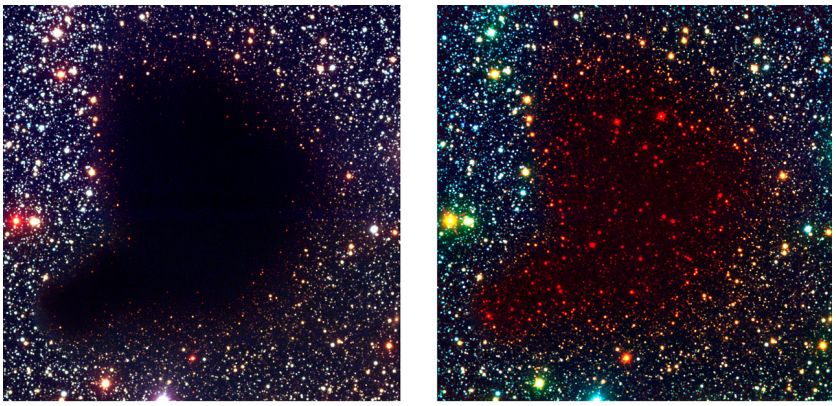


Figure 3.12: The gas globule known as B68 viewed in different optical filters. The left figure shows an image made from three colour bands of optical light (B, V, I) illustrating how the dust blocks out the background light at these short wavelengths. The right figure shows the same object but this time representing the infrared K band as the red colour in the image, showing how the light passes through the cloud at long wavelengths, which are less affected by the presence of dust. Image credit: ESO.

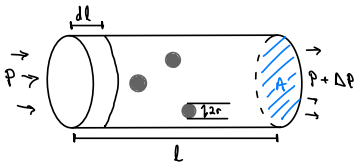


Figure 3.13: Sketch of setup for extinction.

Quantifying Extinction – We can quantify the observational effects of dust in the context of stellar absorption. Figure 3.13 shows the effect of a beam of light with initial power P_0 passing into a cylinder with cross-sectional area A and filled with dust grains, each with a fixed radius of r such that their diameters are the $2r$ illustrated in the figure. The beam emerges from the other side with a diminished power $P + \Delta P$, where $\Delta P < 0$. Our goal is to relate ΔP to the initial power P and the dust properties.

As illustrated, there are four dust grains in this volume. If each grain blocks an cross-sectional area πr^2, then the input power is reduced by the fraction of the cylinder's cross sectional area area blocked by the dust grains:

$$\frac{P + \Delta P}{P} = \frac{A - 3\pi r^2}{A} \tag{3.23}$$

where the 3 arises because there are three dust grains in the cylinder. Thus, in this setup, we can calculate the fractional loss of power in terms of the fraction of the area of the cylinder that's blocked by the dust grains.

The above sketch needs to be formalized by making two amendments. First, we need to allow for an arbitrary number of dust grains in the cylinder, which we represent in terms of the number density of dust grains, n, expressed in terms of "dust grains per unit volume." More subtly, we also need to allow for the case where one grain sits in front of another dust grain, "eclipsing" the grain in the back. This means that the eclipsing pair of dust grains end up blocking less area than $2\pi r^2$. To accomplish this, we restrict our analysis from a full cylinder of length ℓ to a very short cylinder of length $d\ell$. In this very

short cylinder, we can make the assumption that no dust grains block each other since the length can be shorter than a radius. Considering just this tiny cylinder, the fractional loss of power is then

$$\frac{P+\Delta P}{P} = \frac{A - \mathcal{N}\pi r^2}{A} \tag{3.24}$$

where $\mathcal{N}$ is the number of grains in the cylinder. For a number density of dust grains n, we can express $\mathcal{N} = nV_{\text{cyl}} = nAd\ell$ where the last equality follows by replacing the volume of the short cylinder in the problem with the quantities in the sketch.

$$\begin{aligned}
\frac{P+\Delta P}{P} &= \frac{A - nAd\ell\pi r^2}{A} \\
\frac{P+\Delta P}{P} &= 1 - nd\ell\pi r^2 \text{ so that} \\
P + \Delta P &= P - Pn\pi r^2 d\ell \\
\Delta P &= -Pn\pi r^2 d\ell
\end{aligned}$$

where we can turn the (tiny) difference of ΔP into a differentially small value representing the infinitesimal change in P. This, of course, is setting up a differential equation which we can separate and integrate[8] from position o to ℓ resulting in the change from initial power P_0 to the final power P.

[8] Don't worry if this looks wild. The main result is what's important.

$$\begin{aligned}
\frac{dP}{P} &= -n\pi r^2 d\ell \\
\int_{P_0}^{P} \frac{dP}{P} &= \int_0^{\ell} -n\pi r^2 d\ell \\
\ln\left(\frac{P}{P_0}\right) &= -n\pi r^2(\ell - 0), \text{ thus} \\
P &= P_0 \exp\left(-n\pi r^2 \ell\right).
\end{aligned}$$

This is a neat result: the power decreases exponentially as it passes through a column (i.e., the cylinder in Figure 3.13). We can make this even more general by recognizing that the assumed spherical shape for a dust grain doesn't really matter and we can, instead replace this with a more general variable just called the *cross section* σ. For spherical grains $\sigma = \pi r^2$ as above but we can average over irregularly shaped grain populations to come up with an estimate of σ:

$$P = P_0 \exp\left(-n\sigma\ell\right). \tag{3.25}$$

This generalization also allows us to take into account that the changing optical properties of dust grains with wavelength of light λ. Because of scattering and absorption properties, we could assume, for example $\sigma_\lambda = \sigma_0(\lambda/\lambda_0)^{-4}$ where σ_0 and λ_0 are constants and the λ subscript just means that the cross section varies with wavelength as

opposed to being "per unit wavelength" which we used in the the flux density notation.

Formally, a cross section is a very flexible notion defined in terms of a probability of absorption or scattering. When we refer specifically to the cross section set by the shape of the object, we call this the geometric cross section. The cross section concept is quite common through physics[9]. A lot of cross-section interactions are derived for subatomic or atomic scale interactions where the notion of a geometric cross section is poorly defined because of the delocalization of particles on the quantum scales. Instead, we define the cross section based on the probability that a specific interaction will occur. For example, if a beam the spans area A of N_{in} incident particles travels through an experiment with N_{targ} targets, we can measure the number of reactions that occur (scatterings, adsorptions, reactions, etc.), N_{rxn}. Then, we can define the cross section of the particles for those interactions as

$$\sigma = \frac{N_{\text{rxn}}/N_{\text{targ}}}{N_{\text{in}}/A}. \tag{3.26}$$

[9] You might have seen this in the context of the Rutherford Scattering experiment.

These numbers of reactions and incident particles can also be defined in terms of a flux, i.e., per unit time. As long as these times are the same, the expression remains the same. The numerator of this fraction represents the probability of a reaction occurring. Since this probability of a reaction can depend on both the properties of the incident particles (say, the wavelength of the incoming light) and the properties of the targets, the cross section can show dependence on a lot of different physical variables. Here, in astrophysics, we mostly focus on the variation with the incident wavelength of the light, so we will often write the cross section $\sigma_\cdot$ to indicate this functional variation.

In the domain where we consider the absorption (and later emission) of light, we usually contract the argument to the exponential in into a term called the optical depth τ_λ, writing this equation as $P = P_0 e^{-\tau_\lambda}$ where we have written $\tau_\lambda = n\sigma_\lambda \ell$ and explicitly indicated that we might have a wavelength-varying cross section and thus a wavelength-varying optical depth. We can further generalize this expression by recognizing that the density n can vary through space. Hence, tracing things back, this means that the optical depth is best expressed as

$$\tau_\lambda = \sigma_\lambda \int_0^\ell n d\ell = \sigma_\lambda N \tag{3.27}$$

where we have defined the *column density* $N \equiv \int n d\ell$. The column density is the total number of particles per unit area if all the gas along the line of sight were flattened down into a thin layer. The column density is referred to in units of particles per unit area, like

"atoms per m^2." Systems where $\tau_\lambda \ll 1$ are called *optically thin* and systems where $\tau_\lambda > 1$ are called *optically thick*.

Another expression we focus on in the study of opacity and particle interactions in astrophysics is the *mean free path*, which is the mean distance a particle can travel through a medium before it is likely to have a reaction. The mean free path is defined as

$$\ell_{\text{mfp}} \equiv \frac{1}{n\sigma}. \tag{3.28}$$

We will need to compare the mean free path of a particle to other physical scales in the system to determine if certain processes are important.

Finally, we can relate the optical depth to the magnitude system. Using the definition of the magnitude system, we can consider the before-and-after absorption powers in units of power per unit area, dividing out the area of the cylinder, so that

$$\begin{aligned} m - m_0 &= -2.5 \log_{10} \left(\frac{P_0 e^{-\tau_\lambda}}{P_0} \right) \\ &= -2.5 \log_{10} e^{-\tau_\lambda} \\ &= 2.5 \tau_\lambda \log_{10} e \\ &= 1.086 \tau_\lambda \end{aligned} \tag{3.29}$$

We finally refer to the difference in magnitudes before and after the dust extinction simply as "the *extinction*" and we give it the variable A, so that $A_\lambda = 1.086\tau_\lambda$, again using the subscript to indicate that the extinction can, in general, vary with wavelength of light observed.

With this definition of extinction in terms of magnitudes, we can refine our observed magnitude equation by including the effects of dust:

$$m_\lambda = M_\lambda + 5 \log_{10} \left(\frac{d}{10 \text{ pc}} \right) + A_\lambda. \tag{3.30}$$

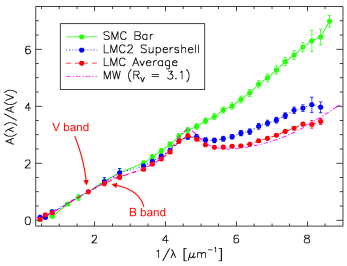


Figure 3.14: Extinction curves for different regions. Adapted from Gordon et al. [2003] under CC-by-SA.

Reddening and Colour Excess – While we have cryptically implied that the extinction of dust can vary with wavelength, we now take advantage of the similarity in the absorption and scattering properties of dust through the local Galaxy. We first adopt a waveband as a reference for the extinction value and then compare the extinction in other wavebands to that value. For historical reasons, we adopt the V band (see Figure 1.4) as the reference band for the extinction, likely having to do with it wavelength right in the middle of the optical part of the spectrum. We then compare the extinction in the V-band to the extinction in other wavelengths. Figure 3.14 shows the dust extinction curves for different wavelengths in different galactic environments. The x-axis units are a little funky: they

plot $1/\lambda$ since this is a natural quantity to use in studying the optics of dust grains. All the curves are normalized to the V band at $1/\lambda \approx 2\ \mu\text{m}^{-1}$. This curve shows that the extinction increases for shorter wavelengths (larger values of $1/\lambda$), including a characteristic "bump" at $\lambda = 217.5$ nm in the near ultraviolet. The different curves show the value for the Milky Way as well as curves for two satellite galaxies of the Milky Way with lower metallicity: the LMC is the Large Magellanic Cloud, and the SMC is the Small Magellanic cloud and the "bar" and "supershell" refer to two different regions in these galaxies. These satellites have lower metallicity than the Milky Way so they probe what happens to dust grains, which are made of heavy elements, in environments where there is less material to make those dust grains.

Overall, there is good agreement between the extinction curves for longer wavelengths (i.e., lower values of $1/\lambda$ in the Figure) but these curves diverge in at shorter wavelengths because of the changing optics of different dust grain populations. This is typically described in terms of the *reddening*, R_V, a measure of the slope of the extinction curve between the V band and the nearby B band, which are both indicated in the Figure. For most dust populations,

$$R_V \equiv \frac{A_V}{A_B - A_V} \approx 3.1. \tag{3.31}$$

The quantity in the denominator is called the *colour excess* and is frequently denoted as $E(B-V)$. The colour excess refers to how the a colour index (traditionally $B-V$) gets increased (i.e., made redder) in the presence of dust grains so that the $B-V$ colour is larger than it would be without dust, $(B-V)_0$.

We can see the effects of reddening in the Gaia data that we have been using from Gaia Collaboration et al. [2018], where their Figure 1 is reproduced in Figure 3.15. That figure shows the full sample of stars with high quality measurements, including those with significant signs of extinction. Comparing Figure 3.15 to Figure 2.7 shows that some features on the HR diagram, notably the red clump get smeared out along a line set by the reddening of interstellar dust. This so-called *reddening vector* is indicated by the arrow in Figure 3.15. The reddening is set by the properties of the dust grains and is measured by the Gaia team to be

$$R_{\text{Gaia}} = \frac{A_G}{A_{\text{BP}} - A_{\text{RP}}} \approx 1.8, \tag{3.32}$$

with some variation depending on the intrinsic colour of the source. This means that for each magnitude of extinction in the Gaia G band, the colour index will shift to the red by $1/1.8 \approx 0.56$ magnitudes. This relationship gives the slope of the line illustrated in the Figure.

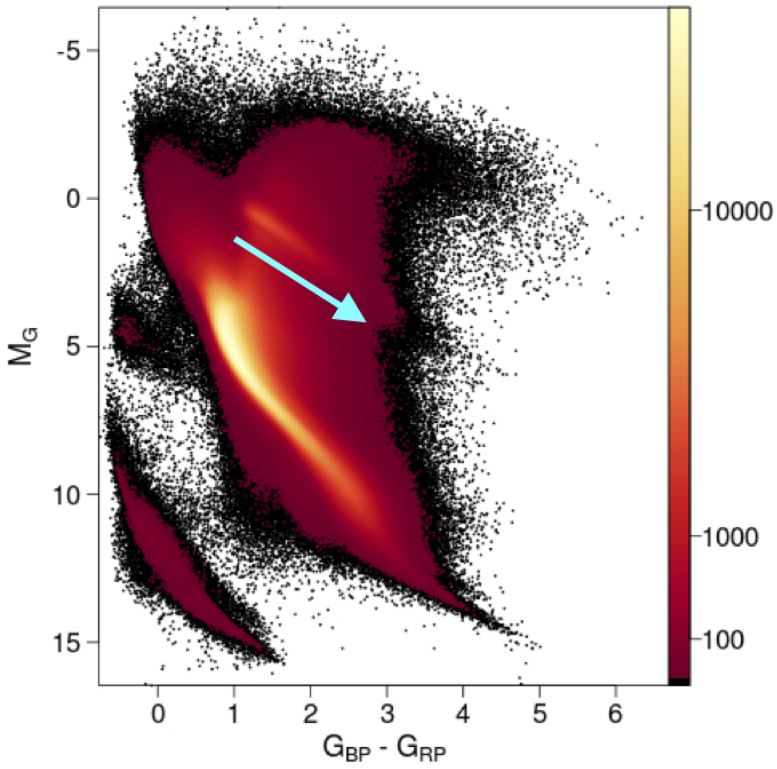


Figure 3.15: Full sample of Gaia data from Gaia Collaboration et al. [2018] shown in their Figure 1. The arrow indicates the effects of reddening.

By measuring the displacement of a star from its expected position along the reddening vector, we can infer the total extinction to the source. This works well for discrete features in the HR diagram like the red clump, but the effect of reddening along the main sequence is more difficult to discern: stars will shift along a line that is approximately parallel to the main sequence and reddening can get confused with the effects of binaries, aging and metallicity. Resolving these ambiguities requires making observations of a region is multiple different wavebands to infer the effects of extinction.

One of the major advances enabled by large scale surveys of stars in the the solar neighbourhood is that we can map out the distribution of where the dust is in space. In this model, we assume that dust is found in clouds of gas that are distributed through the local environment. The means that stars that we can use the differential reddening towards stars that have similar positions on the sky (i.e., similar RA-Dec) but different distances to measure the amount of dust extinction between those two stars. This will allow us to find dust clouds in three dimensions around us and then use that dust map to determine how much reddening is expected for a new star at a given distance. This is the method use by the Gaia team to find their sample of high quality stars with low extinctions that is shown in Figure 2.7.

3.3.3 Star Formation Histories

Thus far, we have focused on how dust opacity and observational effects influence our analyses, but the influence of the star formation history of a system plays a major role in determining the density of stars in the HR diagram. When we say "density," this means the numbers of stars in a set of observations with a given colour and absolute magnitude. We alluded to these effects being relevant when considering the most luminous stars in the HR Diagram, considering why there were so few O and B type stars compared to the F and G type. This can be attributed to the short main sequence lifetimes of these stars. Seeing such stars means that they must have formed recently, or else they would have evolved off the main sequence. For example, if a star has a main sequence lifetime of 10 Myr, there must have been star formation in the past 10 Myr. Furthermore, given the IMF, these high mass stars must have formed with several low mass stars that are also this young and in the vicinity of the high mass stars. We cannot necessarily distinguish them from other stars in the solar neighbourhood from their position in the HR diagram, but star formation will fill up the HR diagram at positions corresponding to all masses, not just the high mass stars. However, these high mass

stars can only be formed through recent star formation so they are unique indicators that the star formation has occurred recently.

In contrast, low mass stars have very long lifetimes, with all stars that have $M < 0.9\ M_\odot$ having main sequence lifetimes longer than the current age of the Universe. Thus, the parts of the HR diagram where these low mass stars are formed are set by the integral of all star formation that has occurred over the lifetime of the Galaxy. The medium mass stars turn out to be the most interesting regime where this population of stars reflect the integrated star formation back to different points in the Galaxy's lifetime, corresponding to the main sequence lifetime of those stars.

We can use these changing main sequence lifetimes to measure the *star formation history* of a galaxy. We describe the star formation history in terms of the star formation rate of the system $\dot{M}_\star$ which represents the mass of stars formed in some part of the galaxy per unit time, also written as the SFR. We can use the IMF to predict the number of stars formed in a mass range by considering the total number of stars formed per unit time as $\dot{N}_\star = \dot{M}_\star / \langle M \rangle$ where $\langle M \rangle$ is the average mass of a star. Then, using the IMF we can predict the number fraction[10] of stars formed in a given mass range, f_M. Then, the number of stars on the main sequence in that mass range should be:

[10] As opposed to the mass fraction.

$$N_{\star,M} = f_M \int_0^{t_{\rm stop}} \dot{N}_\star dt = f_M \int_0^{t_{\rm stop}} \frac{\dot{M}_\star}{\langle M \rangle} dt \tag{3.33}$$

where $t_{\rm stop}$ is either the main sequence lifetime of the star for high mass stars or the age of the Universe for low mass stars (taken as 14 Gyr). By considering stars in different mass ranges with different main sequence lifetimes, we can infer the star formation history of the system by examining the numbers of stars remaining today.

To make this argument more concrete, we can consider two mass ranges of stars, $1.9 < M/M_\odot < 2.1$ and $3.9 < M/M_\odot < 4.1$. The main sequence lifetime of the $M \approx 2\ M_\odot$ stars is about $\tau_2 = 1800$ Myr whereas the $M \approx 4\ M_\odot$ stars have a main sequence lifetime of about $\tau_4 \approx 300$ Myr using the scalings in Equation 2.10. The $M \approx 4\ M_\odot$ stars can give us the average star formation rate over the past 300 Myr, which is the main sequence lifetime of these stars:

$$\langle \dot{M}_\star \rangle_{\tau_4} = \frac{N_{\star,4} \langle M \rangle}{f_4 \tau_4}, \tag{3.34}$$

where $\langle M \rangle$ and f_4 are predicted from the IMF and τ_4 is known from stellar evolution. Similarly, we can find the star formation rate averaged over the past 1800 Myr by examining the $M \approx 2\ M_\odot$ stars:

$$\langle \dot{M}_\star \rangle_{\tau_2} = \frac{N_{\star,2} \langle M \rangle}{f_2 \tau_2}. \tag{3.35}$$

This is neat! We have essentially used the number counts of stars in two different mass ranges to predict the average star formation rate over two different time intervals. Since $\langle \dot{M}_\star \rangle_{\tau_2}$ averages over 1800 Myr and $\langle \dot{M}_\star \rangle_{\tau_4}$ averages over the past 300 Myr, we can also reframe these measurements in terms of the star formation rate in two windows: between 0 and 300 Myr and from 300 Myr to 1800 Myr, where we denote the latter $\langle \dot{M}_\star \rangle_{\tau_2 - \tau_4}$. We just need to recognize that:

$$\langle \dot{M}_\star \rangle_{\tau_2} = \frac{\tau_4 \langle \dot{M}_\star \rangle_{\tau_4} + (\tau_2 - \tau_4) \langle \dot{M}_\star \rangle_{\tau_2 - \tau_4}}{\tau_2} \tag{3.36}$$

and we can solve for $\langle \dot{M}_\star \rangle_{\tau_2 - \tau_4}$:

$$\langle \dot{M}_\star \rangle_{\tau_2 - \tau_4} = \frac{\tau_2 \langle \dot{M}_\star \rangle_{\tau_2} - \tau_4 \langle \dot{M}_\star \rangle_{\tau_4}}{\tau_2 - \tau_4}. \tag{3.37}$$

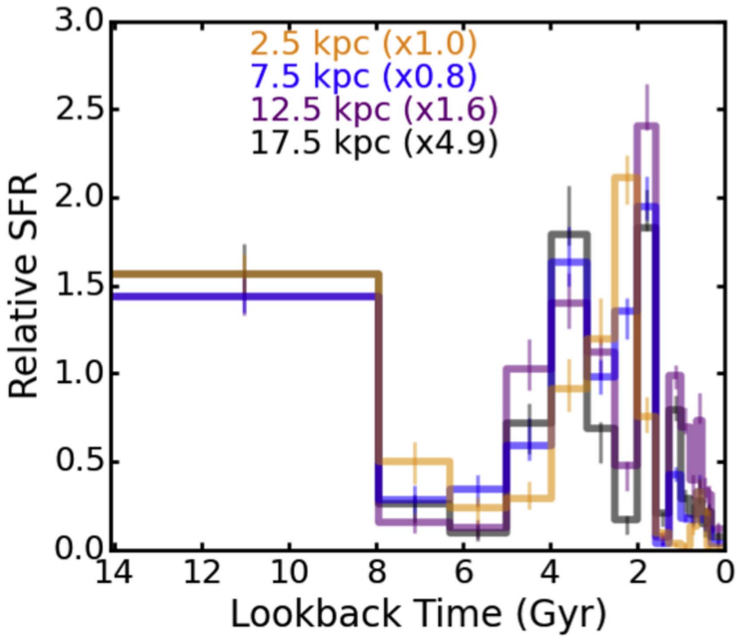


Figure 3.16: Star formation history in the Andromeda galaxy using data from Williams et al. [2017].

Figure 3.16 shows this technique applied to the nearby galaxy Andromeda (M31) using data from the Panchromatic Hubble Andromeda Treasury (PHAT) as presented in Williams et al. [2017]. This study uses data from the Hubble Space Telescope to measure individual medium mass stars in M31 and carry out the same process we outlined above. The different curves show different regions in the galaxy measured in distance from the centre of the galaxy. The "lookback time" is the time period for which these measurements are made. Hence the lookback time we derive in Equation 3.37 would be

0.3 - 1.8 Gyr on this axis. The star formation rate is varying in this galaxy over time, with a high rate early in M31's evolution, a quiescent period, a spike between 2-4 Gyr ago and relatively low star formation rates recently. Note that the age bins get a lot smaller as we get to short lookback times because these correspond to the windows probed by the high mass stars. Their main sequence lifetimes are short and changing quickly with mass, and these stars are quite bright, so this means that we have relatively fine time resolution in this part of the diagram.

As an aside, this type of study is very "clean" in nearby galaxies where we can measure these individual stars because the stars are all approximately at the same distance because the distance to the galaxy is much larger than the differences between the distances. Thus, this study is not strongly subject to the distance biases described in Section 3.3.1.

There are several other types of analysis that leverage this reasoning, where we can infer how galaxies build up mass by using the number of stars with different properties. We can also extend this reasoning to unresolved stellar populations. Figure 3.17 shows how the light from a galaxy changes colour if depending on the star formation history of the system. This compares the light from a star formation rate that is a constant over 14 Gyr and an exponentially decaying star formation rate that follows $\dot{M}_\star \propto e^{-t/\tau_\star}$ where $\tau_\star$ is 1 Gyr and star formation started 14 Gyr ago. We can fit simple models of star formation histories to the observed spectra of galaxies and determine how galaxies assembled their mass.

With this, we have a good sense of how stars look over time. There remain two subtleties to really understand the light we see from galaxies and for inferring their physical properties: the enrichment of metals in galaxies and the interstellar medium. However, these pieces are tightly linked to galaxy evolution so we will treat them in some more detail in what follows. In particular, we will now turn our attention to the Interstellar Medium. The ISM significant reshapes the light we see from galaxies and the material in the ISM sets the star formation rate, and thus the evolution of galaxies.

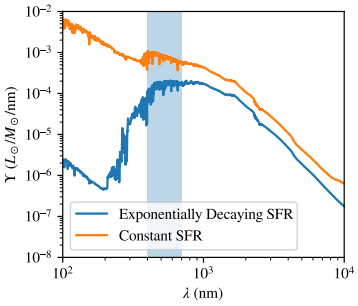


Figure 3.17: Stellar spectra for two different star formation histories. The shaded region indicates the optical portion of the spectrum.

Key Points

- Real stellar populations are more complex because of three main factors: observational effects, dust, and the mixture of stellar populations with different ages and metallicities.
- Observational studies are naturally biased because it is easier to see high luminosity stars to a farther distance,

so these stars are over-represented in the data compared to lower mass stars. This effect must be accounted for by careful consideration of the volume over which the survey is complete to different stars. Complete just means the volume over which nearly all of the actual stars are included in a survey.

- Astrophysical dust is a low mass component of galaxies that significant reshapes the light that we observe. Dust grains absorb short wavelength stellar radiation (Optical, UV) and reradiate it thermally in the infrared.
- Dust grains have small cross sections σ that vary as a function of wavelength. Dust grains absorb / scatter short wavelength light more effectively than long wavelength light. We can see through dust extinction features by observing in longer wavebands like the infrared (Equation 3.12). Dust extinction cause the amount of light received through a cloud to exponentially decrease with the amount of dust in the cloud (Equation 3.3.2).
- The extinction effects of dust are measured in terms of optical depth (Equation 3.27) or the magnitudes of extinction 3.29, which is normally given the variable A and is measured in magnitudes.
- The mean free path is the average value of how far something travels before it interacts with something else in the mdeium. (Equation 3.28).
- The *reddening* of dust is a measure of the slope of the extinction curve as a function of wavelength. For Milky Way dust, there is a relationship between the extinction in one band and in another band (Equation 3.31).
- The star formation history is a measure of the rate of star formation as a function of time in a specific stellar population. The star formation history sets the age distribution of stars in the galaxy.
- Star formation histories are measured by comparing the number of stars with different main sequence lifetimes in a resolved (by counting stars) or unresolved (by measuring the sum of contributing spectra) stellar population.

- Star formation histories show that the star formation rate in galaxies has varied significantly over the age of the Universe.

4
The Interstellar Medium

The interstellar medium (ISM) is the catch-all descriptor for everything in a galaxy except for (1) stars and (2) dark matter. Most the ISM consists of gas in different thermodynamic and chemical states, but the ISM also consists interstellar dust described in the previous chapter as well as cosmic rays and the magnetic field. Here, we will first focus on the different phases of the ISM and then turn to three case studies that highlight important physical processes in understanding the role of the ISM in shaping galaxy evolution.

4.1 The Phases of the ISM

Here, the phases of the ISM typically refers to the state of the gas, which is described in terms of the chemical state of hydrogen – molecular, neutral, or ionized[1] – and its temperature: cold, cool, warm and hot.

[1] These typically use H_2 molecular hydrogen and then spectroscopic notation for the state of atomic hydrogen. Neutral hydrogen, H^0 is denoted H I and ionized hydrogen, H^+ is denoted H II.

The properties of these different phases are summarized in Table 4.1. The table presents a few general trends on which we can focus. First, an most importantly, the pressures of the different phases are all roughly equal, that is $P = nk_BT \sim 10^{-14}$ Pa, where n is the particle density. In ISM land, this pressure is often expressed in the equivalent form of $P/k_B = nT$ in units of K cm^{-3}. Second, the temperatures and ionization states are correlated: cold gas ($T < 100$ K) tends to be molecular whereas hot gas ($T > 100$ K) is ionized. Of course, the ISM is not just hydrogen and contains helium as well as metals in both gases (ionized and neutral) and solids (dust grains).

The other trend that shows up in Table 4.1 is fractions of the mass and the volume occupied by these different phases. Most of the mass of the ISM is found in the atomic (neutral H I) medium (70%), with the warm ionized medium and the cold medium each comprising about 15%. The hot ionized gas in the galaxy contains only a tiny amount of the mass, but this phase fills about half of the volume in the disk of the Milky Way with hot "bubbles."[2] In contrast, the high

[2] In fact, our solar system is currently travelling through the inside one of these bubbles of million-degree plasma, affectionately known as "The Local Bubble."

Name	Temperature (K)	Density (m^{-3})	Mass Fraction	Volume Fraction	Observational Tracer
Hot Ionized Medium	10^6	10^3	—	$\lesssim 0.50$	X-ray emission lines
Thermally unstable (ionized)	$\sim 10^5$	$\sim 10^4$	—	—	UV absorption lines
Warm Ionized Medium	$8000 - 12,000$	10^6	0.15	0.25	H-recombination lines
HII Regions	$8000 - 12,000$	10^6+	< 0.01	—	H-recombination lines
Warm Neutral Medium	$6000 - 10,000$	$10^2 - 10^3$	0.40	0.30	21-cm emission line, atomic fine structure lines, dust emission
Thermally unstable (neutral)	~ 5000	$\sim 10^3$	—	—	21-cm emission line, atomic fine structure lines, dust emission
Cold Neutral Medium	$100 - 3000$	$10^4 - 10^6$	0.30	0.01	21-cm emission line, atomic fine structure lines, dust emission
Cold Molecular Medium	$10 - 100$	$> 10^8$	0.15	0.0005	Tracer molecular lines (CO, OH, HCN), dust emission

Table 4.1: Phases of the ISM in the Milky Way disk

density phases of the ISM occupy much smaller volume fractions. Indeed, this must be true because they have high densities and comparable mass fractions. Thus, the molecular medium of a galaxy fills a tiny fraction of the volume (0.05%) but this tiny fraction is 1/6th of the mass of the ISM and also where all the star formation occurs.

As a caveat, Table 4.1 is specifically for the ISM conditions in the local Milky Way disk. These mass fractions and volume fractions change throughout the Milky Way. For example, the ISM in the inner 500 pc in the Milky Way has nearly all of its mass ($> 90\%$) in the molecular phase. The ISM of some galaxies has almost no neutral or molecular gas and is dominated by the hot ionized medium (for future reference, these are the elliptical galaxies).

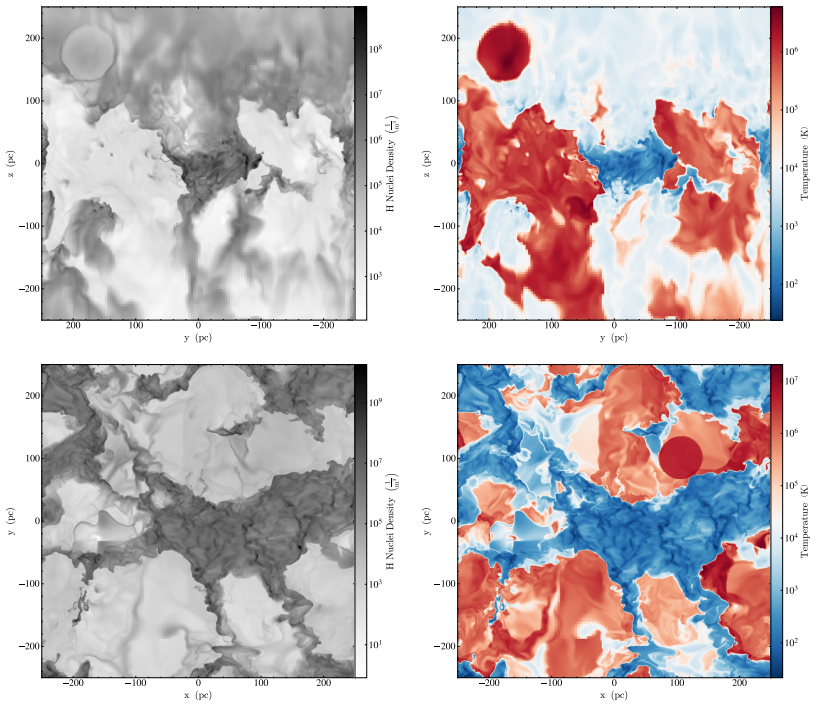


Figure 4.1: Simulation of a portion of the ISM from the SILCC project [Girichidis et al., 2021].

Figure 4.1 shows a slice of a single timestep drawn from a numerical simulation of the ISM conducted as part of the SILCC project, which has some amazing movies[3]! The figure shows two views of the ISM from two directions. Specifically, it shows the hydrogen particle density and the temperature shown as a slice through the $x-y$ plane, corresponding to the midplane of the system. The slice through the $y-z$ plane slows the three dimensional structure of the gas extending out of the plane of the galaxy in the z direction. The simulation has a cold atomic/molecular cloud of gas in its centre, which is surrounded by hot gas found in an ionized state. A hot bubble is visible

[3] https://hera.ph1.uni-koeln.de/~silcc/

out of the plane, which is the signature of a recent supernova explosion. The general trends seen in Table 4.1 are clear in this figure, but the figure also illustrates that the ISM is messy by nature relating to its dynamic nature. The physical conditions span orders of magnitude in range, but the general scaling is that the product of $n_{\rm H}T$ is approximately constant. The "fuzziness" at large values of $|z|$ is a product of the simulation method where these zones are simulated with lower resolution to save on the processing time needed with the supercomputers used in this study.

The study of the ISM is complicated and incorporates a huge swath of physical effects. It is challenging to obtain a complete picture from first principles since the physical conditions span orders of magnitude in energy, length and time scales. However, our growing understanding of the ISM has developed three principles that describe the ISM:

- *The ISM is dynamic.* – This means that the ISM is constantly changing. These changes can consist of motion, with gas flows churning material around in turbulent motions with crossing times (i.e., size of system divided by the velocity of gas) be comparable to or shorter than all the other timescales describing evolution. Ultimately this means that we cannot neglect the time variability of the system in the "real ISM" though our case studies will look at some steady states.

- *The ISM is in equipartition.* – Even though the ISM is constantly changing, the average properties of different ISM phases are in a rough balance. Specifically, the pressures in the ISM are all approximately equal. This does not imply force balance between these phases, but it does mean that the momentum flux (i.e., how pressure is derived) between these phases is approximately equal. Note that pressures and energy densities are dimensionally equivalent: a force per unit area has the same dimensions as an energy per unit volume. Thus, this statement is equivalent to saying that the energy densities of the different phases of the ISM are equivalent. The ISM is not in equilibrium, but it is in a statistical steady state and some processes must regulate that steady state.

- *The ISM is the place where matter and energy exchange while flowing through a galaxy.* – The study of the ISM is particularly important because it is the interface between the material in the intergalactic medium which flows into galaxies, through the ISM and ultimately into stars through the star formation process. When stars drive stellar winds and supernova out into their host galaxies, the energy and momentum injection from this feedback is injected into

the ISM. The dust in the ISM can reprocess over 99% of the radiation from stars in some galaxies, and serves as the main factor that controls how galaxies emit light. Hence, the studies of the ISM are telling us about how galaxies are interacting with themselves and with their environment.

Since we aren't studying all about the ISM, even though we totally should, we will focus on a few specific phenomena observed in the ISM that illustrate three different physical effects that the ISM has on galaxies. Specifically, we will study how the ISM reprocesses stellar radiation through gas as dust (Section 4.2), how fluid flows and shocks shape the ISM 4.4, and how atomic scale processes heat and cool the gas in the ISM 4.3.

Key Points

- The ISM is broadly categorized into phases characterized by the temperature and density of the hydrogen gas. These phases are all in approximate pressure balance (higher density regions have lower temperatures) so that $P_{\rm ISM} = nk_{\rm B}T$ is approximately constant.
- The ISM is evolving on a rapid timescale and is frequently buffeted by the passage of shock waves.
- The ISM is the conduit by which energy and momentum can flow into and and through a galaxy.

4.2 *Case Study: H II Regions*

We will first study photoionized regions around massive stars. Figure 4.2 shows a photoionized region called NGC 604 that is found in the nearby galaxy M33 (also known as Triangulum). Photoionized regions are also called H II regions because the dominant chemical state of hydrogen is ionized and these regions are relatively dense such that they end up being quite bright relative to other components of the ionized medium.

An H II region is the cloud of ionized gas around a young, high mass star or group of stars. These high mass stars have relatively high surface temperatures such that they are emitting a significant amount of energy in photons with $E_\gamma > 13.6$ eV, which is the ionization potential of hydrogen. If such a photon encounters a hydrogen atom, it will ionize it into a free electron and proton.

The physics of an H II region depends on the electronic structure of a hydrogen atom. The energy levels of hydrogen are determined

Figure 4.2: The giant H II region NGC 604 in the nearby galaxy M33. Image credit: Hui Yang (University of Illinois) and NASA/ESA

by the principal quantum number n for the atom and have energies $E_n = -E_0/n^2$ where $E_0 = 13.59844...$ eV is the binding energy of the atom. Since hydrogen serves as as such an amazing introduction to quantum physics, you've probably seen this energy structure before, but in ISM physics, the electronic transitions of hydrogen are particularly important.

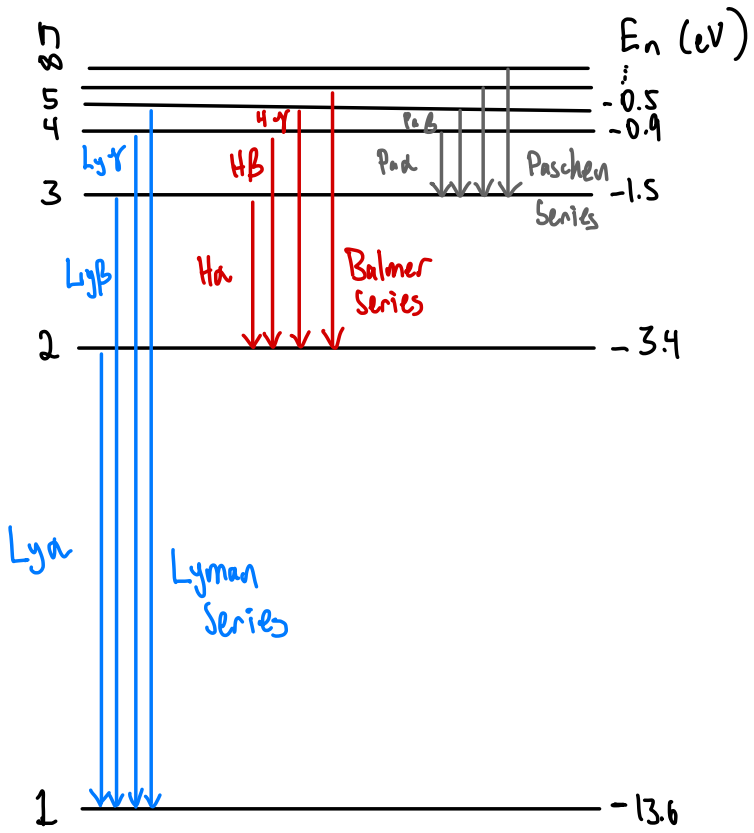


Figure 4.3: Energy levels of hydrogen and their respective spectral series.

Figure 4.3 shows the electronic energy level structure of the hydrogen atom. The arrows represent the electronic transitions of the hydrogen atom between different energy states. When the electron makes a downward transition, the atom will radiate a photon with energy

$$E_\gamma = \Delta E = E_0 \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right) \tag{4.1}$$

where n_i is the initial energy level and n_f is the final energy level such that $n_i > n_f$. Given this photon energy, the wavelength of the photon will then be $\lambda = hc/E_\gamma$. The spectral emission of hydrogen is further organized into spectral series that are classified by the end

state of the transition (the n_f) where each series is given a special name. The $n_f = 1$ lines are called the Lyman lines, the $n_f = 2$ lines are the Balmer lines, the $n_f = 3$ lines are called the Paschen lines, $n_f = 4$ is Brackett, etc. The transitions in each series are labelled with Greek letters starting with α corresponding to the longest wavelength photon in the series, moving on to β, etc. As a special case, since the Balmer lines are prominently in the optical, these are often just called H instead of Balmer. So, the Hα line is the $n = 3 \rightarrow 2$ transition of the hydrogen atom, Hβ is $n = 4 \rightarrow 2$, etc.

Because $1/n_i^2 \rightarrow 0$ as n_i increases, the lines grow progressively closer spaced in wavelength until they reach the limiting case that is called the continuum case corresponding to $n_i = \infty$ with the corresponding photon energy $E_\gamma = E_0/n_f^2$. The photon wavelength corresponding to this transition is often denoted with a c, so that Lyc means the Lyman continuum photon with wavelength $\lambda_{\rm Lyc} = 91.2$ nm and photon energy of 13.6 eV. Similarly, for the Balmer continuum (usually abbreviated Bac, just to break conventions), the photon energy is 3.4 eV. The continuum photon corresponding to n_f is the limiting energy required to ionize a hydrogen atom where the electron is in the n_f state. Table 4.2 gives some of the names and wavelengths for observationally important hydrogen spectral lines that we will occasionally reference. These wavelengths are the measured wavelengths in air; the vacuum wavelengths that you calculate from first principles are slightly shorter. However, we quote the air wavelengths because most spectrographs are operating on the Earth where they are filled with air so these are the wavelengths they measure.

n_f	n_i	Name	λ (nm)
1	2	Lyα	121.6
1	3	Lyβ	102.6
1	∞	Lyc	91.2
2	3	Hα	656.3
2	4	Hβ	486.1
2	5	Hγ	434.0
2	∞	Bac	364.6
3	4	Paα	1875
3	5	Paβ	1281
3	∞	Pac	820.4

Table 4.2: Observationally important H lines.

With this quick reintroduction of the hydrogen spectral series now past us, we can return to the physics of an H II region. The nebulous material that you see in Figure 4.2 is light from the Hα line of hydrogen. This emission line is particularly prominent from ionized regions. The ionizing photons from the star (or rather the "Lyman continuum photons" because we love jargon) split electrons off the hydrogen atoms and these electrons whiz around the protons in a plasma. When an electron gets close to a proton, the reverse process of photoionization can occur where the electron is captured into a bound state, making a hydrogen atom. We call this process *recombination* because the electron is recombining with the hydrogen atom. Recombination can occur from the 'continuum' (i.e., the unbound state) into any of the electronic levels of the atom. If the electron recombines into any state except the ground state, it will then deexcite down toward the ground level, giving off photons for each of the jumps downward. The emission visible in 4.2 is from electrons passing from $n = 3 \rightarrow 2$ and they would then radiation a Lyα photon in

the transition from $n = 2 \rightarrow 1$.

THE STRÖMGREN SPHERE – With this model of photoionziaton and recombination in mind, we can calculate the radius of an H II region based on the properties of the ionizing star, the physics of the hydrogen atom, and the properties of the interstellar medium. This simplified model is called the *Strömgren sphere* after the theorist that first modelled the system: Professor R. Sphere.[4]

Figure 4.4 shows the model of the Strömgren sphere. We assume that there is a star or stars producing ionizing photons in the middle a uniform medium of atomic gas (molecular gas is more challenging because of the dissociation physics of the hydrogen molecule). The ionizing photons radiate outward and ionize gas creating a spherical bubble of ionized hydrogen with a particle number density of hydrogen nuclei $n_{\rm H}$. Since the gas is hydrogen there is an equal density of electrons, $n_e = n_{\rm H}$. We want to calculate the radius of the sphere given $n_{\rm H}$ and the number of ionizing photons that the star(s) produce per second Q_0.

[4] Sorry. But not really.

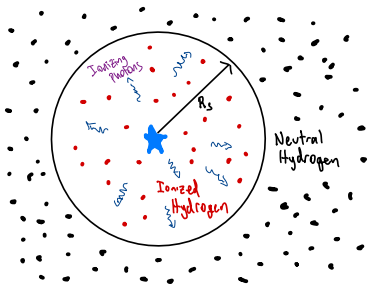


Figure 4.4: The Strömgren sphere model.

The physical model assumes that the region is in steady state such that the number of ionizing photons produced by the star (Q_0) balances the number of recombinations of electrons and protons into bound Hydrogen atoms every second ($Q_{\rm rec}$). If $Q_0 > Q_{\rm rec}$, the region would expand and vice versa. Hence the system is in steady state when $Q_0 = Q_{\rm rec}$. We further model $Q_{\rm rec}$ in terms of the number of recombinations that occur per unit volume per unit time, $N_{\rm rec}$ times the spherical volume for a sphere of radius R_s:

$$Q_0 = Q_{\rm rec} = \frac{4\pi}{3} R_s^3 N_{\rm rec}. \tag{4.2}$$

Clearly, the key physics is the physics of recombination. We describe the rate of hydrogen recombinations that occur per unit volume per second as

$$N_{\rm rec} = \alpha(T) n_{\rm H} n_e \tag{4.3}$$

where the parameter $\alpha(T)$ is called the *recombination coefficient*. This coefficient is a catch-all simplification for all the complicated quantum physics of how electrons get captured by protons. We can cram all this glorious quantum mechanics down into a simple parameterization given by Draine [2011]:

$$\alpha(T) = 2.54 \times 10^{-19} \left(\frac{T}{10^4 \text{ K}} \right)^{-0.8163} \text{m}^3 \text{ s}^{-1}. \tag{4.4}$$

This fitting formula is referenced to $T = 10^4$ K since that is the typical temperature range of gas in H II regions. The recombination rate decreases with increasing temperature because the higher temperatures

mean that the electrons are moving faster. Hence, the probability of capture into a bound state is reduced. Note that the units of 4.4 work out so that Equation 4.3 has units of $\mathrm{m}^{-3}\ \mathrm{s}^{-1}$, i.e., recombinations per unit volume per unit time as desired.

Returning to the model in Equation 4.2, we then have

$$\begin{aligned} Q_0 &= \frac{4\pi}{3} R_s^3 \alpha n_{\mathrm{H}} n_e \qquad \text{so that} \\ R_s &= \left(\frac{3Q_0}{4\pi\alpha n_{\mathrm{H}}^2} \right)^{1/3} . \end{aligned} \tag{4.5}$$

Note that in the last step, we have also invoked the assumption that $n_{\mathrm{H}} = n_e$. This is the size of the Strömgren sphere.

As an example, we can find the size of an H II region that carves out a bubble in the typical volume density in the cold neutral medium with $n_{\mathrm{H}} = 10^6$. We need to know Q_0 for the star and Table 4.3 gives the vital statistics of a few O-type stars on the main sequence. Let's consider an O5V star with $Q_0 = 10^{49.2}$ photons s^{-1}. We assume that the gas temperature is 10^4 K so that $\alpha = 2.54 \times 10^{-19}\ \mathrm{m}^3\ \mathrm{s}^{-1}$. Then, substituting into Equation 4.5 gives $R_s = 2.5 \times 10^{18}$ m $= 80$ pc.

$M/M_\odot$	$\log_{10}(Q_0/\mathrm{s}^{-1})$	T_{eff} (K)	Sp. Type
58.0	49.6	45,000	O3V
38.1	49.2	41,000	O5V
25.3	48.8	37,000	O7V
17.1	48.1	33,000	O9V

Table 4.3: Radiation from massive stars. Adapted from Draine [2011].

Reprocessing Radiation – This particular example is a useful case study for galaxy evolution because of the net action of the H II region. The net effect is to reprocess every ionizing photon from the star into photons in the hydrogen spectral series with energies less than 13.6 eV. Throughout the H II regions there are some neutral atoms even though nearly all of the atoms are ionized. These are the atoms that have recently recombined and their electrons are sitting in the $n = 1$ ground state. When an ionizing photon encounters one of these atoms, it will ionize the electron free of the proton. When that electron recombines, it could recombine back into the $n = 1$ state but this will produce another ionizing photon that will just fly off and find another atom to ionize. However, if it recombines to $n > 1$, it will release a continuum photon with $E < 13.6$ eV that cannot ionize another atom. The electron will then deexcite down to the ground state, and all the photons produced in this process will also have $E < 13.6$ eV. Hence, the net effect of the H II region is to "break up" energy in ionizing photons and redistribute it into the emission in the spectral line series of hydrogen. All these lines correspond to lower photon lower energies.

Figure 4.5 shows the effect of including H II regions and the reprocessing of radiation by nebular emission on a galaxy's SED. The nebular emission includes both the spectral lines of hydrogen as well as several other ions like O II and O II and N II that contribute to the nebular emission. Hence, the Hydrogen spectral lines series are

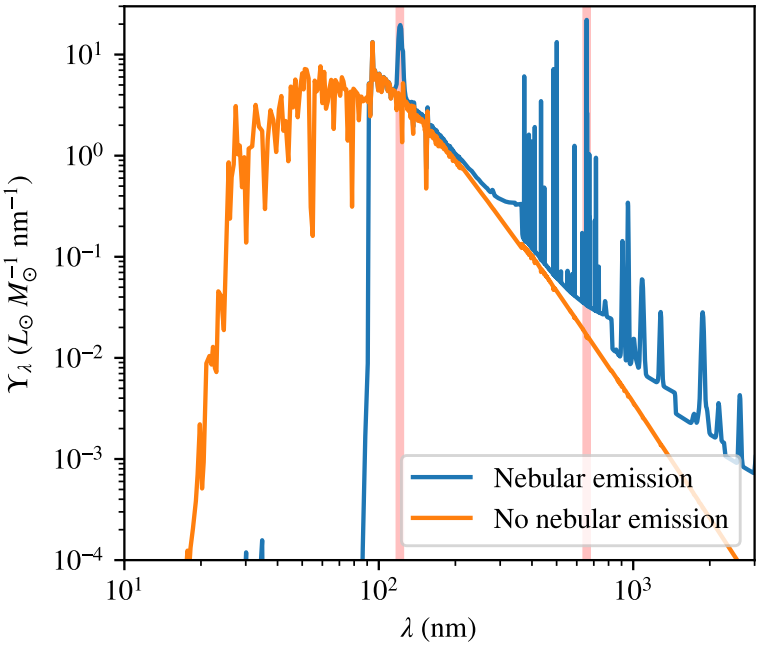


Figure 4.5: SEDs of a young simple stellar population with $\tau = 1$ Myr and $Z = Z_\odot$ shown with and without nebular emission. The presence of neutral hydrogen in the system reprocess the short wavelength radiation into optical emission line radiation. The highlighted regions show Lyα and Hα.

obscured by these other lines. However, the Lyα and Hα lines are highlighted and are prominent in the spectrum. Note that the young population of stars has a lot of emission at $\lambda < 91.2$ nm (photon energies $E > 13.6$ eV) that is completely removed and reprocessed into nebular emission once the effects of neutral gas are included.

Reprocessing by Dust – The reprocessing of radiation is important throughout astrophysics, moving energy from high energy photons into groups lower energy photons. One of the main agents for this reprocessing is interstellar dust. In the previous chapter, we focused on how dust extinguished the light from background stars, but we paid less attention to where the light that was blocked ends up going. A sizable fraction of the light is scattered off in different directions, but photons that are absorbed by dust heat up the grains and the grains reradiate in the infrared part of the spectrum.

The shape of the spectrum is set by the sizes of dust grains. The canonical distribution of dust grains follow a bottom-heavy size distribution for grains of characteristic radius r:

$$\frac{dN}{dr} \propto r^{-3.5} \tag{4.6}$$

meaning there are many more small grains than large grains. The

power-law is bounded by grains with a size of $r_{\max} \sim 250$ nm and $r_{\min} \sim 5$ nm. Since the mass of a grain is proportional to its volume with only a small change due to volume density, most of the mass in the dust grains is concentrated in the largest grains. In contrast, the geometric cross section is proportional to the area of the grain, i.e., r^2 so most of the absorption cross section is found in the small grains. The actual cross section does follow the more complicated scaling with wavelength discussed previously, but this does not change that the small grains do most of the extinction of the light.

This is called an MRN distribution of grain sizes based off the work of Mathis et al. [1977]

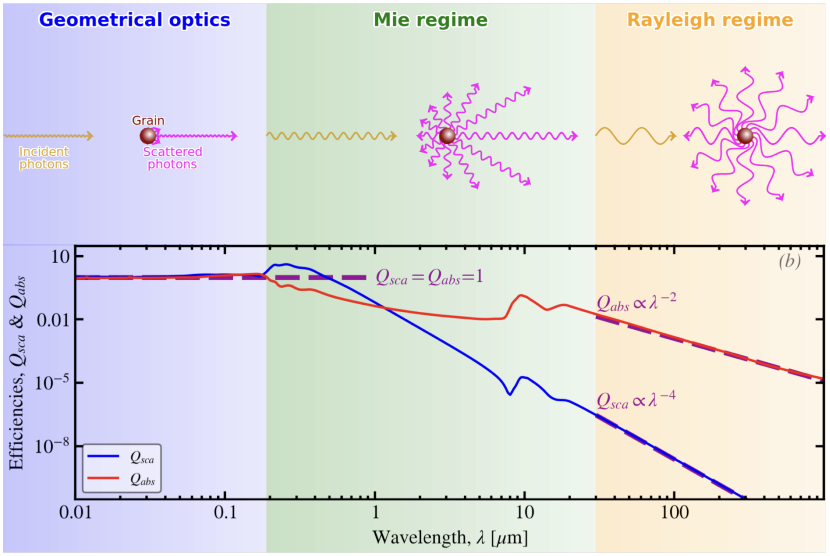


Figure 4.6: Dust cross section vs wavelength for $r = 0.1$ μm adapted from Galliano [2022], under CC-by-SA 4.0.

Figure 4.6 shows the scattering cross section for a dust grain as function of wavelength for a $r = 100$ nm $= 0.1$ μm. In this graph, $Q_{\text{abs}} =\equiv \sigma_\lambda / \pi r^2$, i.e., the ratio relative to The figure illustrates the different optical regimes for scattering including geometric absorption $2\pi\lambda \ll r$, the Mie regime $2\pi\lambda \sim r$ and the Rayleigh scattering and absorption limit $2\pi\lambda \gg r$ where the scattering cross section $\sigma_\lambda \propto \lambda^{-2}$. The absorption properties of dust grains

Once these grains absorb photons they will heat up and reradiate. Like the absorption, the small size of dust grains also makes the emission properties more complicated. Specifically, thermal emitters have a low efficiency for emitting photons with wavelengths comparable to the physical scale of an object: a 200 nm dust grain has reduced efficiency for emitting photons with $\lambda > 200$ nm. The efficiency for absorption and emission drops off as the wavelength gets larger. This means that the emission spectrum for dust is not necessarily a blackbody but need to be modified, leading to a dust population following an analogue to the Stefan-Boltzmann law that

shows that $L_{\rm dust} \propto T_{\rm dust}^{4+\beta}$ where β is a constant set by the size distribution of dust grains with $\beta \approx 2$.

The peak of the emission typically occurs in the 100 to 250 μm range implying dust temperatures that are significantly lower than for the very small grain emission. Typically, dust emission, once fit with the full modified blackbody expression returns $T_{\rm dust} \sim 20$ K for most of the ISM. Dust is a relatively good radiator in the infrared and the amount of emission ends us scaling like $L \propto T_{\rm dust}^{6}$ where again, T^4 would be the standard Stefan-Boltzmann scaling. This strong scaling keeps the dust temperatures in a fairly narrow range. Modern dust models admit a temperature distribution of dust grains to produce the observed SEDs.

The dominant channel for dust grain formation is thought to the dying medium mass stars, from the cooling material in the shed outer layers of such stars. Their optical properties can be changed when found in the cold medium where a mantle of hydrogen ices can build up around each of the dust grains. Dust grains are destroyed by warm and hot gas. At these high temperatures, the kinetic energy of the individual gas particles carries enough energy to break off parts of the grains and the PAHs. This is a slow process and there are dust grains in these regions for some time after being exposed to these harsher conditions. The key point in the formation and destruction properties is that local conditions in the ISM and radiation field can affect the amount of opacity for the dust grains and the value of the index β. Star forming regions tend to be dusty with larger dust grains than the warm and hot regions of galaxies.

Figure 4.7 shows off the emission in the optical and mid-infrared portions of the SED. The three-colour optical image shows the red and blue colours. The centre of the galaxy is dominated by a red bulge and the outer disk of the galaxy is seen as a blue, young stellar population. If you examine the optical image, you can see dust lanes where the dust is extinguishing the light behind it and those same regions show up in emission in the infrared.

This long digression on dust still centres around radiation reprocessing: the ISM is changing the light distribution of galaxies through absorption and reemission of the light in the galaxy. Just like hydrogen gas blocks and reradiates ionizing photons, dust blocks optical starlight and reradiates it as modified blackbody emission throughout the ISM. Figure 4.8 show the effects of dust reprocessing of a galaxy's SED. The dust curve shows how the optical spectrum of stars is blocked by the dust emission: note how the shorter wavelength starlight is blocked by a larger factor than the longer wavelength starlight as we would expect from reddening. The galaxy then reradiates this light at $\lambda > 10^4$ nm through dust emission. The line

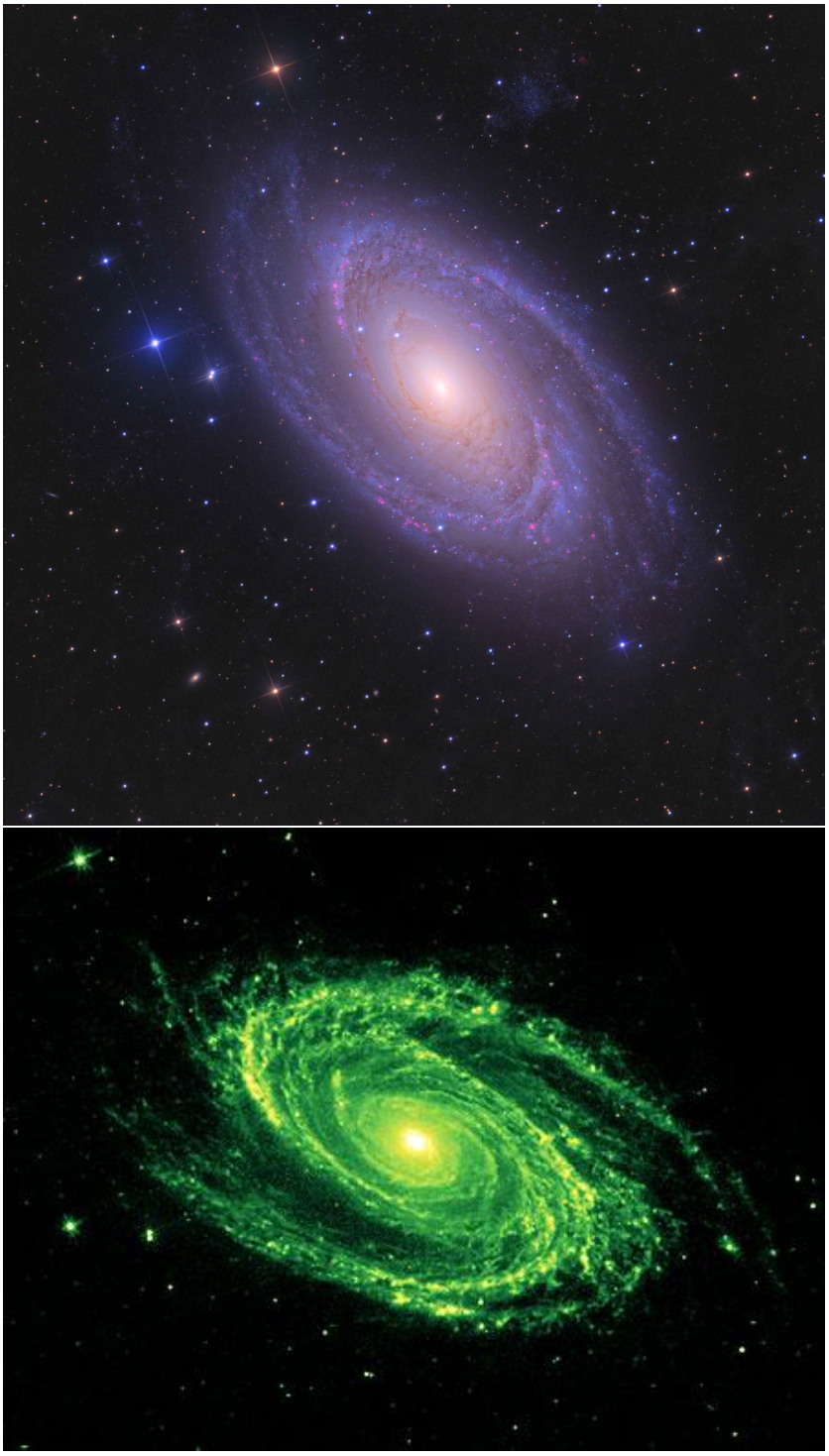


Figure 4.7: M81 seen in the optical (top) and mid-infrared (bottom). The different colours highlight the different stellar populations in the optical, but there are prominent dust lanes seen in the galaxy. Those dust lanes, seen in extinction in the optical, are seen in emission in the mid-infrared.

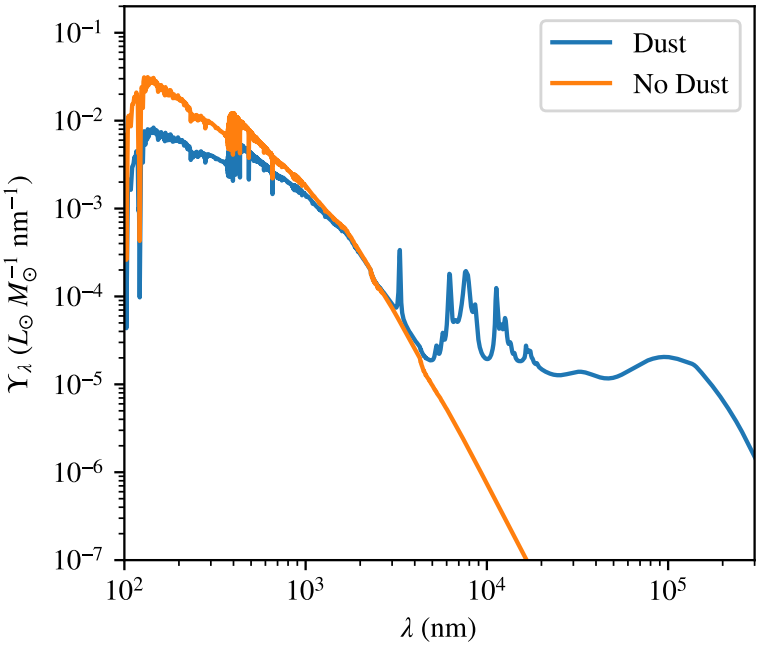


Figure 4.8: Simple stellar populations with and without the effects of dust included. The models show a stellar population with an age of 100 Myr and solar metallicity.

features visible in the dust spectrum at $\lambda \sim 10^4$ nm are the spectral features of PAHs and the modified blackbody emission is visible at longer wavelengths.

Key Points

- Photoionized regions or H II regions are the regions around high mass (and thus high temperature) stars where high energy photons ionize the gas which is balanced by the positive ions and the electrons recombining (Equation 4.4).
- The observed spectrum of photoionized regions is characterized by the presence of emission lines, most notably the spectral series of hydrogen 4.3.
- The size of photoionized regions is modelled using the Strömgren sphere derivation (Equation 4.5), which is an idealized representation of how much gas a UV radiation source can keep ionized.
- The net effect of H II regions are to reprocess radiation from harsh ultraviolet radiation into lower energy photons.

- Dust plays a similar role in reprocessing radiation into the infrared where the dust grains show cross sections that vary with wavelength which also affects their observed spectrum. Typically dust luminosity grows faster than a blackbody with temperature $L_{\rm dust} \propto T^6$ vs the blackbody with T^4. Thus dust is a very efficient radiator and reprocessor of radiation.

4.3 *Case Study: The Temperature of the Ionized and Neutral Media*

In Table 4.1, we note that the main difference between the phases of the ISM is their temperatures. We now can explore what sets that temperature. When we discuss the temperature, we mean the average kinetic energy of an individual particle in that phase of the gas, thus, $\langle K \rangle = \frac{3}{2}kT$ on a per-particle basis is a definition of the temperature T. Hence, heating mechanisms increase the kinetic energy of the gas and cooling mechanisms reduce the kinetic energy. Most the heating in the ISM comes from three main channels: (1) radiation fields, usually from stars, (2) cosmic rays, and (3) shock waves. Cooling mechanisms are usually channels to convert the kinetic energy of particles into radiation that can then leave the system. The H II region is a classic example a balance between heating and cooling and one of the main questions we want to ask is what is the temperature of the ionized gas inside an H II region. The gas is being heated by photoionization: by having a neutral atom at rest get hit with a photon, the kinetic energy of the released electron then goes into heating the gas. The electron whizzes off and collides with other particles in the plasma exchanging kinetic energy and increasing the temperature of the gas, on average. In contrast, the recominbation lines of hydrogen are a cooling mechanism: a free electron travels by a proton and gets captured, and then radiates energy in the form of photons leaving the system. On average, if the rate of heating is higher, the temperature of the gas will increase. The higher kinetic energy increases the frequency of proton-electron flybys (because the gas is moving faster), leading to an increased cooling rate. There is thus an equilibrium temperature where for a given rate of heating, the cooling will balance the heating.

COOLING – We will now make this more concrete by comparing the heating rate from photons ionizing the gas to the cooling rate from spectral line emission. However, we have omitted a detail in our treatment so far that becomes important at this point: there are

metals in H II regions and those metals are important for the cooling of the gas. For example, we frequently observe the spectral lines of ionized oxygen from H II regions. These lines of O II or O III (i.e., O^+ or O^{2+} respectively) are significant coolants. Figure 4.9 shows the some of the energy levels of these two ionization states of the oxygen atom and the spectral transitions in those atoms that are important for cooling. Note that many of the spectral lines are visible in the optical and near infrared.

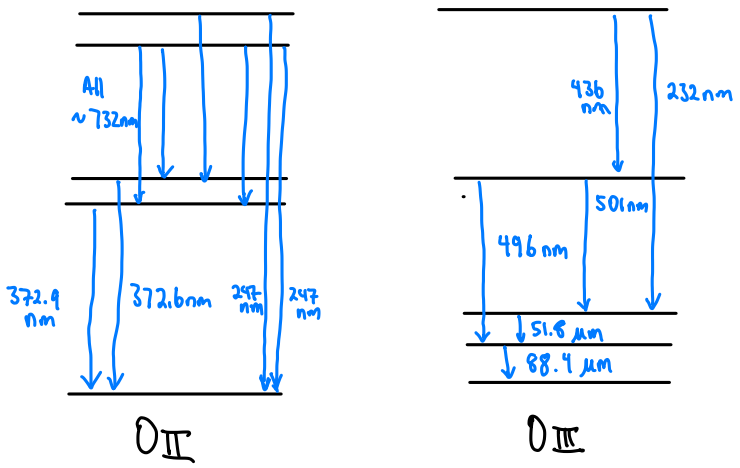


Figure 4.9: Spectral lines of ionized oxygen.

The physical mechanism for these spectral lines cooling is from electrons colliding with the ion and exciting the electron to a higher energy state followed by an electron deexciting in the oxygen ion and giving off a photon.

$$\begin{aligned} e^- + O^+ &\rightarrow e^- + O^+_* \\ O^+_* &\rightarrow O^+ + \gamma \end{aligned}$$

Here O_* indicates an excited state of the oxygen ion. The net effect of this process is that the initial kinetic energy of the electron is converted into the photon energy, and then the photon energy radiates away leaving the system representing a net cooling of the gas.

This cooling rate depends on collisions between the electrons and the oxygen ions so we can model the cooling rate for a specific spectral line of oxygen as $\epsilon_{\rm cool}$ in units of power per unit volume as

$$\epsilon_{\rm cool} = \Lambda_O n_e n_O \tag{4.7}$$

where Λ_0 represents all the microphysics of the electron collision, n_e is the density of electrons and n_O is is the number density of oxygen ions. Note that this equation has the same basic physical scaling as

the recombination rate (Equation 4.3). We can then model the cooling rate Λ_O as $h\nu\langle\sigma v\rangle$ where $h\nu$ is the photon of energy of the cooling transition and $\langle\sigma v\rangle$ average over the collision cross section σ. The average is also carried out of the velocity distribution of the colliders, i.e., the Maxwell-Boltzmann velocity distribution. The reason that the v of the colliders comes inside the average is that the cross section shows a lot of variation with the particle velocities because of resonant interactions. This term $\langle\sigma v\rangle$ is something that can be tabulated for each collision to a specific energy level. The physical meaning of this term is the volume swept out by a target moving at speed v per unit time. To find a collision rate we multiply that volume per time by the number of colliders per volume (to get a rate) and then by the number of targets (to get a rate per volume).

The actual mechanics of calculating all the possible cooling rates is complicated since it requires tabulating all the possible interactions between electrons and ions over all the ionization states and all the possible spectral lines. However, the formulation for the cooling rate means that we just add up all the different cooling rates to get $\epsilon_{\text{cool}} = \Lambda_O n_e n_O$. We can then take $n_e \approx n_H$ and $n_O \approx x_O n_H$ where n_O is the fractional abundance of oxygen by number (typically, $\lesssim 10^{-3}$, see Figure 2.15). Then, these cooling functions can be tabulated in terms of physical parameters including all the other cooling channels like, e.g., nitrogen and carbon cooling lines leading to a cooling rate set by the particle density with all the gory details swept into the cooling function: $\epsilon_{\text{cool}} = \Lambda(T, x_O, x_N, x_C, \cdots) n_H^2$. Figure 4.10 shows different cooling rates for the ions in a typical H II region. Cooling functions tend to follow the functional form $\Lambda \propto T^{-1/2}\exp(-\Delta E/kT)$ where ΔE is the energy level of the top state of the internally excited spectral line.

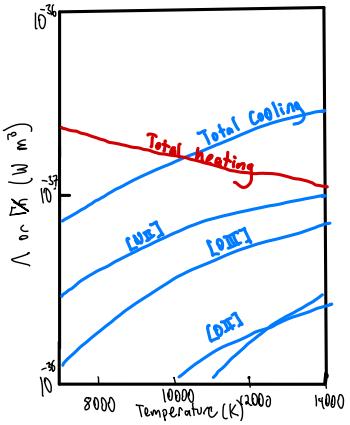


Figure 4.10: Heating and Cooling rates in H II regions based off figures in Osterbrock and Ferland [2006].

HEATING – We carry out a similar process to calculate the heating rate in terms of the particle densities. In H II region, the heating is dominated by the kinetic energy carried by photoionized electrons. In this case, we model $\epsilon_{\text{heat}} = \Gamma n_H$ where Γ is the heating rate per particle from the ionizing radiation of the star. Here, we can calculate this heating rate a little more precisely for hydrogen. During photoionization, we have $\gamma + H^0 \rightarrow p^+ + e^- + K$ where K is the kinetic energy for that photoelectron. We know that $K = h\nu - E_0$ where E_0 is the binding energy of hydrogen. Then, the heating rate will be, integrating over the ionizing radiation of the star:

$$\Gamma n_H = n_H \int_{E_0/h}^{\infty} [f_\nu / h\nu] \sigma_H c (h\nu - E_0) d\nu \tag{4.8}$$

This expression may *look* awful but it can be pulled apart into its constituent pieces. The expression $f_\nu/h\nu$ converts the flux density from

the star passing through a surface into photons passing through a surface following the reasoning on Homework 2. The $\sigma_H c$ term is the same as the $\langle \sigma v \rangle$ seen in the cooling rate, and this cross section ends up depending on frequency[5]. Finally $h\nu - E_0$ is the kinetic energy released in each photoionization after subtracting off the ionization energy of the hydrogen atom. The bounds of the integration run from the frequency required to ionize hydrogen up to ∞ since only these photons will end up contributing to photoionization heating. We can make approximations to carry out this integral, but the important thing is that it depends only weakly on the temperature of the gas. This is generally true of heating process: they tend to depend on the properties of some background field (the stellar radiation field in this case) and a single target being hit by the field. Thus, heating rates tend to depend on the density to the first power and be relatively insensitive to the conditions of the ISM. The data in Figure 4.10 show the true heating rate, which includes the effects of metals and Helium. The heating rates of these species decrease with increasing temperature because the higher temperatures lead to higher ionization states, which cannot be readily photoionized and thus stop contributing to the heating rate.

[5] Specifically, $\sigma_H = 6.3 \times 10^{-22}\ \mathrm{m}^2 (\nu/\nu_0)^{-3}$

In contrast with the heating rate, the cooling rate will depend on n_H^2 and tend to increase with temperature because higher temperatures can excite more internal states of the atoms responsible for cooling. Figure 4.10 shows the balance between heating and cooling in a photoionized region. The point where the two curves intersect is the equilibrium temperature in the H II region, which shows that the model used above, even if not presented in details, can predict a temperature of $T \sim 10^4$ H for H II region as is expected from the ionized ISM.

General Application – The physics of heating and cooling is a great tool for understanding the phase structure of the ISM. For example, neutral medium comes in two distinct phases: cold ($100 < T/\mathrm{K} < 3000$) and warm ($6000 < T/\mathrm{K} < 10,000$). Why are there then two states the neutral medium. This comes from balancing the different heating and cooling mechanisms in this phase. In this case, the heating doesn't come from ionizing photons. Most of these photons are absorbed by recombining atoms inside the ionizing region. Instead, these region are heated by cosmic rays and a diffuse background of starlight called the *interstellar radiation field* (ISRF). The ISRF carries a lot of photons in the near and far ultraviolet, which can move through the neutral gas. These photons are most likely to be absorbed by dust grains, and then they eject an electron carrying kinetic energy, completely analogously to how ionized regions are

heated. This process is called the grain photoelectic effect, since it looks just like the photoelectric effect in metals that was a fundamental experiment in quantum mechanics.

The cooling in the ISM follows the same structure as collisional cooling in HII regions. Here, instead, the cooling comes from exciting fine structure lines within neutral atoms or atoms with ionization potential less than 13.6 eV. The main cooling lines in the neutral medium are [O I] at $\lambda = 63$ μm, [C II] at 122 μm. We put the square brackets around these lines since they are technically forbidden transitions in that they don't follow the selection rules of regular electronic transitions in atoms. However, nothing in quantum is truly forbidden, so these occur rarely.

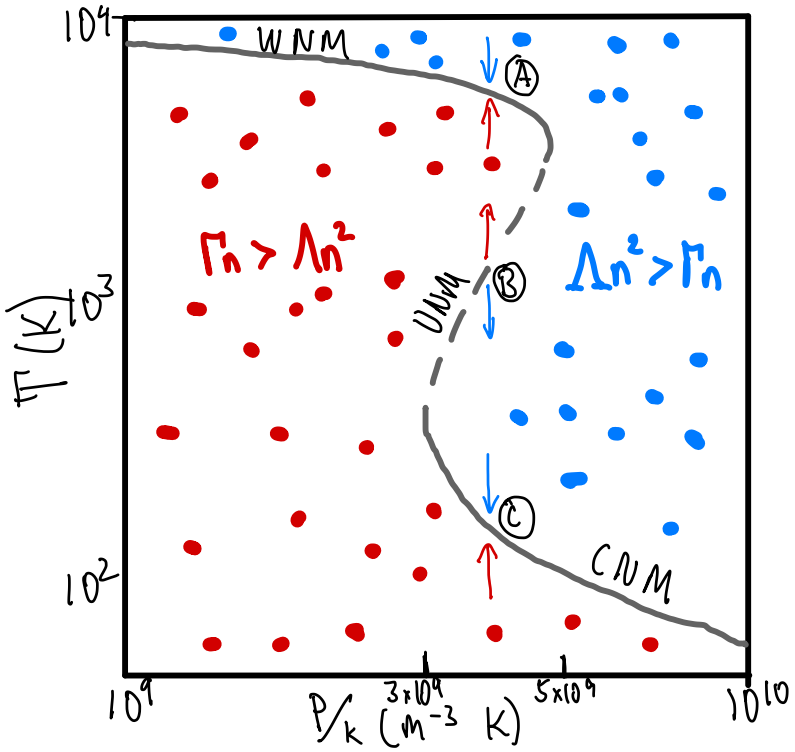


Figure 4.11: Two-phase model for the neutral ISM. The grey line shows the locus where $\Lambda n_{\rm H}^2 = \Gamma n_{\rm H}$.

If we set up the heating and cooling curves for these process (cosmic rays, grain photoelectric effect, cooling by fine structure lines in metals, we arrive at a model for the two phases in Figure 4.11. This figure plots the line where $\Lambda n_{\rm H}^2 = \Gamma n_{\rm H}$, which is the equilibrium line between heating and cooling. The axes here show the temperature of the gas as a function of the pressure (divided by the Boltzmann constant) since this pressure quantity is nearly uniform across the ISM with a value of about 4×10^9 m^{-3} K in the Milky Way. This equilibrium line divides the space into a region where heating is dominant

($\Gamma n_{\mathrm{H}} > \Lambda n_{\mathrm{H}}^2$) in red and where cooling is dominant in blue. The S-shape of the curve leads to the two-phase behaviour of the neutral medium. We can consider this based on studying three separate points in the curve, labelled A, B, C. Gas at point A is in the warm neutral medium with a temperature near $T = 8000$ K. If a perturbation causes the gas to heat up (i.e., move upward from this point in the graph), it will move into a regime where the cooling is dominant so it will return back to the equilibrium point (blue arrow at point A). Similarly if it is cooled, this moves it into the regime where heating is dominant, so it will heat back to point A (red arrow at point A). A similar behaviour occurs for the bottom branch of the graph, which shows the location of the Cold Neutral Medium. At point C in the CNM regime, a similar behaviour occurs as for point A: temperature perturbations will be restored to the equilibrium temperature.

The weird stuff happens at point B, which is called the Unstable Neutral Medium. If a perturbation cools the gas from point B (blue arrow at point B), it enters a region where the cooling dominates the heat so this will pull it father away from the equilibrium line, leading to it cooling all the way down until it becomes part of the CNM. Similarly, warming up from point C leads to the gas continually heating until the gas becomes part of the warm neutral medium. Thus, steady state conditions, no gas should be found in the UNM. Of course, nothing is so neat and simple in the ISM. Almost 20% of the mass in the neutral medium can be found in the UNM which means that it is out of thermal equilibrium and has been displaced there recently. To quantify how recently this displacement must have occured, we can look at the global heating and cooling functions for gas in the ISM.

Figure 4.12 shows the effects of heating and cooling functions across a wide range of temperatures in the ISM accounting for a vast range of physical processes. The absolute values of these curves give some indication of the cooling time for gas. For example, to consider the cooling time for gas in the UNM as above, we can read off the cooling rate for gas at $T \sim 5000$ K: $\Lambda = 3 \times 10^{-37}$ Wm^3. We can then assume a number density for the gas, say $n_{\mathrm{H}} = 10^6$ m^{-3}. From there, we can calculate the cooling time for the gas, which is the time to radiate away energy under a simple linear approximate: $t_{\mathrm{cool}} = E_T/(\Lambda n^2)$. Here, $E_T = \frac{3}{2}nkT$, the thermal energy density of a

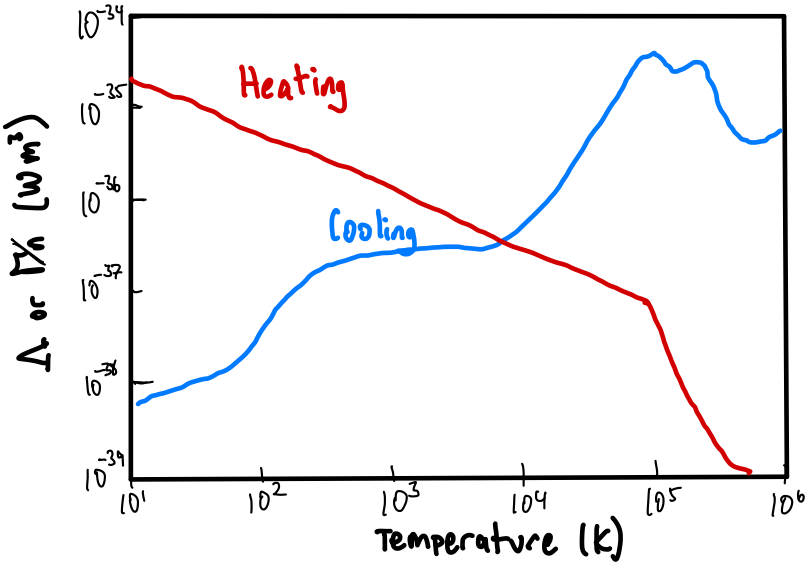


Figure 4.12: Heating and Cooling in the ISM.

gas. Then,

$$t_{\text{cool}} = \frac{E_T}{\Lambda n^2} \tag{4.9}$$

$$= \frac{3kT}{2\Lambda n} \tag{4.10}$$

$$= \frac{3(1.38 \times 10^{-23})(5000\ \text{K}}{2(3 \times 10^{-37}\ \text{W m}^3)(10^6\ \text{m}^{-3})} \tag{4.11}$$

$$= 11,000\ \text{yr} \tag{4.12}$$

This means that gas found in the UNM has been perturbed in the past $\sim 10^4$ yr.

Key Points

- The cooling of the ISM is dominated by the transfer of thermal energy of particles (e.g., electrons) exciting internal energy states in ions, atoms and molecules (e.g., oxygen ions). These excited particles then de-excite, radiating the thermal energy away as a photon. Because this relies on a two-body collision, cooling rates typically have the form Λn^2 where n is the number density of particles.
- The heating of the ISM is predominantly driven by the radiation output of stars heating the gas. Because this only depends on a radiation field and the number of particles, heating rates typically have the form Γn where n is the density of particles.

- The equilibrium temperature of the interstellar medium is set by where heating rates equal the cooling rates.
- The neutral atomic medium, in particular, has a heating and cooling curve that leads to two stable thermal states: a warm and a cold state. Even so, a significant fraction of the ISM's mass is found in the unstable state showing that the ISM is frequently disturbed from equilibrium.
- The cooling time for gas is the characteristic timescale for a gas to lose its thermal energy (Equation 4.9).

4.4 *Case Study: Supernova Explosions*

Our final case study looks at the importance and nature of fluid flows in the interstellar medium though the influence of a supernova shock. Supernovae are one of the important cases for studying stellar feedback, which measures how young, massive stars inject energy and momentum back into the galaxy after their formation. A supernova explosion can be explored in the simple case as "what happens to a diffuse cloud of gas when you suddenly inject $E_{\rm SN} = 10^{44}$ J of energy into the gas?" Specifically, this energy is mechanical energy in the motion of fluid. Recall that roughly $100\times$ this energy is released in neutrinos and $0.01\times$ this energy is released as photons, which are what we see when we observe a distant supernova explosion.

This sudden injection of energy leads to an explosion from the site of the SN. Specifically, the gas will start moving outward at very high speeds and these speeds are significantly higher than the local sound speed, c_s. For the study of shocks, relevant sound speed is the *adiabatic sound speed*

$$c_s = \sqrt{\frac{\gamma k T}{m}} \tag{4.13}$$

where γ is the adiabatic index taken as $\gamma = 5/3$ for a monatomic ideal gas, k is the Boltzmann constant, T is the temperature and m is the particle mass.

The other option, physically, that we could consider here is the isothermal sound speed, which is Equation 4.13 with $\gamma = 1$. The isothermal sound speed is appropriate when the compressions induced by a pressure fluctuation can efficiently exchange heat with their environment. In the case of astrophysical systems, this means through radiation, i.e., they can radiate away their energy or get heated by the environment. Since the heating and cooling times can be relatively fast in some environments (see 4.3 below), sound waves in that case would be isothermal. However, for the case of supernova

explosions, the adiabatic sound speed is appropriate because the shock wave that is generated initially compresses the medium and moves through it so quickly that there is no time for heat exchange.

Returning to the actual supernova shock wave physics, the wave front is propelled outward by the mechanical energy injection, leading to a front moving at highly supersonic speeds. We refer to these supersonic speeds in terms of the sonic Mach number $\mathcal{M}_s \equiv v_{\rm front}/c_s$ where $v_{\rm front}$ is the speed at which the shock front is moving outward. For supernova $\mathcal{M}_s \gg 1$. Physically, what is happening is that the fluid develops a shock front that is moving so fast it arrives before pressure (sound) waves that would move through the fluid as a result of the incoming disturbance. Physically, sound waves are a mechanism reorganize fluids to accommodate disturbances. When you clap your hands, the sound waves propagating outward are the downstream response of air molecules getting out of the way from in between your closing hands.

SHOCK WAVES – Shock waves represent discontinuities in the properties of fluid flows. There are sharp jumps in fluid density, velocity, and energy that happen over sufficiently short timescales that they appear to be a discontinuous jump. Discontinuities are not physical, but what is really happening is that properties of the flow are changing on length scales that are significantly shorter than the mean free path of a particle in the fluid, meaning that the gas is not actually behaving as a fluid. The full definition of a fluid requires all length scales in the problem to be significantly longer than the mean free path. We deal with this in the study of the ISM by ignoring the details of what is actually happening in the shock and just treat the up and down stream sides of the shock.

Figure 4.13 shows an image of a supernova shock wave propagating into the ISM. The shock is moving from the lower right corner of the image toward the upper left. The image is taken in [O III] (blue) and [N II] (red) emission lines in the optical portion of the spectrum, which shows the shock wave cooling off.[6] The shock wave itself is actually slightly upstream from the emission and the light is coming from the area behind the shock. The structure nonetheless illustrates that shocks are quite thin and have a lot of heavy duty physics happening inside of them. Astrophysics usually approximates the shock waves by considering the conservation laws of mass, energy, and momentum far away from the shocks. This treatment is straightforward, but we will skip over it here in the interest of time. However, I'll note where the improved treatment leads to modifying our understanding below.

[6] Since it is cooling, the shock is not adiabatic, so we will get to exactly what is happening here later.

Figure 4.13: The Veil Nebula as imaged by HST. Image credit: ESA/Hubble & NASA, Z. Levay

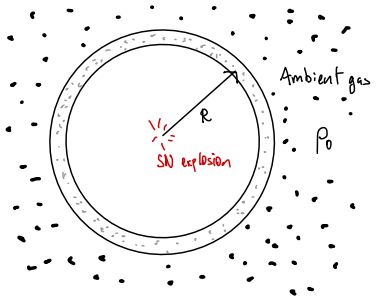


Figure 4.14: Setup for the Sedov solution.

THE SEDOV SOLUTION – Here, we solve for the rate of expansion of a supernova remnant into the interstellar medium. We will make the assumption that the shock front is adiabatic so that it conserves energy as it propagates. Figure 4.14 shows the setup for the Sedov explosion. The supernova deposits an energy of $E_{\rm SN} \approx 10^{44}$ J into the local ISM. The ISM has a mass density of $\rho_0 = \mu m_{\rm H} n_{\rm H}$. Here, $m_{\rm H}$ is the mass of a hydrogen atom and μ is the mean particle mass. For pure hydrogen gas, $\mu = 1$. For ionized hydrogen $\mu = 0.5$ and for pure molecular hydrogen gas, $\mu = 2$. Including helium and metals makes neutral gas have $\mu = 1.4$ and molecular gas has $\mu = 2.4$. The supernova shock front is a thin shell a distance R from the site of the initial supernova explosion. The size of the supernova shock is increasing to R is a function of time t. We want to find out $R(t)$ given the properties of the explosion and the local medium.

The nature of the supernova shock is that it when it hits the ambient ISM, the shockwave drives the ambient medium outward in an explosion, accelerating it from rest to moving outward with speed $V = dR/dt$. Thus, the mechanical energy of the initial explosion translates into the kinetic energy of the mass swept up into the supernova shell. We can relate the explosion energy to the kinetic energy of the shell: $E_{\rm SN} = \frac{1}{2}MV^2$ where M is the mass in the shell. The amount of material in the shell from the initial star turns out to be relatively small. Instead, we consider the mass of the shell to be the

total mass of material cleared out by the explosion:

$$M = \rho_0 \frac{4}{3}\pi R^3$$

Then, the kinetic energy of the shell can be related to that of the intial explosion:

$$\begin{aligned} E_{\rm SN} &= \frac{1}{2}\left(\rho_0 \frac{4}{3}\pi R^3\right) V^2 \\ &= \frac{2\pi}{3}\rho_0 R^3 \left(\frac{dR}{dt}\right)^2 \end{aligned} \tag{4.14}$$

The approximation we are making by ignoring the details of the shock means that the prefactor of $2\pi/3$ should actually be $3\pi/4$ if we included the full shock physics. The difference in the amount of energy going into the kinetic energy is that this treatement neglects the heating of the gas. An equal amount of energy goes into heating up the material, ultimately increasing the temperature to $T > 10^6$ K. If you look up other solutions to the supernova explosion problem, this is why our answer is different.

From Equation 4.14, we isolate for the kinematic variables:

$$R^3 \left(\frac{dR}{dt}\right)^2 = \frac{3E_{\rm SN}}{2\pi\rho_0} = \text{const.}, \tag{4.15}$$

which is a differential equation that can be solved for $R(t)$. Any time we see a product of a variables and its derivatives, then a trial answer of $R(t) = Bt^a + C$ is a good trial solution. We plug this solution into the equation of motion and work out a solution. First, we note that $R \to 0$ and $t \to 0$ (the supernova stars out as being very tiny), so that C=o. Furthermore, $dR/dt = Bat^{a-1}$ so that

$$\frac{3E_{\rm SN}}{2\pi\rho_0} = (Bt^a)^3 \left(Bat^{a-1}\right)^2 \tag{4.16}$$

$$= B^3 t^{3a} \cdot B^2 a^2 t^{2a-2} \tag{4.17}$$

$$= B^5 a^2 t^{5a-2}. \tag{4.18}$$

From this last equation, we can set $5a - 2 = 0$ because we know that the left hand side of the solution is a constant to the right hand side must scale like t^0. Then, we know that $a = 2/5$. We can then substitute this in to solve for B:

$$\frac{3E_{\rm SN}}{2\pi\rho_0} = B^5 \left(\frac{2}{5}\right)^2 \tag{4.19}$$

$$= B^5 \frac{4}{25}, \text{ so} \tag{4.20}$$

$$B^5 = \frac{75E_{\rm SN}}{8\pi\rho_0} \tag{4.21}$$

$$B = \left(\frac{75E_{\rm SN}}{8\pi\rho_0}\right)^{1/5}. \tag{4.22}$$

This leads to our final results for the radius and the velocity:

$$R(t) = \left(\frac{75E_{\rm SN}}{8\pi\rho_0}\right)^{1/5} t^{\frac{2}{5}} \quad (4.23)$$

$$V(t) = \frac{2}{5}\left(\frac{75E_{\rm SN}}{8\pi\rho_0}\right)^{1/5} t^{-\frac{3}{5}}. \quad (4.24)$$

This answer makes good sense. The radius of the shell increases over time but slows down as it gathers more mass. The $1/5$ power on the constant means that the dependence on the supernova energy is relatively weak. To double the speed of the explosion requires $32\times$ as much input energy.

The Radiative Phase – As noted, the full treatment of shock physics both accelerates the gas but also heats up the material to high temperatures. The initial heating is to very high values but the supernova shock cools down to $T < 10^6$ K. According to Figure 4.12, this actually increases the heating rate significantly so that the shocked material ends up cooling down through radiation, which leads to faster cooling. Inside the shock, hotter gas is pushing outward with higher pressure and the net effect is that this compresses the radiating shock layer. The compression raises the density, the shell thins out, and cools quickly to temperatures near $T \sim 10^4$ H where the cooling curve drops off again. This is why the emission lines seen in Figure 4.13 are actually the same cooling lines we see in H II regions.

Under such conditions, the energy conservation that was key to the Sedov solution no longer holds. Instead we move to a phase where the momentum of the material is conserved. Again, the shock wave is sweeping up material into a shell, except this time the shell is thin because of the pressure behind the front. The thin shell is often said to be in the "snowplow phase" because it is sweeping up material. Here, we can assert that the total momentum of the material is constant so that:

$$\rho_0 \frac{4\pi}{3} R^3 V = \text{const.} \quad (4.25)$$

We then assume that the thin shell phase starts from some initial radius R_0 and time t_0, which is where the shell gets to in the energy-conserving phase. We can set the momentum as being the same as the momentum at that time:

$$\rho_0 \frac{4\pi}{3} R^3 V = \rho_0 \frac{4\pi}{3} R_0^3 V_0 \quad (4.26)$$

$$R^3 \left(\frac{dR}{dt}\right) = R_0^3 V_0 \quad (4.27)$$

$$\left(\frac{R}{R_0}\right)^3 \frac{1}{V_0}\frac{dR}{dt} = 1. \quad (4.28)$$

If we define a variable $r \equiv R/R_0$, this becomes a separable differential equation:

$$\frac{R_0}{V_0} r^3 \frac{dr}{dt} = 1, \text{ so} \tag{4.29}$$

$$\frac{R_0}{V_0} \int_1^r r'^3 dr' = \int_{t_0}^t dt' \tag{4.30}$$

$$\frac{r^4}{4} - \frac{1}{4} = \frac{V_0}{R_0}(t - t_0) \tag{4.31}$$

$$r^4 = 1 + \frac{4V_0}{R_0}(t - t_0) \tag{4.32}$$

$$r = \left[1 + \frac{4V_0}{R_0}(t - t_0)\right]^{1/4} \tag{4.33}$$

$$R = R_0 \left[1 + \frac{4V_0}{R_0}(t - t_0)\right]^{1/4}, \text{ and thus} \tag{4.34}$$

$$\frac{dR}{dt} = V_0 \left[1 + \frac{4V_0}{R_0}(t - t_0)\right]^{-3/4} \tag{4.35}$$

Comparing to the energy conserving phase (Equation 4.24), we see that the shock front expands more slowly, which makes sense because it is losing energy to radiation.

The key variable that sets the transition is the speed of the shock and the transition point is typically taken as $V_0 = 250$ km s^{-1}. By using this in Equation 4.24, we can find the transition point between the energy- and momentum-conserving phases.

THE EVOLUTION OF SUPERNOVA REMNANTS – With these two models, we can ask the question of how long does a supernova remnant actually last and how big does it get. As a reference case, we will consider $E_{\rm SN} = 10^{44}$ J and $n_{\rm H} = 10^6$ m^{-3} for pure, neutral hydrogen at a temperature of $T = 8000$ K. This means that $\rho_0 = m_{\rm H} n_{\rm H}$. We want to understand how big the supernova remnant gets before it merges with the local ISM. The condition for this blending is that the speed of the shock reaches the thermal speed in the gas at this temperature: $c_s = 10.5$ km s^{-1}. Once the remnant reaches those speeds, sound waves will push back on the shock and it will dissipate into the local gas. Substituting these values into Equations 4.24 gives:

$$\frac{R}{1 \text{ pc}} = 0.36 \left(\frac{t}{1 \text{ yr}}\right)^{2/5} \tag{4.36}$$

$$\frac{V}{1 \text{ km s}^{-1}} = 1.4 \times 10^5 \left(\frac{t}{1 \text{ yr}}\right)^{-3/5}. \tag{4.37}$$

This expansion in the energy-conserving phase transitions to the momentum conserving phase at $V_0 = 250$ km s^{-1} (statement without proof). Solving the velocity equation for the time when the shock

slows to this V gives $t_0 = 38,000$ years at which point the radius will be $R_0 = 24$ pc. From there, we can substitute these values into Equation 4.35 and solve for the time when the velocity slows to $V = 10$ km s^{-1} which gives $t_f = 1.7 \times 10^6$ yr at which time $R = 71$ pc. Again, we end up with timescales that are $\lesssim 1$ Myr, a common time scale for evolution in the ISM. Supernova remnants are visible in Figure 4.1, with radii between 20 and 50 pc, indicating their relatively late stage in their evolution.

At any given time, we estimate (on homework) that $\sim 5\%$ of the volume of the galaxy disk is inside a supernova remnant. Given each remnant will last 4×10^5 yr, this implies that, approximately every 4×10^5 yr$/0.05 = 8$ Myr, a piece of the ISM will be hit with a shock from a supernova explosion. This is a relatively short period of time and drives the reason why the ISM is so dynamic: gas clouds cannot reliably settle into equilibria for any timescales longer than the time between shocks. On a galactic scale, this is a short time period compared to the evolution timescale for stars (Myr to Gyr) or the time for galactic dynamic effects to matter (100 Myr).

Key Points

- Shock waves are one of the main agents shaping the interstellar medium. Shocks arise any time gas is moving faster than the local sound speed (Equation 4.13) leading to discontinuous jumps in the conditions of the gas.
- Supernova are one of the main ways of producing shocks in the ISM. They are frequent and high energy, injecting shocks into the gas.
- A supernova shock goes through two main phases: an energy conserving phase early on where the gas does not have time to radiate significantly as it expands and a momentum conserving phase where radiation becomes important. The energy conserving shock front (Equation 4.24) and the momentum conserving shock front (Equation 4.35) both have idealized solutions.
- Shocks keep the ISM from settling into an equilibrium and are one of the main sources of mixing and energy injection into the gas of the galaxy.

4.5 *Summary*

These three case studies summarize several key interactions in the ISM. H II regions and dust grains illustrate the processing radiation field to different parts of the electromagnetic spectrum and the resulting heating and cooling of the gas phases. Supernova remnants show the importance of shock waves in the ISM and imply that the ISM is dynamic and frequently stirred up by the passage of shock waves. These case studies ignore several other effects that are important in a full treatment of the ISM: magnetic field, plasma physics, and chemistry just to name a few. We have not discussed the full role of the ISM in the context of galaxy evolution and will return to the questions of star formation and accretion from the intergalactic medium later when we discuss the evolutionary processes of the system as a whole. However, these three case studies give us the tools and vocabulary we need to understand most of the key physics in the ISM.

5
Phenomenology

So far, we have covered observational astrophysics and the two visible constituents of galaxies: stellar populations and the interstellar medium. We now have the pieces needed to interpret the observations of galaxies, both our own Milky Way and more distant systems.

We should begin with a definition of a galaxy to codify the idea that galaxies are indeed "things." Broadly speaking, galaxies are discrete, gravitationally bound collections of (in order of mass contribution) dark matter, stars, gas, dust, and high energy particles. The gravitational binding is key: it makes galaxies into discrete objects that we can actually inventory and describe and separates them from "phenomena" which can be easy to identify but difficult to rigorously define (such as the spiral arms in galaxies).

The dominant mass contribution in a galaxy comes from its dark matter halo. Dark matter is (thus far) dark and we can only infer its presence from its gravitational interaction with baryonic matter. However, galaxies and galaxy clusters provided the earliest evidence for the existence of dark matter setting off the continuing search for what dark matter actually is. By process of elimination, we believe that dark matter follows the weakly interacting massive particle paradigm where dark matter actually is a particle with a large mass scale that is travelling non-relativistically with typical speeds of $v_{\rm DM} \sim 300$ km s^{-1}. This typical speed also comes from the study of galaxies in the context of *cosmology*, i.e., the study of the evolution of the Universe. If dark matter was travelling significantly faster, it would spread out more across the Universe and thus would not form deep gravitational wells into which gas can fall and form into galaxies. Thus, we simply wouldn't have the galaxies that we observe today if dark matter travelled quick, which is referred to as "hot dark matter."

Instead, the galaxy population that we see critically relies on dark matter being "cold" and our current model for galaxy evolution is called the Λ-Cold Dark Matter (ΛCDM) paradigm. The Λ is the

standard variable used to represent the cosmological constant in the general relativity equations that govern the Universe, and is the component called *dark energy*. This dark energy is a relatively recent innovation from the late 1990s which was invoked to explain the observation that the Universe's expansion was apparently accelerating. Such acceleration could be explained mathematically by sticking what is essentially an integration constant into the equations that describe the Universe's evolution[1] Placing that constant in the equations yielded solutions that matched the Universe's expansion and, through the theory of General Relativity, the geometry of the Universe. However, physicists do not actually know what Dark Energy actually is. For purposes of understanding galaxies, the main thing that the cosmological constant does is tune our models so that they can reproduce observations the expansion of the Universe after the Big Bang. This expansion the rate at which defines matter is being pulled apart from cosmological expansion and the initial conditions of what gravitation must overcome to form and build up galaxies. Dark matter is far more important to individual galaxy evolution.

[1] More on this in ASTRO 430.

Galaxies are fundamentally defined by dark matter halos since these establish the gravitational potential of the system and the baryonic matter that we observe is the clichéd tip-of-the-iceberg that indicates where dark matter halos are found. It is an ongoing study as to how the dark matter halos are sub-structured and how they evolve over the age of the Universe (though the final portion of this course will address those questions).

Because of the darkness of said matter, we must focus our study of galaxies on the matter that we can see, inferring what we can about how that matter and the dark matter evolves. In this context, we focus on the baryonic matter that is luminous: the stars and ISM. While the behaviour of this matter does not feed back and shape the evolution of dark matter halos significantly, it does undergoing significant interactions and evolution of its own. Furthermore, the luminous baryonic matter is found deep within the dark matter halos. There, the density of dark matter is comparatively low and the behaviour of the luminous matter is shaped primarily by its own self-interactions (gravity, stellar feedback, star formation, etc.).

5.1 A Generic Galaxy

Figure 1 shows a generic (awesome) galaxy with its basic components labelled. We usually discuss galaxy components in terms of the galaxy *bulge*, *disk*, and *halo*. These components are governed by different physical processes and are thus frequently treated independently.

The galaxy bulge consists of a spheroidal distribution of stars. In

Figure 5.1: The Sombrero Galaxy (NGC 4050) which highlights the basic parts of a galaxy. Image credit: NASA, ESA, and The Hubble Heritage Team STScI/AURA).

general, these are relatively red in their optical colours and are *dynamically hot*, which means that the motions of the stars in the system have random motions that are comparable in magnitude to their ordered motions. The red colour indicates a older stellar population ($> 10^9$ yrs since significant star formation). The bulge also hosts one or more supermassive black holes (SMBH) at the centre of the galaxy. SMBHs have masses $10^4 < M_\bullet/M_\odot < 10^9$ and larger galaxies host larger black holes. Observations of the stellar mass of the bulge is proportional to $M_\bullet$, which is fundamentally suspicious. While the black hole is huge on a mass scale, it is incredibly compact. The Schwarzschild radius of black hole is

$$R_S = \frac{2GM}{c^2} = 3 \text{ km} \left(\frac{M}{M_\odot} \right) \tag{5.1}$$

which implies that the $4 \times 10^6 M_\odot$ black hole in our Milky Way has $R_S = 1.2 \times 10^7$ km, i.e., only $\sim$ 0.1 AU. The scale of the bulge is vastly larger (kpc scale or almost 10^8 in length scale) and the black hole's gravitation does not strongly influence the stars except within $\approx 10 R_S$ or so. Thus, *something* must be connecting these scales together and we will need to explore why there is this intimate link between such different mass and length scales in galaxies.

In many galaxies, the bulge will be collocated with a *bar* feature that consists of standing dynamical pattern that looks like a long, linear feature of stars. Figure 5.2 shows an example barred galaxy.

Figure 5.2: The barred galaxy NGC 1300. Image credit: NASA, ESA, and The Hubble Heritage Team STScI/AURA).

The interstellar medium of bulges tends to be dominated by the warm and hot ionized media though colder phases (neutral and molecular gas) are frequently found associated with the bar.

The disk of the galaxy is "dynamically cold" which means that the random motions of stars have magnitudes that are small compared to

the rotational motion around the centre of the galaxy. This statement means that the orbits of most disk stars are well aligned and all the stars in the disk are generally travelling in one direction around the galaxy's centre, albeit at different speeds depending on where the stars are relative to the galactic centre. Disks have most of their matter concentrated into a relatively thin layer with a typical scale of 300 pc for the stars and 100 pc for gas. Disks contain significant populations of older stars, but they also tend to host the cool and cold phases of the galaxy's ISM. Phases of the ISM are described by the chemical and thermal state of hydrogen. The cool/cold phases of the ISM consist of hydrogen in its atomic and molecular (H_2) states with characteristic temperatures of $T < 10^4$ K. The coldest phases of the ISM ($T < 10^2$ K) are the sites of star formation.

The cool/cold ISM has about 1% of its mass in dust. In the optical, dust features are visible because they block out light from the galaxy's stars that are located behind the dust lanes. The dust extinction is one of the easiest ways to identify the presence of significant cool / cold gas in galaxies. The disk feature that is easiest to identify in Figure 5.1 is the dust lane.

Finally, the halo is the faintest component of the galaxy and is typically difficult to discern in most imaging of galaxies. Though the halo contains a relatively small fraction of galaxy mass, it has the largest spatial extent of the galaxy components and is the dynamically "hottest" component of a galaxy where star motions are almost entirely random. The halo also hosts *globular clusters*, which are uncommonly dense and old clusters of stars. Both the halo and the globular cluster population appear to be the oldest parts of galaxy, making them a key target for studies of galaxy formation and evolution.

It is important to remember that these components can be collocated. For example, there are halo stars currently found in the disks of galaxies. In our own Milky Way, we can find these halo stars in the Solar neighbourhood by looking for stars with low metallicity (e.g., $[Fe/H] < -2$) with large proper motions that indicate they are travelling on an orbit that is heavily inclined with respect to the galaxy's rotational motion.

Property	Range	MW
Baryonic Mass	$10^7\ M_\odot$ to $10^{11}\ M_\odot$	$5 \times 10^{10}\ M_\odot$
Dark Matter Mass	$10^9\ M_\odot$ to $10^{14}\ M_\odot$	$1 \times 10^{12}\ M_\odot$
Luminosity	$10^6\ L_\odot$ to $5 \times 10^{10}\ L_\odot$	$2 \times 10^{10}\ L_\odot$
Size scales	0.5 kpc to 30 kpc	10 kpc
Ages	10^9 to 1.3×10^{10} years	12 Gyr

Table 5.1: Characteristic properties of galaxies.

Table 5.1 shows some of the characteristic scales that describe

galactic systems. Of note, there is a wide range in masses (4 orders of magnitude) and a similarly wide range in dark matter halo masses. The baryonic matter is a few percent of the total mass of the galaxy, but the ratio between the baryonic matter and the dark matter is a function of galaxy mass. Finally, galaxies are old with ages comparable to the age of the Universe (14 Gyr). This means that galaxy formation had to occur relatively early on in the history of the Universe.

Key Points

- The main parts of a galaxy like our own are the disk, bulge and halo. The centre of a galaxy contains one (or sometimes more) supermassive black holes.
- Most of the mass in a galaxy is in dark matter, followed by stars, and then the gas.
- The stellar populations in the bulge and halo are older than the stars in the disk with signs that the halo formed first and ceased forming stars relatively long ago. The bulge is not actively star forming and then the disk has a significant amount of gas and is actively forming new stars.

5.2 *Defining the Galaxy Population*

In the past 20 years, the population of galaxies has become progressively better defined and the Sloan Digital Sky Survey (SDSS) has proven to be the primary agent driving our refined understanding. Because the SDSS played such a central role, we can spend some time understanding how it worked operationally. The SDSS is actually an umbrella of several different surveys, but the major survey we discuss here is retroactively called the Legacy survey. It consisted of two main parts: an imaging survey and a spectroscopic survey. The sky coverage of the imaging survey is shown on the right.

https://sdss.org

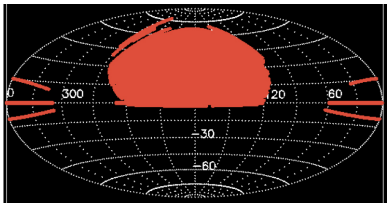


Figure 5.3: Coverage of the Legacy component of SDSS. It's big.

The figure shows the coverage in celestial coordinates. It covered an unprecedentedly large portion of the sky. Furthermore, the imaging survey consisted of a drift scan where the SDSS imaging telescope was pointed at part of the sky and the Earth's rotation carried the sky over different parts of the camera. Cleverly, the data from the camera was digitized ("read out" in the astronomical parlance) simultaneously with the sky objects moving across the camera. This allows for excellent image stability since the telescope was not moving around and tracking on different parts of the sky. Figure 1.2 showed the SDSS camera.

Even better, the camera was divided into several different detectors each column of filters had a filter in front of it. Figure 1.4 shows the SDSS filter set as *ugriz* bands (we will often drop the $'$ that appears on the filters since this just distinguishes between slightly different versions of the same filter). The filter designations roughly correspond to their transmitted colours: u is ultraviolet, g is green, r is red, i is infrared and z is zeven-more-infrared (or something). This configuration allowed for a huge part of the sky to be imaged stably, and the imaging was done in a suite of filters across the optical band. As such, the SDSS was a wonderful survey of the stellar properties of galaxies and was in the vanguard of the era of survey astronomy.

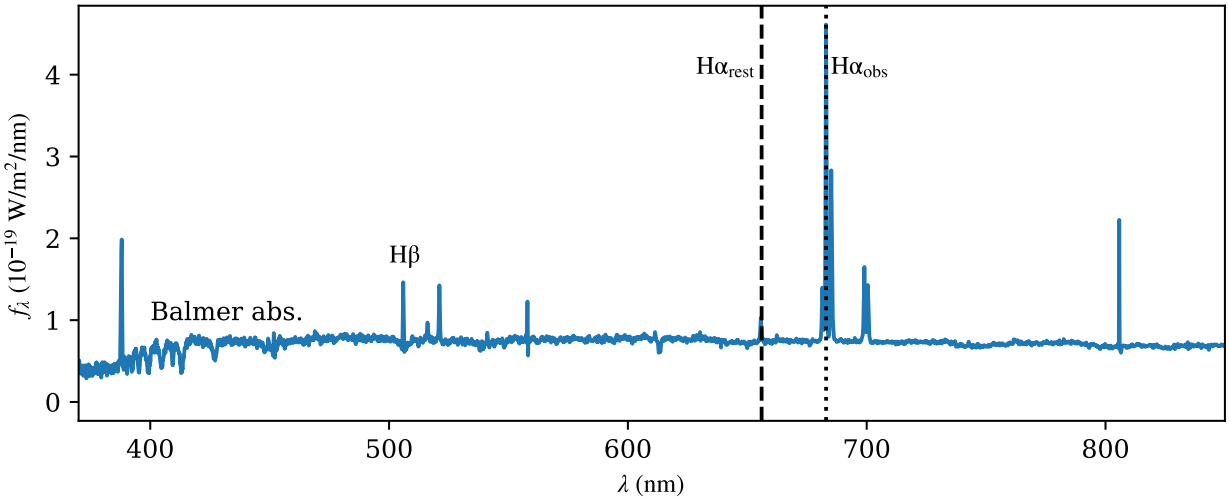


Figure 5.4: Spectrum of a star-forming galaxy taken from the SDSS MANGA survey.

The SDSS also included a spectroscopic followup campaign, which selected targets from the imaging survey and observed them with a fibre-fed spectrograph, allowing many objects in the same part of the sky to be observed simultaneously. A sample spectrum is shown in Figure 5.4. This spectrum combines light from stars with light from the interstellar medium. In general, the absorption lines seen in the spectrum, most prominently near $\lambda = 400$ nm are the Balmer absorption lines in A and B type stars in the galaxy. The spectrum shows several emission lines including the Hα and Hβ lines. The Hα line is in the middle between two lines of [N II] in the spectrum. This emission is from H II regions in the galaxy. The Hβ line shows a combination of absorption from stellar photospheres and emission from photoionized gas.

5.3 *Cosmological Redshift*

The other notable feature in Figure 5.4 is that the wavelengths of the spectral lines are shifted from their expected rest wavelengths. For example the Hα line is found at $\lambda = 683$ nm vs the $\lambda_{\rm rest} = 656$ nm at which it is expected. We can use the observed wavelengths and the known rest wavelengths to determine the line-of-sight velocities of the targets. This determines the spectroscopic redshift z of an object:

$$\frac{\lambda_{\rm obs}}{\lambda_{\rm rest}} \equiv 1 + z \tag{5.2}$$

For $z \ll 1$, $v_r = cz$. Hence, for the spectrum shown above

$$\frac{\lambda_{\rm obs}}{\lambda_{\rm rest}} = \frac{683\ {\rm nm}}{656\ {\rm nm}} = 1.044 = 1 + z$$

so that $z = 0.044$ or a recession velocity of $cz = 13,000$ km s^{-1}.

For most of the galaxies in the SDSS, this recession velocity is dominated by the Hubble flow from which we can estimate a distance using the *Hubble law*:

$$v = cz = H_0 d \tag{5.3}$$

where H_0 is the Hubble constant:

$$H_0 = 70.4 \pm 1.4\ {\rm km/s/Mpc} \tag{5.4}$$

The Hubble Law arises because of the isotropic expansion of the Universe so that all galaxies are moving away from all other galaxies and the rate of expansion is proportional to the distance between the galaxies. This version of the Hubble law is a small order approximation to the full expression. As $z \to 1$, relativistic effects become important and we need to a full cosmological model to be able to relate the cosmological redshift to distance. In applying Equation 5.3, remember that it is only truly valid for $z \ll 1$. The cosmological redshift is formally tracking the scale factor of the Universe. If (and only if) we ignore the gravitational interaction between two galaxies, we can relate how far apart these galaxies were in the past relative to how far they are apart now to the redshift:

$$\frac{d_{\rm past}}{d_{\rm now}} = \frac{1}{1+z} \tag{5.5}$$

Returning to the small-order expansion to the Hubble law, we can use this to calculate distances to objects. For example in using the spectrum in Figure 5.4, we can use $v = 13,000$ km s^{-1} so that

$$d = \frac{cz}{H_0} = \frac{13,000\ {\rm km\ s^{-1}}}{70.4\ {\rm km\ s^{-1}\ Mpc^{-1}}} = 190\ {\rm Mpc}.$$

Combining the Hubble law with the absolute magnitude relationship gives the ability to measure absolute magnitudes for galaxies with spectroscopic redshifts.

$$M_r = r - 5\log_{10}\left(\frac{cz/H_0}{10\text{ pc}}\right). \tag{5.6}$$

One word of caution: the cosmological redshift is not exactly the same as a Doppler shift, though we frequently treat them as such. The redshifting in the cosmological case occurs during the wave propagation. The reason we must be cautious is that galaxies are "moving" due to cosmological expansion *and* they are moving due to other physics like gravitational interaction with other galaxies. These latter motions also create redshifts (and blueshifts). In the parlance of extragalactic astrophysics, these are often called the *peculiar motions* simply because they deviate from the Hubble law.

Redshift and Lookback Time – One of the truisms of astronomy is that, because of the finite speed of light, when we are looking at distant objects, we are seeing them as they were in the past. In astrophysics, we use the cosmological redshift as a proxy for how far we are looking back in time. This is particularly important for the study of galaxy evolution because the galaxy population has been constantly evolving over the course of the Universe.

We can relate the redshift to how far into the past we are looking when we see an object at a given redshift. This time in the past is called the *lookback time*. Figure 5.5 shows the relationship between the cosmological redshift and lookback time. This relationship is set by how the mass and energy in the Universe regulate expansion since the Big Bang under the theory of General Relativity. Again, you can find out where this relationship comes from in ASTRO 430.

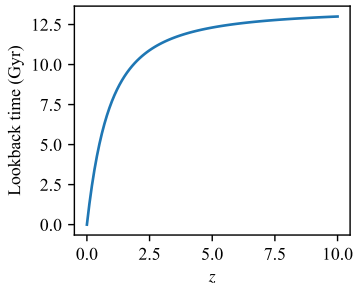


Figure 5.5: Lookback time vs. redshift for a ΛCDM cosmology.

We refer to the local Universe as the $z = 0$ Universe, meaning that all the systems in it are sufficiently close that their cosmological redshifts are small and we are seeing contemporaneous galaxies as our own. However, as the redshift grows larger, we are seeing the Universe in the past when galaxies were in a different state of evolution. At $z = 1$, we know that the separation between galaxies in the Universe is half of what it is today (Equation 5.5) and that we are seeing the galaxies as they were 7.7 Gyr ago, when the Universe was less than half its current age. We are routinely imaging galaxies at $z = 6$ – commonly called "high redshift" or "high-z" – when the Universe was 7% of its current age and galaxies are just forming. Hence, cosmological redshift and lookback time is an amazing asset in understanding galaxy evolution because we can observe the Universe as it was at different ages and infer how the population of the galaxies changes.

Key Points

- Galaxy surveys like SDSS helped define the large-scale galaxy population. These are a combination of spectroscopic and photometric surveys in a series of optical and IR bands.
- The primary tool for measuring distances on the scales of galaxies is the Hubble flow which arises from cosmological expansion. Locally, this allows us to use a linear scaling between the redshift (Equation 5.2) and distance through the Hubble constant (Equation 5.3).
- The redshift is related to the lookback time through the history of the Universe's expansion (Equation 5.5).

5.4 *The Galaxy Colour-Magnitude Diagram*

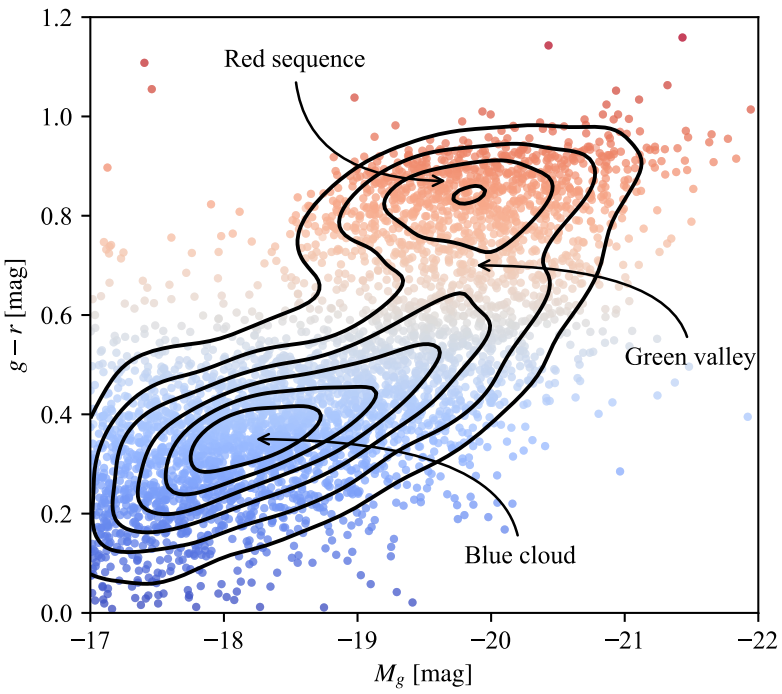


Figure 5.6: The colour-magnitude diagram for 8200 galaxies drawn from the SDSS toward $\alpha = 180°, \delta = +20°$. The contours show the density of the data.

The colour-magnitude diagram (CMD) is an essential tool for studying stars. In Chapter 2 and 3 we saw how this diagram illustrated the relationship between intrinsic properties of stars. By analogy with the stellar population, we can construct a CMD for galaxies.

It's a sensible thing to do: the optical light of galaxies comes from stars so the galaxy CMD encodes information about the ensemble of stars in the system. Figure 5.6 shows the CMD for galaxies in the SDSS. Note that, because of astronomer's general hatred for consistency and sensible conventions, the diagram axes are transposed from the standard stellar CMD. Relatively red galaxies are found at the top and relatively blue galaxies are at the bottom. Low luminosity galaxies are found on the left and high luminosity galaxies are seen on the right.

The galaxy distribution is bimodal with two distinct peaks: one at lower luminosities and bluer colours and the other at higher luminosities and red colours. The red peak appears to follow a structured relationship for which it is sometimes dubbed the *red sequence* owing to the apparent relationship between luminosity and colour. The blue peak is more broadly distributed and is called the *blue cloud*. Galaxies in between the two peaks are said to occupy the *green valley* as annotated on the right.

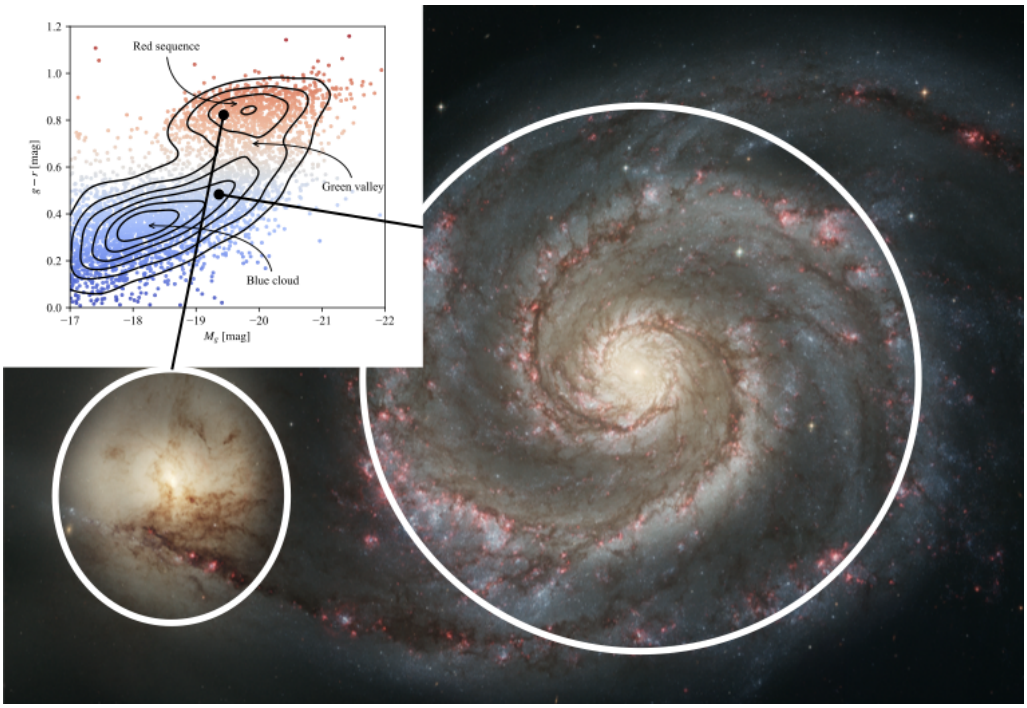


Figure 5.7: The interacting pair of galaxies NGC 5194 and NGC 5195 also known as the Whirlpool galaxy. Optical image credit: NASA and European Space Agency.

To get a sense of the meanings of these different distributions, we can consider two galaxies: NGC 5194 and NGC 5195, the famous Whirlpool interacting galaxy shown in Figure 5.7. The corresponding positions of each galaxy is indicated in the inset galaxy CMD. The main blue galaxy in NGC 5194 and is located in the top end of the blue cloud. It has extremely active ongoing star formation as in-

dicated by its blue colour, well-defined dust lanes, and clear spiral structure. In contrast NGC 5915 is the less luminous galaxy and is located in the red sequence, but at the lower end of it. It shows little signs of dust, though there is a feature toward the centre likely associated with the merger. There do not appear to be any dynamical structures in the galaxy with a relatively featureless ellipsoidal shape. These illustrate how different galaxies appear in the CMD. A critical point of departure between stars and galaxies (in addition to the transposed axes) is that a galaxy's location in the CMD does not establish all of the conditions for a galaxy. Two galaxies can occupy the same place in the diagram yet have very different appearances or *morphologies*.

The primary determinant of a galaxy's colour is the star formation history of the galaxy. Broadly speaking, if a galaxy has had recent star formation (within the past 10^8 years) its overall colours will be relatively blue and it will not appear in the red sequence. Otherwise the galaxy will appear in red, which can be regarded as the default colour of galaxies. Since star formation is associated with the presence of the cool / cold ISM and such ISM appears preferentially in disk galaxies, we tend to see that the blue cloud galaxies have a disk-like component. Of note, galaxies undergoing star formation tend to deplete their gas reservoirs on timescales of 2 Gyr implying that sustained star formation at the current rate requires replenishment of gas.

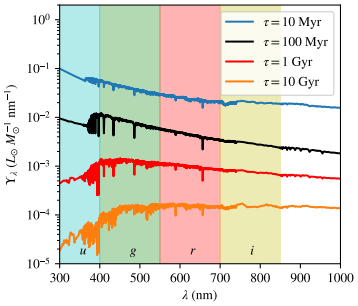


Figure 3.10 showed why the colours of galaxies are set by the star formation rate, which we reproduce in the margin here. That figure showed the emergent flux distribution for a population of stars at times $t = 10^7, 10^8, 10^9, 10^{10}$ years after formation in the SDSS bands. The SEDs of the population change through the course of stellar evolution with general trend that the light becomes fainter and redder over time. It is important to note that most of mass in stars is found in the lowest mass and longest lived stars in the system, meaning that the mass of the system that is contained in stars doesn't change significantly over the time that the light changes. The mass-to-light ratio thus gets significantly larger with an aging population. Another notable feature of young stellar populations is that they put out a large amount of radiation in ultraviolet wavelengths. The Figure also shows that once a stellar population ages past about 10^8 years, it is red.

Figure 5.8 to 5.10 show two representative galaxies from each of the red sequence, green valley, and blue cloud respectively. The key point in these figures is that each pair of galaxies has been chosen to have contrasting morphologies for comparable colours and luminosities. For the pair of red sequence galaxies, one shows a significantly

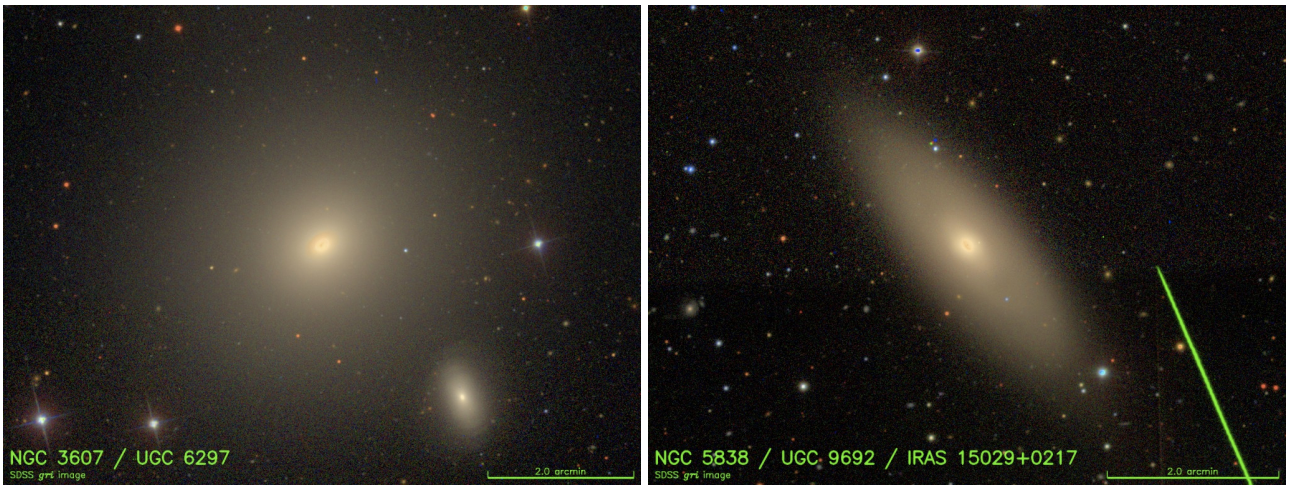


Figure 5.8: Example red sequence galaxies drawn from the NGC imaging page of David Hogg. The galaxies show a variety of morphologies but all show the characteristic red colour of old stellar populations. The linear feature in the right panel probably an airplane or a satellite. Get used to seeing more of these with satellite constellations like

higher aspect ratio (long axis compared to short axis) compared to the other. However both show the characteristic colours of old stellar populations and are free from obvious dynamical features (spiral arms) or dust lanes.

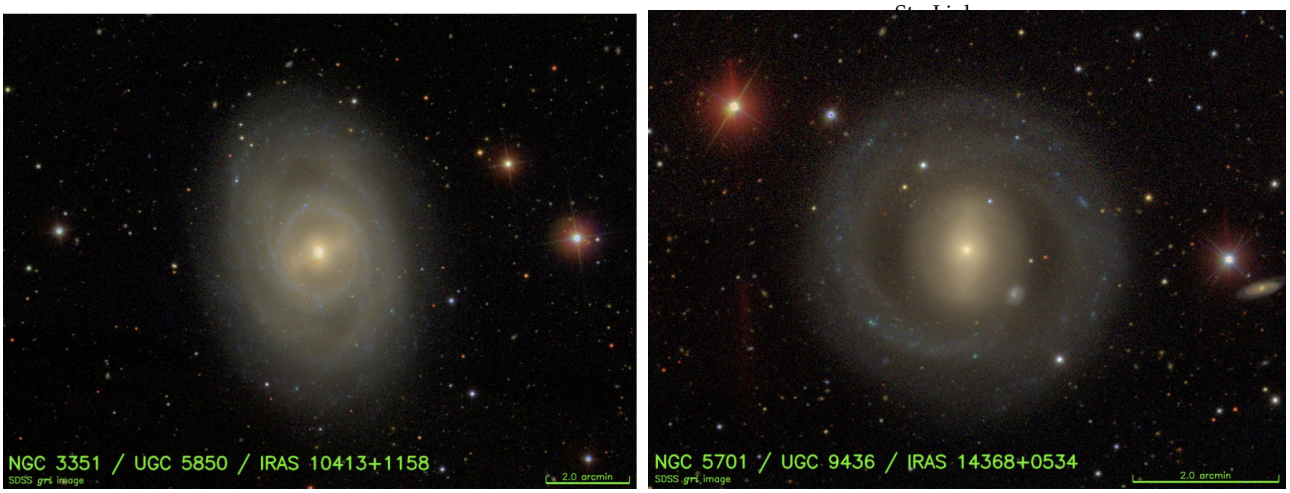


Figure 5.9: Example green valley galaxies drawn from the NGC imaging page of David Hogg.

The green valley galaxies have different dynamical features. They show a significant central bulge that appears to have the colours of an old stellar population. The bulges appear to be surrounded by disks that have colours associated with recent star formation. Both galaxies are barred but they appear to have different spiral arm patterns and the right galaxy appears to have ring-like structure.

Finally, the two blue cloud galaxies show markedly different morphologies. The first galaxy appears to have no stellar structure and

Figure 5.10: Example blue cloud galaxies drawn from the NGC imaging page of David Hogg.

would be termed *Irregular* whereas the second galaxy has disk-like structure but no bulge. The disk appears to be bluer than the green valley galaxies, which is an effect of there not being as much of an underlying old stellar population in the blue cloud galaxies.

Comparing the green valley galaxies to the blue cloud galaxies shows that both tend to have disk-like structures and the green valley is just a redder version of the blue cloud. Overall though, we are struck with the variety of morphologies in these broad categories of galaxies so it appears that morphology is another key observable of galaxies.

Key Points

- Galaxy redshift surveys show that galaxies follow a colour-magnitude diagram (Figure 5.6) showing a luminous red sequence and a lower luminosity blue cloud with a underpopulated region in between called the green valley. Galaxies on the red sequence are not actively star forming, leading to their redder colours. These galaxies formed most of their stars early in the Universe. Galaxies on the blue cloud have significant amounts of cold ISM and are actively star forming, hence their blue colours.
- Galaxy colour are primarily an indication of their recent star formation. As a stellar population ages, its mass-to-light ratio changes (Figure **??**) so that the population becomes redder and overall less luminous.

5.5 *The Luminosity Function*

Figure 5.6 shows the distribution of galaxies in the colour-absolute magnitude plane. If we ignore the colour of the galaxies and simply ask how many galaxies of a given luminosity (absolute magnitude) there are in the figure, we can measure the *luminosity function*. The luminosity function is analogous to the initial mass function for stars in that it is a probability density function that measures the number density of galaxies found in intervals of luminosity:

$$\frac{1}{V}\frac{dN(L)}{dL} = \frac{dn(L)}{dL} \tag{5.7}$$

where V is the volume covered by the study, N is the number of galaxies between L and $L + dL$, and $n(L) \equiv N(L)/V$ is the number density of galaxies with a luminosity between L and $L + dL$.

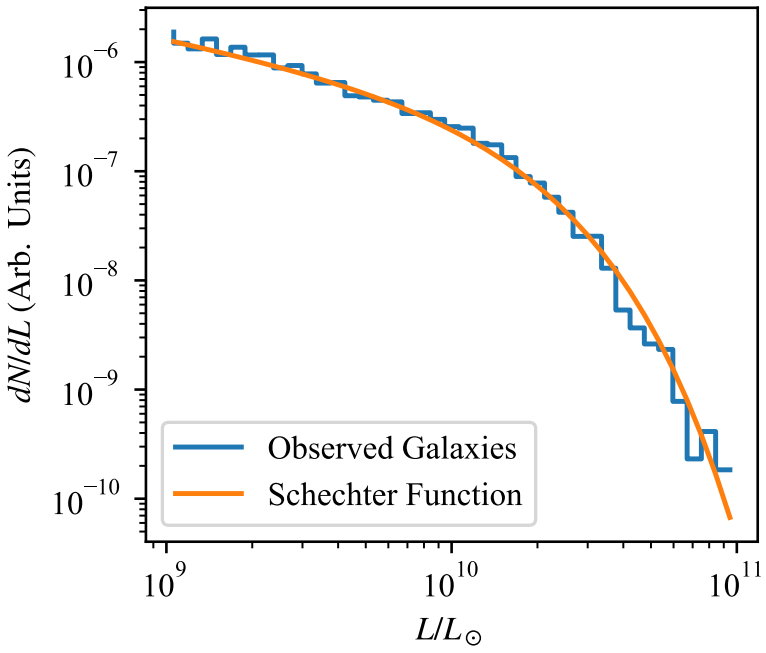


Figure 5.11: The luminosity function of SDSS galaxies.

Figure 5.11 is proportional to the luminosity function for the stars in Figure 5.6. The "proportional" arises because I couldn't determine the survey volume given how I sampled the SDSS data, but it is a constant factor off from the the luminosity function, which is usually measured in units of Galaxies/Mpc^3/$L_\odot$. This figure also includes a fit to the data using a functional form called the Schechter function, which was developed as an empirical fit to the galaxy luminosity

function by an astronomer named Schechter. The form of this function is

$$\frac{dn}{dL} = \frac{\phi}{L_\star}\left(\frac{L}{L_\star}\right)^{\alpha} \exp\left(-\frac{L}{L_\star}\right) \tag{5.8}$$

where ϕ and $L_\star$ are constants and α is an index. Typically, $-2 < \alpha < -1$ and $L_\star \approx 2 \times 10^{10}\ L_\odot$ (Figure 5.11 shows the fit for $\alpha = -1.5$ and $L_\star = 2 \times 10^{10}\ L_\odot$). Both the index and $L_\star$ vary with galaxy population. For example, galaxies in the early Universe (high-z) systems show lower $L_\star$. One of the major goals of galaxy evolution theory is to explain the functional form of this luminosity function and why it varies.

Key Points

- The galaxy luminosity function is the analogue of the initial mass function for stars. However, instead of measuring the number of stars at different stellar masses, it represents the number density of galaxies at different luminosities.
- The luminosity function typically follows a Schechter function 5.8 form but, unlike stars, the parameters of the function change over time in the Universe and in different environments.

5.6 *Morphological Classification*

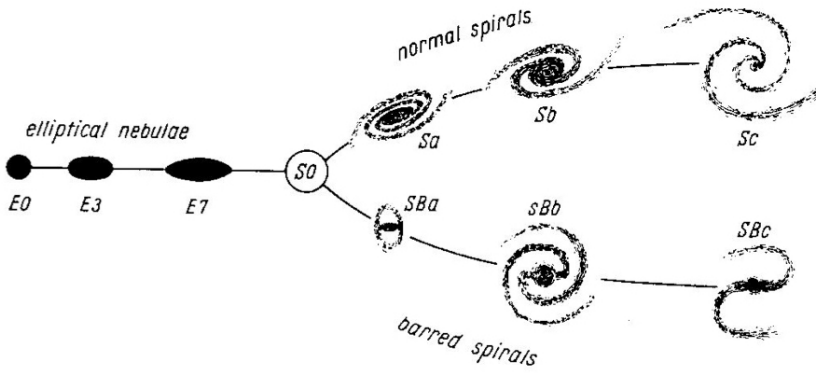


Figure 5.12: Morphological classification of galaxies as shown in Hubble's *Realm of the Nebulae* (1936).

The galaxy CMD is actually a rather recent innovation in the scientific description of galaxies. A far older approach is to leverage the differences in morphologies to describe different galaxies. Figure 5.12 shows Hubble's classification on galaxies based on their shape in the classic *tuning fork* diagram. Galaxies are broadly classified into

elliptical or spiral galaxies with So (lenticular) galaxies being the intermediate case. The spiral galaxies are further divided into barred and unbarred types based on the presence of a (wait for it) bar. The tuning fork structure is an unfortunate holdover from the original conjecture that this was an evolutionary sequence. While we do believe that major mergers can indeed transform one galaxy type into another, the pathway that Hubble envisaged is not longer well described. However, the linguistic legacy lives on with elliptical galaxies being called *early types* and spiral galaxies being called *late types* and the farther along the tuning fork and the relative ordering of early to late being left-to-right in the figure shown above.

Astronomers still use the Hubble classification scheme for naming galaxy and a good deal of information is encoded in these short designations. The Hubble scheme has been refined and extended to a fairly detailed system laid out by de Vacauloeurs and used in the Third Reference Catalog of Bright Galaxies [RC3; de Vaucouleurs et al., 1991].

The designation for the elliptical galaxies is determined by the aspect ratio of the surface brightness distribution (Figure 5.13):

$$\varepsilon = 1 - \frac{b}{a} \tag{5.9}$$

where b and a are the observed major and minor axes of the galaxy respectively. The number after the E is just $10 \times \varepsilon$.

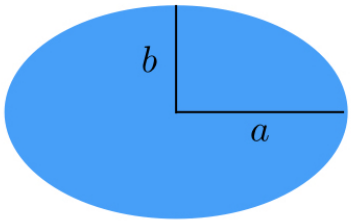


Figure 5.13: The geometry used to determine the ellipticity of a survey brightness profile.

Key Points

- Galaxies are frequently classified by their shapes, but this is becoming less important over time. This is mostly important for understanding the literature on galaxies.
- The main divisions of galaxies in morphological classification are ellipticals vs spirals with lenticulars as intermediate between these two cases. Spiral galaxies are further subdivided based on whether they have a bar or not.

5.7 *Our Galaxy*

Because it is *our* Galaxy, we focus a lot of attention on the Milky Way which has advantages and disadvantages for studying the the physics of galaxies as a whole. The major disadvantage is that we are stuck inside the Galaxy and therefore cannot gain a broad perspective on the entire system. The top-down image of the Galaxy that we showcase in Figure 1.9 is an artist's rendition that has been pieced together through decades of observations of the Galaxy. The reality

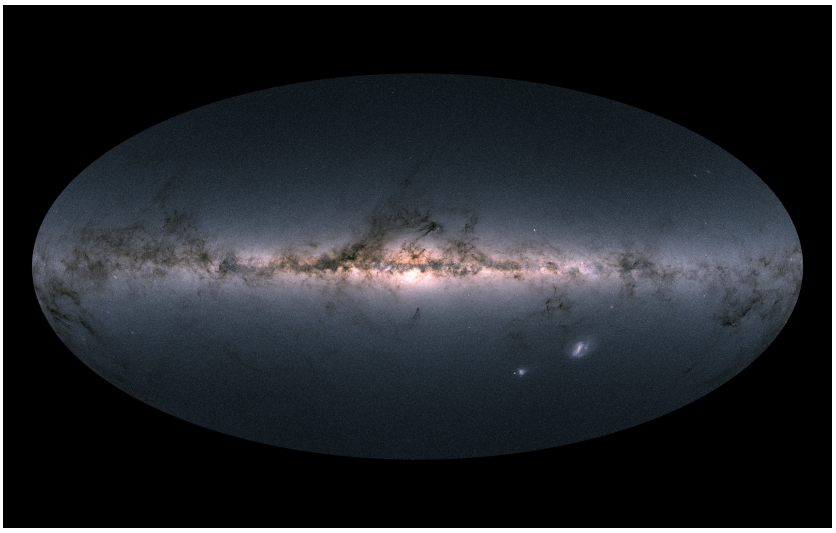


Figure 5.14: The Milky Way as seen by the Gaia mission. Image credit: Gaia Data Processing and Analysis Consortium (DPAC); A. Moitinho / A. F. Silva / M. Barros / C. Barata, University of Lisbon, Portugal; H. Savietto, Fork Research, Portugal.

of observations look more like Figure 5.14. We can identify the plane of the Milky Way and the increased brightness toward the centre of the Galaxy, which is in the middle of the image. However much of the perspective on the Galaxy is obscured by dust which blocks out the optical light.

Observations of at wavelengths of $\lambda \sim 2\ \mu$m are good probes of the stellar content of galaxies. The wavelengths are long enough that the extinction due to dust is minimized. In the K-band at 2 μm, $A_K = A_V/10$. Looking through the disk of the galaxy, where $A_V = 30$ along some lines of sight, this extinction in the near infrared can still be significant. However, we can still get a minimally biased view of the stellar content of the Milky Way. Figure 5.15 shows an image of the Milky Way disk in the near infrared wavelengths which shows much more clearly the structure of our Galaxy as a disk which a bright nuclear structure toward the centre. Careful examination of the central bulge feature shows suggests that it is not perfectly spheroidal but has a slight "peanut" shape, which was one of the first indications that our own Galaxy is a barred galaxy.

The Galaxy's disk-like geometry implies that the natural coordinate system for mapping out its physical structure is a cylindrical polar coordinate system with an origin centred on the middle of the Galaxy. We refer to this coordinate system as the *galactocentric* coordinate system. This is different from the Cartesian coordinate system we developed in Chapter 2 (Equations 1.29 to 1.31), which was centred on the Sun using the Galactic coordinate system on the sky.

Figure 5.16 shows the Galactocentric coordinate system. In this coordinate system, the position of an object is defined in terms of

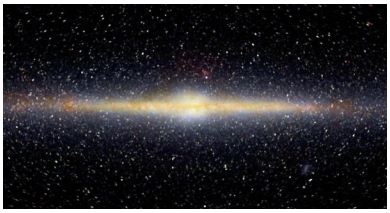


Figure 5.15: The Diffuse Infrared Background Explorer image of the Milky Way in the 1.5 to 4 μm bands. Image credit: E. L. Wright (UCLA), The COBE Project, DIRBE, NASA.

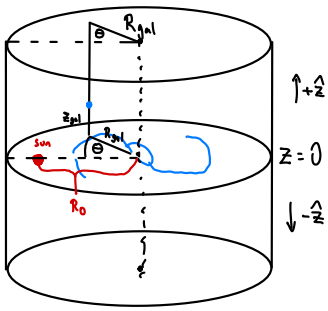


Figure 5.16: Sketch of the Galactocentric coordinates.

coordinates $(R_{\rm gal}, \Theta, Z_{\rm gal})$. We prefer this geometry since, in a disk like system, the variations in physical quantities are relatively constant in terms of the angular coordinate Θ and the system's physical properties are described by its variation in $R_{\rm gal}$ and $Z_{\rm gal}$.

We can further relate the galactocentric coordinate system to the Galactic coordinates using triangle geometry as is shown in Figure 5.17. The key points in this diagram are the Sun's position and the Galactic centre (GC). We have measured that the Sun is located a distance $R_0 = 8.178 \pm 0.035$ kpc [Gravity Collaboration et al., 2019] from the Galactic centre. The circle with radius R_0 centred on the GC is called the *solar circle*. If we know the distance to an object elsewhere in the Galaxy (d), we can use the law of cosines to measure how far it is from the Galactic centre:

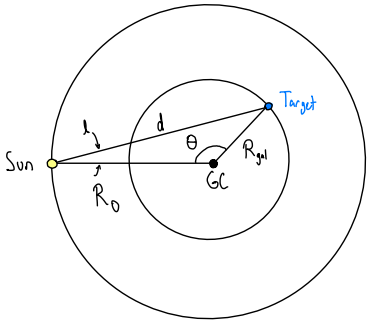


Figure 5.17: Top-down view of Galactocentric coordinates.

$$R_{\rm gal}^2 = R_0^2 + d^2 - 2R_0 d \cos \ell \tag{5.10}$$

where ℓ is the Galactic longitude as discussed in Chapter 2. If necessary, we can also find the angular coordinate Θ from the the law of sines:

$$\frac{d}{\sin \Theta} = \frac{R_{\rm gal}}{\sin \ell} \tag{5.11}$$

Sometimes different parts of these triangles are unknown and the solutions must seek the remaining variables but the fundamental geometry remains the same. For example, we can use Galactic rotation and the Doppler shift to measure the motions of atomic hydrogen gas. By assuming the gas is on circular rotation, it is possible to use the Doppler shift to measure $R_{\rm gal}$ for a gas cloud. In this case, we can then solve Equation 5.10 to find the distance in a method called measuring the *kinematic distance*.

5.7.1 Mass Density Profiles

Thanks to our position inside the Galaxy, we can make a measurement of the density distributions of matter. We focus here on the Milky Way but we also assume that other disk galaxies follow a similar distribution of matter. This general functional approach is appropriate with modification for different stellar populations and the various phases of the ISM. The general form is:

$$\rho(R_{\rm gal}, \Theta, Z_{\rm gal}) = \rho_0 \exp\left(-\frac{R_{\rm gal}}{R_d}\right) f\left(\frac{Z_{\rm gal}}{H}\right). \tag{5.12}$$

Here, ρ is the mass density of matter in, e.g., $M_\odot$ pc^{-3} or equivalent units. The characteristic length for the radial distribution is called the *scale length* and is denoted R_d. The standard function has no dependence on the angular coordinate Θ, i.e., the disks are assumed

to be azimuthally symmetric. This is clearly not entirely true because of spiral arm and bar-like features. Despite their stark appearance, their influence on the mass distribution is relatively small: spiral arms are only a $\sim$ 20% perturbation on the mass though they are much sharper contrast in the light. The vertical distribution function – $f(Z_{\rm gal}/H)$ – has several commonly adopted functional forms that depend on the matter being studied. The dimension H is called the *scale height*. Taking $\zeta \equiv Z_{\rm gal}/H$, some common forms of $f(\zeta)$ are

$$f(\zeta) = \exp(-\zeta^2/2) \quad \text{Gaussian} \tag{5.13}$$

$$= \exp\left(-|\zeta|\right) \quad \text{Exponential} \tag{5.14}$$

$$= \text{sech}^2(\zeta/2) = \frac{4}{e^{-\zeta} + e^{\zeta} + 2} \quad \text{“sech-squared”} \tag{5.15}$$

The sech^2 profile (Equation 5.15) is a physically motivated form for a self-gravitating layer of material and is derived in the Appendix (9.4). It is also called the “sech-squared” profile. This is only included for completeness and we won't dive into this physics. But the physics does reward us with a little more insight that says we can derive a scale height for the sech^2 layer

$$H = \sigma_Z / \sqrt{8\pi G \rho(Z_{\rm gal} = 0)} \tag{5.16}$$

where σ_Z is the vertical velocity dispersion and $\rho(Z_{\rm gal} = 0)$ is the mass density at the midplane at that location. A graph of the sech^2 layer density profile and other functional forms is shown in Figure 5.18. All these functional forms have a value of $f(\zeta) = 1$ at $\zeta = 0$.

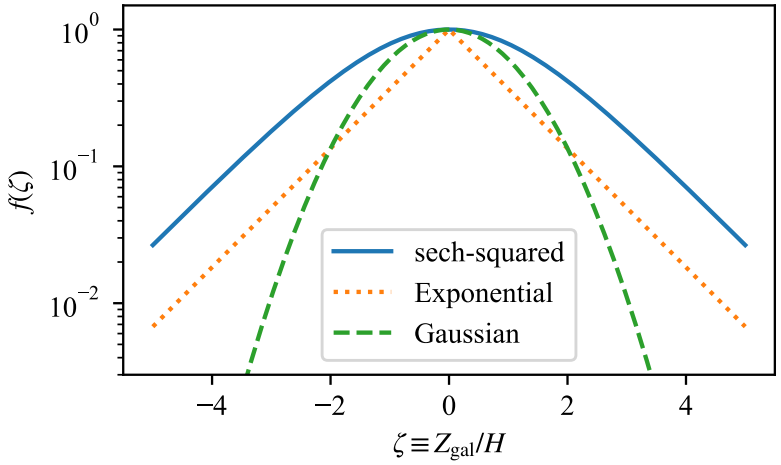


Figure 5.18: The vertical density profiles for material distributions in a disk.

For external galaxies, we do not measure the mass density directly, but instead we see all the mass along a given line of sight. If we look

directly through a face on galaxy, we can measure the *surface density* profile, Σ, where

$$\Sigma(R_{\rm gal}, \Theta) = \int_{-\infty}^{\infty} \rho(R, \Theta, Z_{\rm gal}) dZ_{\rm gal} \quad (5.17)$$

$$= \int_{-\infty}^{\infty} \rho_0 \exp\left(-\frac{R_{\rm gal}}{R_d}\right) f\left(\frac{Z_{\rm gal}}{H}\right) dZ_{\rm gal} \quad (5.18)$$

For a sech^2 profile, this reduces to

$$\Sigma(R_{\rm gal}, \Theta) = 4H\rho_0 \exp\left(-\frac{R_{\rm gal}}{R_d}\right) \quad (5.19)$$

where we define $\Sigma_0 = 4H\rho_0$. The different functional forms have different coefficients in this relationship. For a Gaussian, $\Sigma_0 = \sqrt{2\pi}H\rho_0$ and for an exponential $\Sigma_0 = 2H\rho_0$. Practically, for external galaxies, the surface density is what we measure and we assume a vertical distribution of matter and use that to infer the midplane density.

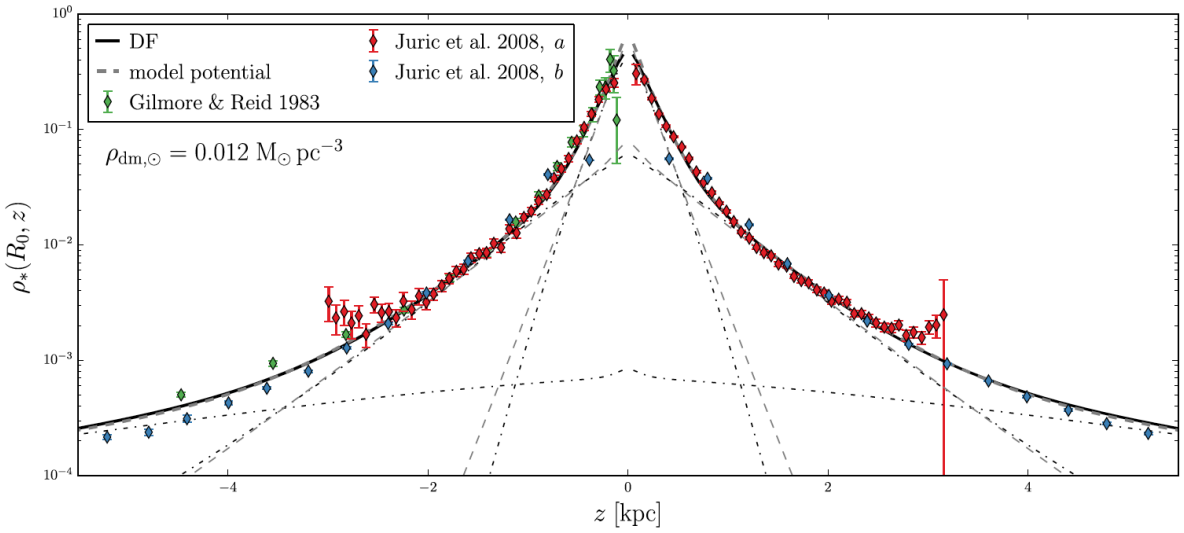


Figure 5.19: Vertical distributions of stars in the solar neighbourhood. From Bland-Hawthorn and Gerhard [2016].

Scale Heights – If we consider stars in the solar neighbourhood, we can look at their different vertical distributions. Figure 5.19 the density of stars as a function of $Z_{\rm gal}$ for $R_{\rm gal} = R_0$. The distribution function here does not actually follow any single one of the model distributions, i.e., it is not a sech^2, Gaussian or exponential profile. However, it appears to be a composite of three different profiles, each of which is well modelled by a sech^2 or exponential profile with different scale heights. These three components are called the *thin*

disk, the *thick disk*, and the *halo* in order of increasing scale height and decreasing mass density in the midplane. At the solar circle, $H_{\text{thin}} = 300 \pm 50$ pc and $H_{\text{thick}} = 900 \pm 180$ pc. The model for the halo is not disk like but is spheroidal so it does not have a scale height.

Table 5.2 gives the midplane mass densities of various components of matter at the solar circle based on dynamical estimates for the total mass and counting up the baryonic matter (stars and gas). These data are adopted from Bland-Hawthorn and Gerhard [2016].

Component	Density ($M_\odot$ pc^{-3})
Thin Disk	0.041
Thick Disk	0.002
Gas	0.041
Dark Matter	0.013
Halo	0.00004
Total	0.097

Table 5.2: Mass densities at $R_{\text{gal}} = R_0$ and $Z_{\text{gal}} = 0$.

The different disk components also have different dynamical and metallicity properties, which we will explore in later chapters. Briefly, the thin disk is the youngest and dynamically cold of the stellar components. The halo is the oldest population and is dynamically hot and the thick disk is intermediate between these two on nearly all axes.

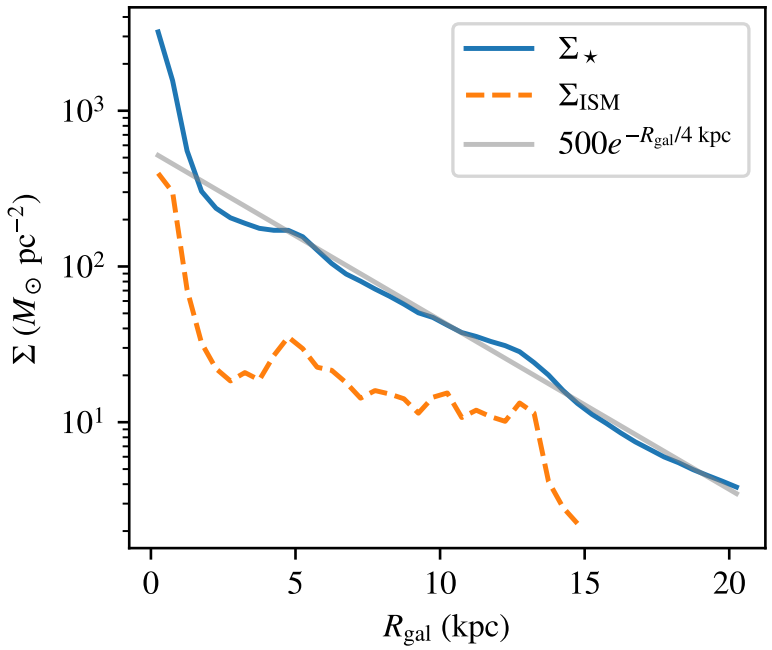


Figure 5.20: Radial profiles of matter for the galaxy NGC 4321

Radial Profiles – The radial profiles of galaxies are typically exponential in their total gas and stellar content. Figure 5.20 shows the average surface density of matter in the nearby galaxy NGC 4321. This is plotted on a logarithmic y-axis so that exponential profiles appear as straight lines. For $R_{\text{gal}} \gtrsim 2$ kpc, the mass surface density profile for the stars appears to follow an exponential distribution with a disk scale length of $R_d \approx 4$ kpc illustrated in by the grey line

in the figure with $\Sigma_0 = 500\ M_\odot\ \mathrm{pc}^{-2}$. Inside $R_{\mathrm{gal}} < 2$ kpc, the surface density profile rises sharply, associated with the stellar bulge in this galaxy.

The total gas surface density in the ISM is also plotted in Figure 5.20. We can note that $\Sigma_{\mathrm{ISM}} \sim \Sigma_\star/10$ and there is a hint of an exponential profile in the gas as well. This is not nearly as smooth as the stellar profile. Analysis of the Milky Way yields that the stellar scale is $R_d = 2.6 \pm 0.1$ kpc [Licquia and Newman, 2016] for the thin disk and $R_d = 2.0 \pm 0.2$ kpc for the thick disk [Bland-Hawthorn and Gerhard, 2016].

BULGES AND ELLIPTICALS – The surface density profile for bulges and elliptical galaxies follows a slightly different model. This model is empirical. While disks have a vertical density distribution that can be approximated from physics and their exponential distribution in the radial direction fits well with galaxy formation theories, the more spheroidal systems are described by a surface density profile of

$$\Sigma = \Sigma_0 \exp\left\{-b_n\left[\left(\frac{R}{R_e}\right)^{1/n} - 1\right]\right\} \tag{5.20}$$

This profile is called the Sérsic profile and the index n is called the Sérsic index. Typically, n ranges from 1 to 4, which corresponds to the transtion between an exponential disk ($n = 1$) and the standard profile for an elliptical galaxy which has $n = 4$. In the latter case, this profile is sometimes called the *de Vaucouleurs profile*.

The increase in brightness at the centre of NGC 4321 that is visible in Figure 5.20 is the bulge of the galaxy. In general, a disk galaxy's brightness profile is just the sum of a Sérsic profile in the centre and an exponential disk in the outside. Elliptical systems have no associated disk and irregular galaxies are... well... irregular. However, a galaxy can look like a real train wreck in the blue part of the optical if it is undergoing patchy recent star formation. If the system is observed in the near infrared, it ends up looking a lot smoother, since the near infrared traces the mass of the total stellar population better.

Figure 5.21 shows the surface density profiles for different values of the index n. The constant b_n has been chosen in each case so that all the profiles are normalize to the same value at $R = R_e$, which is called the *effective radius*, which is the radius that contains half of the mass (or light if we are just looking at the light of the galaxy). Galaxies with $n = 4$ are more centrally concentrated than galaxies with $n = 1$ and this index is a useful measure of galaxy structure that is less subjective than Hubble classification. In general, large values of the index $n \sim 4$ correspond to older stellar populations and more massive red sequence galaxies. Galaxies with $n \sim 1$ are usually found

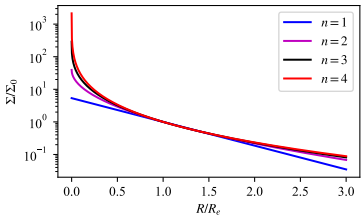


Figure 5.21: Surface density distributions for different Sérsic profiles.

in the blue cloud and the green valley galaxies have intermediate values.

Unfortunately, for elliptical and bulge systems, there isn't a clear relationship between the surface density distribution and the three-dimensional mass density distribution, i.e., there is not a great analogue to equation 5.12.

HALOS – The final part of the galaxy to explore is the halo. The stellar content in the halo of a Galaxy is measured to typically follow that of a spheroid where

$$\rho(R_{\rm gal}, \Theta, Z_{\rm gal}) = \rho_0 \left(\frac{m}{m_0}\right)^n \quad \text{where} \quad m^2 = R_{\rm gal}^2 + \frac{Z_{\rm gal}^2}{q^2} \tag{5.21}$$

where q is a flattening parameter. In the Milky Way, $q \sim 0.6$ [Helmi, 2008]. This spheroidal system consists of stars on randomized orbits and is thought to be closely connected to a galaxy's dark matter halo density.

Key Points

- For disk galaxies, we quantify their structure using a cylindrical-polar coordinate system (Figure 5.16).
- The mass density profile of a disk galaxy follows an exponential profile radially and a compact profile vertically which can be modelled with several different forms (Equations 5.12 and 5.15).
- The vertical stellar structure of the Milky Way shows that the stars are organized into an older thick disk with a larger scale height and a younger thin disk with a smaller scale height. Most of the mass in the solar neighbourhood is in the thin disk and gas.
- Bulges, halos, and elliptical systems all follow different density distributions that we describe empirically (Equations 5.20 and 5.21).

5.8 *The Open Questions*

What comes above represents a huge chunk of information of what we see about galaxies. Embedded in this information is the story of galaxy physics and evolution. These observations raise several questions that we aim to understand through the application of physics. We ultimately aim to build a picture of galaxy evolution that can explain these observations and lay out the pathways that connect the

origin of the Universe in the Big Bag to the population of galaxies we see today. In the next chapters, we aim to understand:

1. Why do galaxies have the shapes that they do?
    - How are stars and gas moving through galaxies?
    - What is different about the history of the bulge and disk of a galaxy?
    - Why are the masses of the black holes in the centres of galaxies correlated with their bulges?
2. What processes shape the light we receive from galaxies?
    - What distinguishes the red sequence from the blue cloud galaxies?
    - Are these galaxy populations connected through evolution?
3. Why do galaxies have the contents that they do?
    - Why are galaxies mostly stars with only a little bit of gas?
    - What physics makes the masses galaxies change over time?
    - What physics changes the metallicity of galaxies and the enrichment of the material in a galaxy with the products of stellar evolution?
    - What makes a galaxy build up mass faster or slower?
4. How do the answers to these questions change as the Universe evolves?

6
Dynamics and Secular Evolution

Given our list of questions from the previous chapter, we first want to ask the hardest question: why do galaxies have the shapes that they do. To answer this, we turn to the study of *galactic dynamics*. Here, the term *dynamics* simply means through forces, typical the gravitational force though particle collisions (i.e., electromagnetism) are important for gas and fluid collisions. Here, we will consider the evolution of systems of stars under their mutual interaction via gravity.

First, we consider the physics governing a pair of stars interacting and (spoilers) we will find that it's not something we have to consider deeply. Instead, we can treat stellar motions in terms of a statistical mechanical framework, which is the heart of galaxy dynamics. Overall, we have to hold two ideas in our mind to get a good sense of galaxies. First, the statistical perspective which describes the evolution of the ensemble, but also we should consider the behaviour of individual stellar orbits which gives some intuition about how a given star has traversed the galaxy. Prior to this discussion, we should consider the typical conditions for stellar systems to contextualize our calculations. Specifically, we consider the number of stars N, their characteristic mass $\langle M \rangle$, their volume density $n_\star$, and their typical (rms) speeds $\sigma_v = \sqrt{\langle v^2 \rangle}$

	Galactic Disk	Globular Cluster	Open Cluster
N	10^{11}	10^6	10^2
$\langle M \rangle (M_\odot)$	0.2	0.1	0.2
$n_\star$ (pc)$^{-3}$	0.1	10^4	10
σ_v (km s^{-1})	30	10	10

Table 6.1: Characteristic scales for the properties of stellar systems.

For clarity, we should be formal about what we mean by the *velocity dispersion* σ_v. This is the standard deviation around the mean velocity and we can consider it projected into coordinate axes x, y, z.

Let's consider the velocity dispersion in the x direction.

$$\sigma_{v,x}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(v_{x,i} - \bar{v}_x)^2 \tag{6.1}$$

where $v_{x,i}$ is velocity vector component in the x direction for the ith star and the sum runs over the N stars in the system. The variable $\bar{v}_x$ is mean velocity in the x direction:

$$\bar{v}_x = \frac{1}{N}\sum_{i=1}^{N} v_{x,i}. \tag{6.2}$$

There are similar equations for the y and z direction. We can then find the total velocity dispersion

$$\sigma_v^2 = \sigma_{v,x}^2 + \sigma_{v,y}^2 + \sigma_{v,z}^2. \tag{6.3}$$

We often describe a velocity dispersion as *isotropic*, meaning it is the same in all directions. In that case $\sigma_{v,x}^2 = \sigma_{v,y}^2 = \sigma_{v,z}^2$ so that $\sigma_v^2 = 3\sigma_{v,x}^2$ or for any coordinate axis.

6.1 Two-body interactions

Two unbound stars that approach each other will interact under (Newtonian) gravitation. Formally, we would consider this as a two particle system interacting on an unbound, hyperbolic orbit. This is a well studied and solved problem but the dynamics of the orbit become important when the scale of the closest approach between the two stars brings the stars sufficiently close that their gravitational potential is comparable to their mutual kinetic energy. For stars of mass m_1, m_2, the system has a total mass of $M = m_1 + m_2$ and a *reduced mass* of

$$\mu = \frac{m_1 m_2}{m_1 + m_2}. \tag{6.4}$$

For these definitions, the gravitational energy is $U_{\rm grav} = GM\mu/r$ and the kinetic energy is $K = \mu v^2/2$. The separation on which these two energies become comparable is

$$r = \frac{2GM}{v^2}. \tag{6.5}$$

Considering a pair of stars in the Galaxy disk with typical properties as summarized in Table 6.1, this spatial scale is 1 AU$\lesssim 10^{-5}$ pc. Let's call this scale r_c for the separation required for a stellar "collision" to occur. These are also called *strong interactions* and perturb stellar trajectories significantly.

How frequently, then, do stars pass within 1 AU of each other? We could probably infer that collisions are probably rare relative to

Estimation pro-tip: One of the fun numerical coincidences in galaxy dynamics is that 1 km s^{-1} $\approx$ 1 pc Myr^{-1} and that you can express the gravitational constant $G \approx 1/222$ pc^3 Myr^{-2} M$_\odot$. This allows for easy calculation of many quantities using $G \approx 1/222$ with units of velocities in terms of km/s, times in Myr, masses in solar masses and distances in pc.

the age of the solar system (4.6 Gyr) owing to the whole "still goin' round the Sun" argument. A star passing within 1 AU of our Sun would probably not be conducive to remaining in a nice circular orbit. To pin some more numbers on this estimate, we use a mean free path argument. If we consider a star moving through a stationary field of other stars at the characteristic relative motion of the stellar population, it will travel along a straight line and have a close encounter with any star that is within a distance of r_c of the trajectory. This interaction volume is thus a cylinder of radius r_c and a length equal to vt_c. The collision time t_c is defined as the time that elapses before this cylinder contains 1 star from the static field with volume density $n_\star$. Considering the typical summed mass of stars to be twice the average mass of a star ($M = 2\langle M \rangle$) and the characteristic relative speed of stars to be given by their random motions $v = \sigma_v$

$$(\pi r_c^2 \sigma_v t_c) n_\star = 1 \tag{6.6}$$

$$t_c = \frac{1}{\pi r_c^2 \sigma_v n_\star} \tag{6.7}$$

$$= \frac{v^3}{4\pi G^2 \langle M \rangle^2 n_\star}. \tag{6.8}$$

Or, scaling this equation to typical values gives

$$t_c = 7.0 \times 10^{15} \text{ yr} \left(\frac{\sigma_v}{30 \text{ km s}^{-1}} \right)^3 \left(\frac{\langle M \rangle}{0.2\ M_\odot} \right)^{-2} \left(\frac{n_\star}{0.1 \text{ pc}^{-3}} \right)^{-1} \tag{6.9}$$

The casual observer will note that t_c is much larger than the Hubble time of $t_\mathrm{H} = 13.6$ Gyr, i.e., which we take as the characteristic age of the Universe.

This argument shows that stars do not have *strong* interactions with other systems over the course of their lifetimes. Instead we consider stars in a statistical fashion governed by the interaction of a single star with a background gravitational potential. The tools of statistical mechanics serve us well here because we can consider the ensemble properties of stars to understand their evolution. Plus, a given system is comprised of many stars ($N_\star \sim 10^{11}$ in our Galaxy), so statistics is a good way to go.

It is worth while comparing stars to other systems undergoing collision such as neutral gas particles, which is the domain of statistical descriptions of gas. Neutral gas particles collide via dipole interaction when the particles approach each other. The range of this induced dipole interaction is short, comparable to the scale of the individual particles $\ell \sim 0.1$ nm. Outside this range, the particles are non-interacting and the collisions within this range are elastic. We can use this information to calculate the collision time for particles in a gas following a similar formalism as above. We use a statistical

description of a gas to relate the characteristic velocity of the gas to the temperature of the gas: $\sigma_v = \sqrt{kT/m}$. Note that we are using a one-dimensional velocity dispersion so this will not have the typical factor of 3.

$$t_c = \frac{1}{\pi r_c^2 \sigma_v n} \tag{6.10}$$

$$= \frac{\sqrt{m}}{\pi r_c^2 \sqrt{kT} n} \tag{6.11}$$

$$= 1.1 \times 10^3 \text{ yr} \left(\frac{m}{m_\text{H}}\right)^{1/2} \left(\frac{T}{100 \text{ K}}\right)^{-1/2} \left(\frac{n}{10^6 \text{ m}^{-3}}\right)^{-1} \tag{6.12}$$

Here, we have non-dimensionalized the equation to conditions typical of neutral gas in our Galaxy. We can compare this to gas on the planet were you are reading this where $T \sim 300$ K, $n \sim 2 \times 10^{25}$ m^{-3}, and $m \sim 29\ m_\text{H}$ where $t_c = 5$ ns. In both cases, the collision times are much shorter than the other relevant timescales in the system, so that the gas behaviours are collisionally dominated and must be treated in that fashion. What makes this relatively easy is that the interactions have a small length scale and elastic collisions. Hence, statistical mechanics.

The intermediate case is for plasmas where the study is complicated by two effects. The ionized nature of the material means that the particles have long range interaction effects from electromagnetic forces, and the electrons have a significantly different mass scales from the ions and so respond to forces significantly differently. Furthermore, collisions are important between particles in many domains. Fortunately, we don't have to consider plasmas much in this class, but respect your local plasma physicists.

Key Points

- A *collision* between a pair of stars occurs when they pass near enough to significant change the direction of their orbits. Collisions act to randomize stellar motions.
- Collisions between stars in most environments are extremely rare.
- In contrast, collisions between gas particles, even in interstellar spaces, still occur frequently enough to randomize gas motions.
- We call stars a *collisionless* fluid but gas in the ISM is *collisional*.

6.2 *Potential Theory*

Returning to the case of stars, we are in the physical domain where strong collisions are rare, but the particles do have long-range interactions. The zeroth-order response of the stars to each other is the one-to-many response of the star to the aggregate force exerted by all the other stars in the system. At long ranges, these forces can be approximated as a smoothed gravitational potential, further abetted by the presence of dark matter in the system which contributes to the potential. Dark matter is, in general, important but non-dominant through most of the visible parts of galaxies with the effect of the dark matter becoming larger outside of galaxy centres.

Given a spatial distribution of stars, we consider the smoothed distribution of mass density $\rho \approx \langle M_\star \rangle n_\star$. On scales much larger than the typical separation between stars, this is a good approximation. If we consider the ratio of forces between the Sun and the nearest star system α Cen (both approximately $M \sim 1\ M_\odot$ separated by 1.3 pc) compared to the force between the Sun and the aggregation of stars in the Galaxy considered to be a point mass at the Galactic centre gives ($M \sim 5 \times 10^{10}\ M_\odot$ at a distance of 8.5 kpc) shows that

$$\frac{F_{\text{gal}}}{F_{\alpha\ \text{Cen}}} = \frac{M_{\text{gal}}}{M_\odot} \left(\frac{1.3\ \text{pc}}{8500\ \text{pc}} \right)^2 \sim 1200. \tag{6.13}$$

Thus, the Sun's motion is dominated by the aggregate forces of all the stars in the Galaxy rather than the nearby stars.

For the smooth density distribution, we can leverage the mathematics of the Coulomb force. Both gravity and electric force are $1/r^2$ forces that depend on the product of the charges (considering mass as 'gravitational charge') and a coupling constant. In this section, we specifically consider a spherical-polar coordinate system where the origin is at $r = 0$. The polar and circumferential terms (θ and ϕ) don't matter since we will consider spherically symmetric mass distributions. We can relate the mass distribution of a system, represented by a spatially varying distribution of density $\rho(\mathbf{r})$ to a gravitational potential $\Phi(\mathbf{r})$. Here the gravitational potential precisely analogous to an electrical potential (i.e., voltage) in that it is a gravitational energy per unit mass ('gravitational charge'). Thus, Φ has units of J/kg. We can also write down a force per unit mass as $\mathbf{g}$. Here, we are being sly and using $\mathbf{g}$ as the gravitational field that can act on a test mass, so this has dimensions of force per unit mass analogous to the electric field being force per unit charge. We use the variable g because a force per unit mass is just an acceleration, or so Newton would say.

There is an elegant mathematical relationship between ρ and Φ and the odds are mixed that you have actually seen this math al-

ready. If you have taken an introductory electrodynamics course (PHYS 381 here at U. Alberta), this will be familiar. If not, then you can skip the rest of this paragraph. It's only here to draw a parallel to something you've seen. If you are pushing onward, we note that by simply repunctuating a bunch of math that you may have already seen[1], we can arrive at Poisson's equation for gravity:

[1] Like what's in Griffith's *Introduction to Electrodynamics*

$$\nabla^2\Phi(\mathbf{r}) = 4\pi G\rho(\mathbf{r}), \tag{6.14}$$

and also the standard field-potential relationship that $-\nabla\Phi = \mathbf{g}$ as well as $\nabla \cdot \mathbf{g} = -4\pi G\rho(\mathbf{r})$. This looks just like its electrodynamical counterpart $\nabla \cdot \mathbf{E} = \rho_{\text{charge}}(\mathbf{r})/\epsilon_0$.

Density Distributions – Poisson's equation leads to the study of potential-density pairs that represent self-gravitating structures. There are a few useful representative pairs that we can use in this class. We should star by considering the simplest case, which is the gravitational potential for a point mass, which is just a mass M at the origin. In potential theory, this would follow a delta-function mass distribution $\rho(\mathbf{r}) = M\delta^3(\mathbf{r})$, but this has the outcome that is more familiar from first-year physics:

$$\Phi(\mathbf{r}) = -\frac{GM}{r} \tag{6.15}$$

where r is the distance away from the point mass at the origin.

A standard pedagogically useful form is the singular isothermal sphere. It is spherically symmetric so only depends on distance from the origin r:

$$\rho_{\text{SIS}} = \rho(r_0)\left(\frac{r}{r_0}\right)^{-2}. \tag{6.16}$$

The singular isothermal sphere is non-physical as the density diverges as $r \to 0$ and the mass diverges as $r \to \infty$. However, central density divergence is commonly referred to as *cusp* in the context of galaxy dynamics.

The Plummer sphere instead uses an inner radius scale, a_{P}, to "soften" the density distribution:

$$\rho_{\text{P}}(r) = \frac{3a_{\text{P}}^2}{4\pi}\frac{M}{(r^2 + a_{\text{P}}^2)^{5/2}}. \tag{6.17}$$

where M is the total mass of the system. The Plummer sphere is a good, but simple description for star clusters.

Finally, a common description used for dark matter halos of galaxies is the Navarro-Frenk-White (NFW) density distribution

$$\rho_{\text{NFW}}(r) = \frac{\rho_{\text{NFW}}}{(r/a_{\text{NFW}})(1 + r/a_{\text{NFW}})^2} \tag{6.18}$$

that has a characteristic density and length scale of $\rho_{\rm NFW}$ and $a_{\rm NFW}$.

CIRCULAR VELOCITY CURVES – For the spherically symmetric density distributions (and **only** for spherically symmetric distributions), we can also consider the circular orbit speed for an object at a radius r denoted V_c.

$$\frac{V_c^2(r)}{r} = -\mathbf{g}(r)\hat{r} = \frac{GM(r)}{r^2} \tag{6.19}$$

where the first equality is just the statement that the centripetal acceleration is provided by the gravitational acceleration and the second equality follows by considering Poisson's equation for the gravitational potential. The important thing in this equation is the $M(r)$ notation which is the mass of all material inside a radius of r. Formally, we will define:

$$M(r) = \int_0^r ds\, 4\pi s^2 \rho(s). \tag{6.20}$$

With this advanced version of the mass distribution, we can get away with treating the density distributions with all the sophistication of first-year physics because of Newton's theorem which holds that *for a spherically symmetric mass distribution*: the force on a test mass m at a radius of r from the centre has magnitude $GmM(r)/r^2$. The mass at $r' > r$ exerts no net force on the object. In this case, we arrive at the equation for a circular velocity curve for a spherically symmetric mass distribution:

$$V_c^2(r) = \frac{GM(r)}{r}. \tag{6.21}$$

We consider this orbital speed since it can be compared to observations using the Doppler shift of emission from stars and gas. By examining the velocity profiles of objects as a function of radius, $V_c(r)$, the density distribution of matter can be inferred. While the exact mechanics of this relies on characterizing the the departures from spherical symmetry for non-circular object, the spherical model serves as a good reference.

The first case to consider here is the point-mass density distribution and associated gravitational potential. In this case $M(r) = M$ for $r > 0$, so that the circular velocity curve is

$$\begin{aligned} \frac{V_c^2}{r} &= \frac{GM}{r^2} \\ V_c &= \sqrt{\frac{GM}{r}} \end{aligned} \tag{6.22}$$

In this case, V_c drops off proportional to $r^{-1/2}$, leading to two important conclusions. First, when we are in a region outside of a spherical mass distribution, it behaves as if it were a point mass at the origin

and will follow this rotational profile. Or, if $\rho(r > R) = 0$, then $V_c = \sqrt{GM/r}$ for all $r > R$ where M is the total mass in the system. This holds even from non-spherical systems if you consider distances sufficiently far from the object so that it can be treated as a point mass at the origin. Second, this serves as a limit for the rate at which the circular velocity curve can decline. Since there is no negative mass, $\rho \geq 0$, and thus, the velocity curve cannot fall off with distance faster than $r^{-1/2}$.

As another example, the singular isothermal sphere the orbital speed is a constant. We can see this by considering $M_{\rm SIS}(r)$

$$\begin{aligned} M_{\rm SIS}(r) &= \int_0^r ds\, 4\pi s^2 \rho_{\rm SIS}(s) && (6.23)\\ &= \int_0^r ds\, 4\pi s^2 \rho(r_0) \left(\frac{s}{r_0}\right)^{-2} && (6.24)\\ &= 4\pi r_0^2 \rho(r_0) \int_0^r s^2 s ds && (6.25)\\ &= 4\pi r_0^2 \rho(r_0) \int_0^r ds && (6.26)\\ &= 4\pi r_0^2 \rho(r_0) r && (6.27) \end{aligned}$$

Inserting this into Equation 6.21, we get

$$V_{\rm SIS}^2(r) = \frac{4\pi G r_0^2 \rho(r_0) r}{r} = 4\pi G r_0^2 \rho(r_0) \qquad (6.28)$$

which is a constant because the mass in each shell between radii of r and $r + dr$ is a constant and the growth in mass precisely balances the $1/r$ drop off in the effectiveness of the force.

We can thus, by integrating the mass distributions using Equation 6.20, calculate the circular velocity curves for any spherically symmetric mass distribution. Moreover, we can also interpret the circular velocity curves as the contributions different mass distribution. If we have two density distributions that give $M_1(r)$ and $M_2(r)$, we can find the result velocity curve:

$$V_c^2 = \frac{G[M_1(r) + M_2(r)]}{r}. \qquad (6.29)$$

Of course, not all density distributions are spherically symmetric, galaxy disks being the obvious notable exception that is currently beyond the scope of this approach. We simply lack the time to go through this in full detail; the math is not much harder than what we've seen so far.

Key Points

- Because collisions are irrelevant, we can study stellar orbits by considering their motion in a smooth galactic potential.
- The mass density distribution sets the galactic potential. We wrote down three standard mass density distributions (Equations 6.16, 6.17, 6.18).
- For spherically symmetric mass distributions, the circular orbit velocity is set by the amount of mass in the distribution within a radius r of the centre (Equation 6.21).

6.3 *Relaxation*

We have made an argument up to here that only the far-field distribution of stars matters for the motion of a star and that is broadly true on the scale of a *dynamical time*, which is the characteristic time for the global potential to significantly change the direction of a particle's motion. This is a rough timescale and can be estimate as $t_{\rm dyn} = mv/F$ where mv is the momentum of the particle and F is the force from the potential. We're being all fancy here because this is just velocity over acceleration but we used a up above and I'm trying to be good about notation. For a star on a circular orbit, this timescale is $t_{\rm dyn} = v/(v^2/r) = r/v \sim t_{\rm orb}$ where $t_{\rm orb}$ is the orbital timescale. Indeed, r/v is in general a good estimate of the dynamical time. Since this is related to the time it takes to traverse a system, we also call this the *crossing time* $t_{\rm cross}$.

However, if you consider several stars in orbit around the centre of a galaxy, a group of stars a distance r from the centre will all feel approximately the same net force from the potential and thus all undergo orbits together. These stars are co-moving and can be treated as moving together in a common rotating frame. For stars in the vicinity of the Sun, the stars are moving around the centre at a circular speed of $v_c = 220$ km s^{-1}. However, as suggested by the collision time argument using a velocity dispersion σ_v, the stars have some random velocity components relative to each other and are thus fluctuating closer and farther apart from each other. These relatively nearby stars thus, exert changing forces stars passing by each other. Over time, these random fluctuations will exchange energy between the stars using the long-range gravitational interaction. We know these forces are weak (as above) but over a long enough timescale they can perturb the velocity of passing stars. But how long is this *relaxation timescale*? This timescale represents the time that it takes

for a star to exchange significant energy and momentum with other nearby stars.

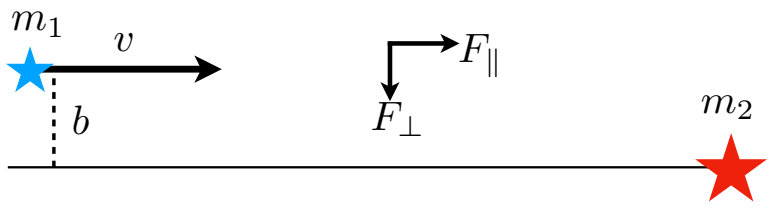


Figure 6.1: The impact parameter geometry.

To estimate the timescale, we consider the gravitational interaction of two stars at large distances. Since the interaction is necessarily weak, we use an *impulse approximation* to calculate the change in the star's velocity. Specifically, the interaction time is considered to be relatively short compared to the full evolution of the system so the change in momentum can be calculated through the total impulse. The problem geometry is illustrated at Figure 6.1. The star m_1 is moving relative to m_2 at a speed v. The impact parameter b is the distance between the line going through the centre of m_2 and the path of star m_1. The gravitational force $\mathbf{F}$ can be decomposed into two components that run parallel and perpendicular to the initial direction of motion: $F_\parallel$ and $F_\perp$. The force component $F_\parallel$ accelerates m_1 toward m_2 until closest approach at time $t = 0$ and then then deccelerates the star afterward. The force $F_\perp$ actually changes the direction of the motion of m_1. Using Newton's Law of Gravitation, we can express

$$F_\perp = \frac{Gm_1m_2}{b^2+(vt)^2}\frac{b}{\sqrt{b^2+(vt)^2}} = \frac{Gm_1m_2b}{[b^2+(vt)^2]^{3/2}} \tag{6.30}$$

The change in the perpendicular momentum is then just the impulse of this component of the force, i.e., we integrate over all time.

$$\Delta p_\perp = \int_{-\infty}^{\infty} dt \frac{Gm_1m_2b}{[b^2+(vt)^2]^{3/2}} = \frac{2Gm_1m_2}{bv}. \tag{6.31}$$

We can then calculate the angle of deflection that m_1 is diverted by through this interaction as

$$\phi = \frac{\Delta v_\perp}{v} = \frac{\Delta p_\perp}{m_1 v} = \frac{2Gm_2}{bv^2}. \tag{6.32}$$

Taking our typical stellar parameters for the Galaxy as given in the table at the beginning of the chapter and an impact parameter of 1 pc, we see that $\phi \approx 4 \times 10^{-6}$ rad which is a post hoc justification of the impulse approximation. High-fives all around.

We will consider an object to be relaxed once the original trajectory of the motion is changed by a significant angle, which we can take to just mean that the accumulated change in velocity is equal in magnitude to the original velocity. But this process of weak scattering off distant objects is actually a random process so we will fulfill the relaxation condition once the root-mean-squared change in the velocity equals v^2. Thus, we calculate the average $\langle \Delta v_\perp^2 \rangle$ as a star passes through a field of nearly stars with density $n_\star$ with characteristic speed v. These stars will be found at a full range of different impact parameters, so we integrate over this distribution to find the rms scatter after time t.

$$\begin{aligned}
\langle \Delta v_\perp^2 \rangle &= \int_{b_{\rm min}}^{b_{\rm max}} db\, n_\star v t (\Delta v_\perp)^2 2\pi b && (6.33)\\
&= \int_{b_{\rm min}}^{b_{\rm max}} db\, n_\star v t \left(\frac{2Gm_2}{bv}\right)^2 2\pi b && (6.34)\\
&= \frac{8\pi G^2 m_2^2 n_\star t}{v} \int_{b_{\rm min}}^{b_{\rm max}} \frac{db}{b} && (6.35)\\
&= \frac{8\pi G^2 m_2^2 n_\star t}{v} \ln\left(\frac{b_{\rm max}}{b_{\rm min}}\right). && (6.36)
\end{aligned}$$

Under the condition that $\langle \Delta v_\perp^2 \rangle = v^2$, we can solve for the relaxation time $t_{\rm relax}$.

$$t_{\rm relax} = \frac{v^3}{8\pi G^2 m_2^2 n_\star} \frac{1}{\ln(b_{\rm max}/b_{\rm min})}. \qquad (6.37)$$

The quantity $b_{\rm max}/b_{\rm min}$ has been left as undefined up to now and is generally given the variable name $\Lambda \equiv b_{\rm max}/b_{\rm min}$. The exact values that determine $\ln\Lambda$ are only roughly defined, but we don't try too hard because of the choice of the parameters appears in the logarithm. Typically, $b_{\rm min}$ is taken as the distance required for a (rare) strong encounter, i.e., r_c in determined by Equation 6.5, which for the Galaxy is approximately 1 AU in the stellar neighbourhood. The outer scale for encounters $b_{\rm max}$ could take a range of values. For example, in a disk, it could be as small as the thickness of the disk in the perpendicular direction (300 pc) or as large as the size of the galaxy, 20 kpc. For these two cases, we would obtain $\ln\Lambda = 17.9$ to 22.1. This change is unimportant for our rough estimates: the equation is shaped by the other quantities.

Comparing Equation 6.37 to 6.8 shows that

$$t_{\rm relax} = \frac{t_c}{2\ln\Lambda}. \qquad (6.38)$$

This is a key result: the collective action of random weak encounters with a field of nearby stars is far more important than strong collisions between pairs of stars. In the time it takes for a star to have one

strong encounter, it will have its velocity redirected 40 times by weak collisions. Inserting our characteristics scalings gives

$$t_{\rm relax} = 7.0 \times 10^{14}\ {\rm yr} \left(\frac{v}{30\ {\rm km\ s^{-1}}}\right)^3 \left(\frac{\langle M\rangle}{0.2\ M_\odot}\right)^{-2} \left(\frac{n_\star}{0.1\ {\rm pc^{-3}}}\right)^{-1} \left(\frac{\ln\Lambda}{20}\right). \tag{6.39}$$

Despite all this work, we still find that the characteristics relaxation time for the solar neighbourhood is much longer than the age of the Galaxy / Universe. The relaxation time is the typical time it takes for stars to exchange and randomize their motions, to this isn't surprising. Over the relaxation time, the circular motion of the galaxy should be randomized into a spherical distribution. We're still in a disk, so the system is not relaxed.

However, globular clusters have significantly higher stellar densities and correspondingly shorter relaxation times, of order 10s of Myr to a few Gyr depending on the cluster. These system have had time to exchange energies and becomes significantly relaxed. This leads to several interesting outcomes for people who study objects in such clusters (Prof. Heinke is quite the fan). First, there are significant numbers of stellar collisions over the lifetime of the clusters leading to creation of peculiar binary systems. Further, mass segregation occurs where low mass stars are found at larger typical radii than high mass stars. This is a consequence of energy equipartition where every star has statistically the same total energy.

VIRIAL THEOREM – Some further insight can be gleaned through the application of the *virial theorem*.

$$\frac{1}{2}\langle \ddot{I}\rangle + 2\langle K\rangle + \langle U_g\rangle = \sum_i \langle \mathbf{F}_{{\rm ext},i}\cdot\mathbf{x}_i\rangle. \tag{6.40}$$

Here, the angle brackets indicate a time average, the moment of inertia of the system is given by I, the kinetic and gravitational potential energies are given by K an U respectively and the final term represents the work done by external forces on the system. For an isolated ($\mathbf{F}_{{\rm ext},i} = 0$) galaxy in steady state ($\ddot{I} = 0$), we obtain the simple virial theorem result that

$$-2\langle K\rangle = \langle U_g\rangle \tag{6.41}$$

Considering an ensemble of N stars each with mass m in a spherically symmetric density distribution with characteristic radius R and a one-dimensional velocity dispersion of σ_v, we can write

$$K = \frac{3}{2}Nm\sigma_v^2 \tag{6.42}$$

$$U_g = -\beta\frac{GN^2m^2}{R}. \tag{6.43}$$

We have incorporated a constant β of order unity to account for the details of the mass distribution. The calculation of β for a spherically symmetric mass distribution is analogous to calculating the energy of self-assembly for a charge distribution in the study of electrodynamics. Specifically, the calculation is for bringing a thin spherical shell of mass $dM = 4\pi(r)^2\rho(r)dr$ from ∞ to the surface of the object mass $M(r)$ and radius r

$$U_g = -\int_0^\infty dr\, 4\pi r G\rho(r)M(r). \tag{6.44}$$

Then,

$$M(r) = \int_0^r ds\, 4\pi s^2\rho(s). \tag{6.45}$$

By expressing the result of this double integral as a function of the total mass (i.e., M and radius R) the value of β can be determined.

We can then use the virial theorem to return to our expression for the relaxation time and substitute in

$$\sigma_v^2 = \frac{GNm\beta}{3R}. \tag{6.46}$$

We use this expression to calculate the ratio of the relaxation time to the dynamical time (harmonizing notation such that $\sigma_v = v$ and $m_2 = m$):

$$\frac{t_{\rm relax}}{t_{\rm dyn}} = \frac{v^4}{8\pi G^2 m^2 n_\star R\ln\Lambda} \tag{6.47}$$

$$= \frac{\beta^2 G^2 N^2 m^2}{(R^2)(72\pi G^2 m^2 n_\star R\ln\Lambda)} \tag{6.48}$$

$$= \frac{\beta^2}{54}\frac{N}{\ln\Lambda}. \tag{6.49}$$

We can further show that

$$\Lambda = \frac{R}{r_c} = \frac{\beta GNm}{3v^2}\frac{v^2}{2Gm} = \frac{\beta N}{6} \tag{6.50}$$

Thus, the virial theorem gives us (with $\beta = 1$ for compactness),

$$\frac{t_{\rm relax}}{t_{\rm dyn}} = \frac{6N}{\ln(N/6)}. \tag{6.51}$$

For galaxies with $N = 10^{11}$ and long dynamical times, relaxation only a long term aspiration. However, globular clusters with $N = 10^6$ will relatively quickly, only about 5000 crossing times. Interestingly, open clusters with $N \sim 10^2$ have relaxation times comparable to their crossing times. Since open clusters are also associated with being recently formed and have crossing times comparable to the main sequence lifetimes of the highest mass stars, the dynamical

evolution of these systems is in an uncomfortable domain where our approximations are looking like a generally Bad Idea. Thus, we'll need to investigate these with numerical simulations.

Key Points

- As stars move through the galaxy, their random motions grow over time through the process of *relaxation*. Relaxation is caused by many small perturbations from long-range encounters with other stars.
- A relaxed system shows random motions without signs of a preferred velocity direction.
- Relaxation times are shorter than collision times but much longer than dynamical times.
- For isolated, steady-state systems, the virial theorem relates the gravitational and kinetic energies $2K = -U_g$.
- The Milky Way disk is not relaxed but globular clusters can be.

6.4 *The Shapes of Galaxies*

These dynamical processes are fundamentally what sets the shapes of galaxies. When we look out and see a galaxy, this isn't a fixed body of mass. Instead, the shapes are showing where the stars are located on their orbits within a galactic potential *right now*. As all the stars orbit under the collective influence of each others mass and the mass of the dark matter halo (and to a lesser degree the gas in the ISM, they set a gravitational potential and then collectively orbit in that potential. The process of relaxation and long range perturbations ends up slowly increasing their random motions.

We can start by considering the initial conditions in a star forming disk galaxy. Since gas is collisional (i.e., the time between particle collisions is short), it dissipates its random motions through gas viscosity and settles into a thin layer in the galaxy. For example, the scale height is $H \sim 50$ pc for star forming gas in the Milky Way. The action of supernova remnants and other components of stellar feedback end up stirring up the gas disk to keep it settling to an even smaller thickness. It is in this thin layer that the stars form. The gas is orbiting the centre of the galaxy on nearly circular orbits at a speed set by the circular velocity at that $R_{\rm gal}$ and a velocity dispersion appropriate for the cold molecular medium, $\sigma_v < 10$ km s^{-1}. Thus, the stars form in that gas layer have rotational motions $V_c \gg \sigma_v$ and the

motions are well ordered. Thus, the stars that form from that gas also have low velocity dispersions and high circular velocities.

As soon as stars form, they transition from a collisional to a collisionless fluid and are free from the effects of feedback, gas motions, and the radiation of other stars. They are cast loose into the broader galactic potential to start orbiting the system and are subject to the formalism developed around stellar dynamics as given above.

We can best probe the motions of stars in the solar neighbourhood where the combination of parallax measurements, proper motions and radial velocity surveys allow us to determine the phase space structure of stellar motions. We typically plot these in terms the components of a star's velocity vector referenced to the *local standard of rest* or LSR. The LSR is a hypothetical point at the Sun's location that is on a perfectly circular orbit around the centre of the galaxy. Notably the Sun's orbit is not perfectly circular. We refer to the components of the velocity vectors of stars using a coordinate system typically labelled with components $\mathbf{v} = (U, V, W)$ where U is the velocity vector component pointing toward the galactic centre, V is the velocity component toward the direction of Solar motion and W is the velocity vector out of the plane, pointing in the Galactic north direction (see Figure 6.2). In Chapter 1, we gave these these different names $v_{\rm x,gal} = U$, $v_{\rm y,gal} = V$, $v_{\rm z,gal} = W$. This is annoying now but both conventions can be used in the literature. The Sun has a velocity vector with respect to the LSR of $\mathbf{v}_\odot = (-10.0, 5.2, 7.2)$ km s^{-1}.

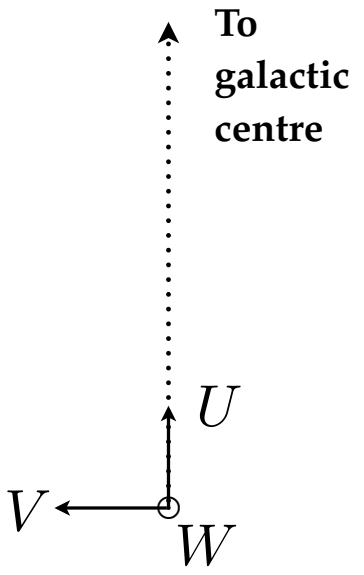


Figure 6.2: Coordinate system used to describe stellar velocity components. The Sun orbits the Galactic centre in the V direction in this figure.

The Sun is thus not on a perfectly circular orbit nor is it orbiting in a Keplerian potential, i.e., where it is orbiting a single point mass at the centre of the Galaxy. Instead, the galactic potential and the orbit are more complicated. Figure 6.3 shows the shape of the solar orbit over time. The sun oscillates in and out of the Galactic plane ($Z_{\rm gal} = 0$) and orbits the Galactic centre on an orbit that does not "closed." In other words, when the Sun returns to the line of $Y_{\rm gal} = 0$ after completing a 2π rotation around the centre of the Galaxy, it is not at the same distance $R_{\rm gal}$ that it used to be. Indeed the solar orbit moves in the space between $8 < R_{\rm gal}/{\rm kpc} < 10.4$ over the course of time. This behaviour is in contrast with the orbits in the solar system where the planets trace out elliptical orbits that nearly perfectly "close" on themselves so that the Earth traces out the same orbit year after year. The amplitude of the Sun's vertical oscillation is about 100 pc up and down out of the Galactic plane. This vertical motion is what gives the galaxy disk its thickness. The Sun is a member of the thin disk population being relatively high metallcity (solar metallicity, in fact) and having relatively small random motions compared to the LSR $\sigma_v \approx 13$ km/s with $V_c \approx 220$ km/s

The stars all around us are on orbits like this one. In their orbits,

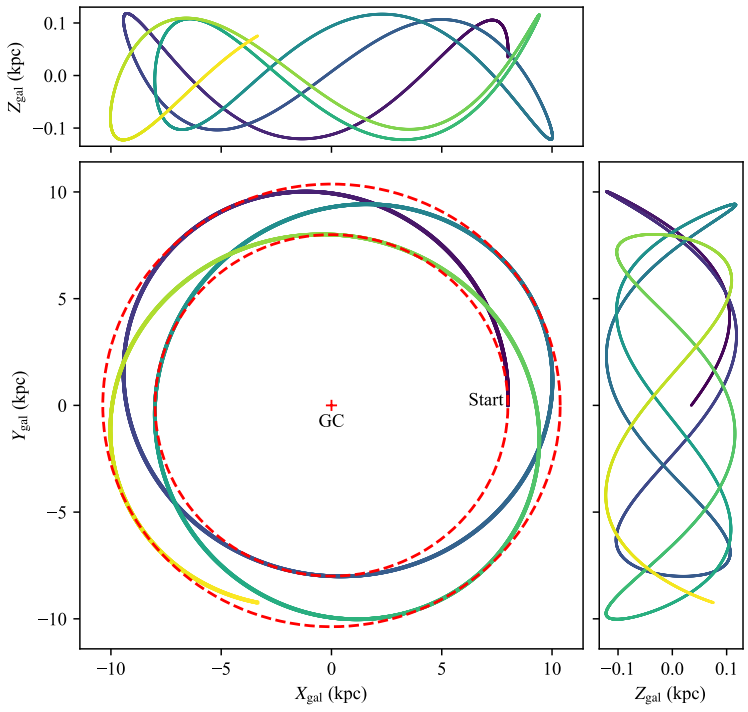


Figure 6.3: Orbit of the Sun around the Galactic Centre. The colour of the point indicates the progression of time where the darker colours are the start of the orbit and the lighter colours are later in the orbit. The two red dashed circles show $R_{gal} = 8$ kpc and $R_{gal} = 10.4$ kpc.

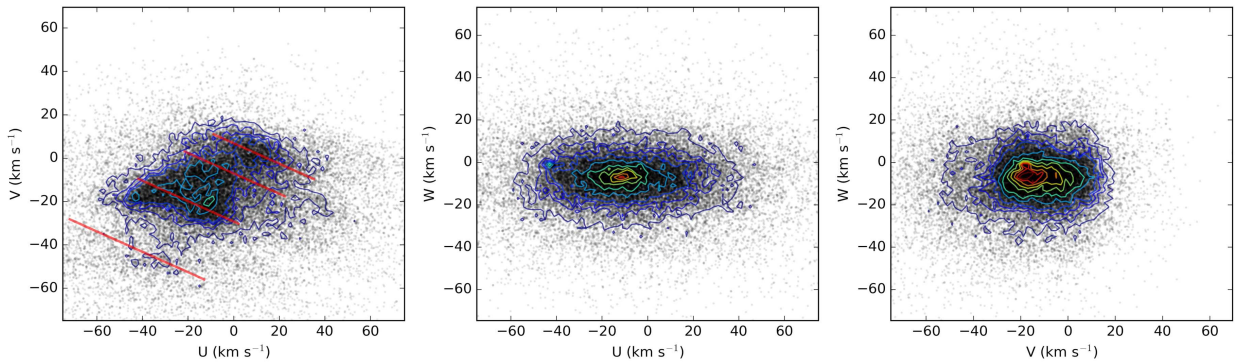


Figure 6.4: Components of stellar velocity vectors in the solar neighbourhood. The figure is drawn from Riedel et al. [2017]. The contours and lines highlight stellar kinematic features seen in these diagrams. The U vs V plot highlight stellar streams in the solar neighbourhood. The V vs W highlights asymmetric drift.

they are moving up and down out of the disk and moving in and out radially. Each of the minor encounters with a passing neighbour changes the shape of this orbit a little bit, gradually increasing the deviation from a perfectly circular orbit. In doing so, this smooths out a stellar disk in the radial direction and slowly increases the thickness in the vertical direction.

Velocity and Stellar Orbits – By examining the proper motions and radial velocities of the stars around us, we can map out the motions of stars in three dimensions. While there are not obvious structures to the stellar population in position space, when we examine features in velocity space we start to see patterns that are consistent with the pictures of stars forming at with small amounts of random motion and these motions getting larger over time.

Figure 6.4 shows the velocities of 40 000 stars with respect to LSR after correcting for Solar motion with respect to the LSR. The panels in this figure highlight the peculiarities of the Galaxy not being relaxed. In the U vs V plot, the data are clearly not centred on zero with significant structure pointing toward negative U and V. Moreover, these data are structured: there are several groups of stars, all moving with similar kinematics around the Galaxy in clusters found in *phase space*. These stars can be found all over the sky and their similarity only emerges once you consider their motions in a Galactic frame of reference. These *stellar streams* and *moving groups* are relics of their formation and are the result of when stellar clusters become unbound. Under normal conditions, it will be many dynamical times (orbital periods) before these stars are randomized into their velocities. Indeed these stars frequently show signatures of forming together, having similar chemical compositions and ages.

In the V vs W plane, another feature emerges, namely very few stars (including the Sun) are travelling on a perfectly circular orbit

and, on average, the stars are lagging behind the LSR in its perfectly circular march around the Galactic centre at 220 km s^{-1}. Relative to the LSR "pace marker", stars tend to lag behind in a process called *asymmetric drift*, so called because the offset it not symmetric around $V = 0$. The origin of asymmetric drift becomes clearer once these data are paired with stellar age measurements. Such data show that the older the stars are, the more the stars lag the LSR. This 'lagging' isn't the stars going more slowly on circular orbits. Instead, these objects are moving around on non-circular orbits with significant motion in the U and W motion.

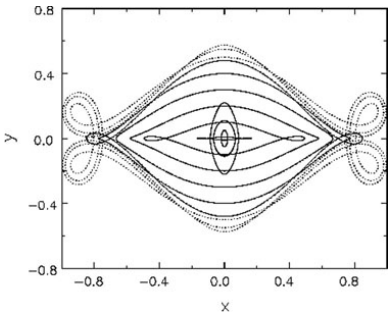


Figure 6.5: Stellar orbit families in a bar-like potential from Sellwood [2014]. Dynamics is awesome yet terrifying.

BARS – In the centres of galaxies, the shapes of stellar orbits can develop in a *stellar bar*. In the bar, stellar orbits can take on some truly bizarre shapes such as those shown in Figure 6.5. That figure shows orbits in a bar potential in a rotating frame that is fixed with respect to the stellar bar. Over time the bar rotates around the centre of the galaxy with a fixed *pattern speed* so that the shape rotates as a solid object. However, the individual stars are not fixed in one place with respect to the bar. Instead, they take on the weird orbits shown in Figure 6.5. Bars are self-sustaining and these orbital patterns keep the bar position in the same configuration on average. Our Milky Way galaxy has a bar in the centre.

Stellar bars appear to be relatively long lived stellar dynamical patterns. Looking at the incidence of bars in massive galaxies over cosmic time ($z = 0$ to $z \sim 0.6$), the fraction appears constant at 30%. However, the bar fraction in low-mass galaxies increases over the same interval. This census is consistent with a picture where a bar is an emergent phenomenon that develops after a galaxy has formed and settled. Bars also appear to be thicker than the disk component of the galaxy, as is suggested by the shape of our own Milky Way (Figure 6.6).


Figure 6.6: The the Diffuse IR Background Explorer image of the Milky Way. The structure of the bulge is visible and the "peanut" shape provided some of the first evidence that the Milky Way was a barred galaxy.

Bars represent families of orbits such as those shown in Figure 6.5. Since the light of a galaxy traces where the stars are found and the orbits have distinctly non-elliptical shapes in the rotating frame, they take on the bar-like shape in the light. The orbits sketched in the Figure are grouped into families, with the most obvious bar-like set called the x_1 (solid lines) and the perpendicular set of orbits being called the x_2 family. The x_1 orbits are particularly interesting because stars on these orbits can be on nearly-radial orbits moving from the end of the bar into the interior of the galaxy in one quarter of a dynamical time. This class of orbits is interesting from the perspective of feeding *active galactic nuclei* (AGN). AGN result from accretion onto the central supermassive black holes in galaxies. These orbits can channel gas from the disk of the galaxy into the interior rather

quickly. There is not, however, clear evidence that AGN are more frequent in barred galaxies, which undermines a neat and clean picture of bars as the dominant feeding mechanism.

Bars appear to be a rather simple disk instability found inside the co-rotation radius for the pattern. Bar systems appear to be "material," namely the same stars are participating in a bar over the course of several dynamical times, which stands in contrast with spiral arms (discussed next). They are trapped onto elongated orbits and their mutual self-gravity keeps them on those orbits. Once formed, bars also are unstable to a buckling instability where their thicknesses are increased by self-gravitational buckling (bending in the vertical direction).

Both spiral structure and bar instabilities (plus a wealth of other, less significant effects) redistribute the mass and angular momentum within galaxies. This changes the shapes of the stellar mass and light within galaxies. Moreover, since galaxies are also filled with gas, these instabilities move the gas around but also form new generations of stars when that mass is concentrated.

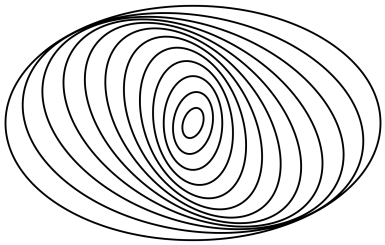


Figure 6.7: Spiral patterns emerging from converging stellar orbits. Image credit: Wikipedia/DbenBenn licensed under CC-SA-3.0.

Spiral Arms – Spiral arms are similar to stellar bars in that the orbits of stars in a disk of galaxies tend to align together. Where stellar orbits converge, the local mass of the galaxy will increases and the collective action of stars will pull more orbits close to this location leading to a perturbation. Figure 6.7 shows how a nested set of elliptical orbits with different orientations could converge together and form a spiral pattern. Since orbits aren't closed, as they are in this figure, the representation is more illustrative than perfectly accurate, but it is accurate in the general sense that the alignment of orbits where stars converge gives a small increase in density. Typically, these perturbations in mass density are small, only a few percent change in the local density of the disk. However, spiral arms are typically quite striking when viewed in images of galaxies.

There is not, at present, a fully consistent model that explains how spiral arms in galaxies develop, but a gravitationally unstable disk appears to be a necessary though not sufficient condition. While we won't dive into this in a lot of detail, the condition of gravitational instability just means that the disk will tend to fragment into smaller structures. The instability arises because of gravity: the disk will concentrate into rings or clumps if the material in the disk can gather into structure and gravitationally collapse into concentrated structures faster than the shearing motions in a galaxy disk can pull the material apart. This shear arises because the rotation curves of disks tend to be close to constant ("flat") because of the presence of dark matter, i.e., V_c is a constant. As such, the angular frequency of rota-

tion falls off with radius $\Omega = V_c / R_{\rm gal} \propto R_{\rm gal}^{-1}$, or simply that orbits closer to the centre of the galaxy complete their orbits faster than orbits farther out. This means that there will be a shearing motion in the disk so that stars and gas at inner radii will catch up and pass stars and gas at outer radii.

The consensus is that spiral arms are gravitationally driven density waves in the stellar disk of stars (not just the gas). The emerging view is that spiral structure is *transient* and a given spiral pattern only lasts a few disk rotations [Sellwood, 2014], which is largely motivated by simulation work. This model stands in contrast to the idea that spiral arms are *material arms*, where the same material is propagating in the arm around the centre of the galaxy. The major challenge to the idea of material arm spiral structure is the "windup problem," which just points out that spiral patterns can't effectively persist in a shearing disk. Simulations of spiral structure finds that the pattern speed model is a good one for spiral structure, namely spiral arm perturbations propagate around the galaxy at a near-constant pattern speed.

Spiral instabilities show two key effects of the disks of galaxies. The inflate the random motions of stars, closing the gap between our expected and observed velocity dispersions (Section 6.5). They also serve to transport angular momentum outward the the galaxy, allowing a larger amount of matter to move inward.

While the origins of spiral structure remain obscure, it is clear that dynamical interactions with other galaxies can drive spiral structure through tidal interactions. The classic Whirlpool galaxy has its spiral structure driven by the tidal field of the passing M51b (e.g., Figure 6.8). It may be that all spiral patterns are initiated by close tidal interactions, either from low luminosity dwarf galaxies or even dark matter subhalos.

The reason for this outsized influence of these small spiral arm perturbations in galaxy images comes from star formation. The stellar orbits lead to a local potential minimum in the galaxy disk. This perturbation gravitationally attracts stars but also the galaxy's ISM, which it compresses. This compression triggers the formation of the cold neutral medium, the cold molecular medium and then a generation of newly formed stars. Since newly formed stars include high mass stars with short lifetimes, the sites of recent star formation are bright and relatively blue compared to the rest of the galaxy disk. This is illustrated in Figure 6.8, which shows the Whirlpool galaxy in blue (left) and near infrared (right) light. The spiral structure of the galaxy is much more prominent in the blue light since only the young, high mass stars emit prominently in this band. In contrast, these stars don't stand out as much in the near infrared since nearly

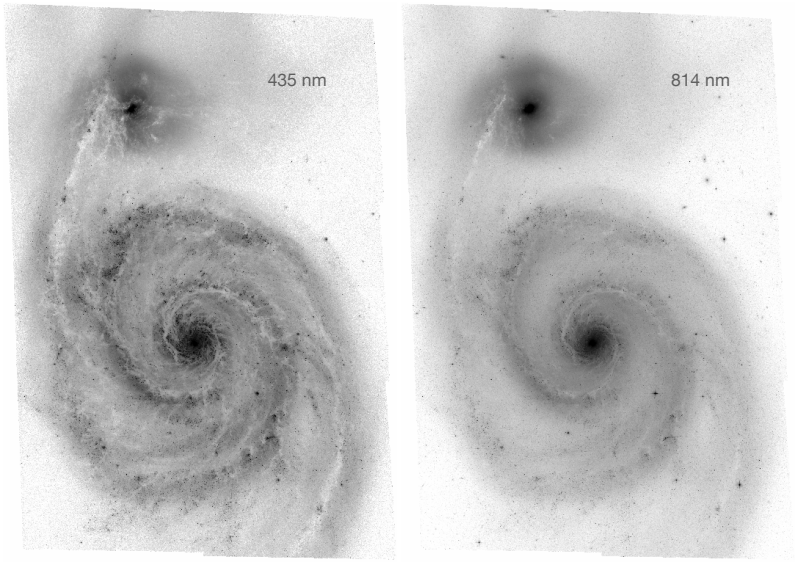


Figure 6.8: The Whirlpool Galaxy (M51) seen in two different wavelengths ($\lambda = 435$ nm and $\lambda = 814$ nw). Spiral structure is much more prominent in the shorter (bluer) wavelengths.

all stars emit significantly. The high mass stars remain comparatively luminous but the sheer number of the lower mass stars emitting in this band drowns out the high mass stars.

It is instructive to compare the timescales involved. In the Milky Way's disk, the orbital (dynamical) time for the Sun is 250 Myr. The formation time for stars within molecular clouds is < 3 Myr and the lifetime of a high mass star is < 10 Myr. This means that stars an form and the high mass stars from that formation process will not make it around the galactic disk before ending their lives. Indeed, once triggered in a spiral arm, these high luminosity,high mass stars will only make it through $< 5\%$ of a rotation before dying. Thus, the spiral shapes in galaxies are not as significant as they appear: they are just where star formation makes the galaxy light up and then fade away.

Elliptical Galaxies and Bulges – The final shape of galaxy to consider is that of the spheroidal systems like elliptical galaxies, the bulges of galaxies, and their halos. These systems have large random motions with little or no sign of ordered motions. Their shapes reflect this random motion and the orbits of galaxies within their potentials. In this case, stellar orbits are be moving around the centre of galaxies with no preferred direction. Thus, stars will frequently pass each other that are orbiting in opposite direction. There can still be some rotation to these systems but it is typically small.

In spheroidal systems, the random motions of the stars in different dimensions (i.e., σ_x vs σ_y vs σ_z) can have different magnitudes

and this leads to the elliptical shapes of the galaxies. If a galaxy has $\sigma_x > \sigma_y$ then the galaxy will have a larger extent in the x-direction compared to the y-direction. The random nature of the motions leads to the spheroidal shape. This scenario is in contrast with collisional systems like dense gases where random motions in one direction will be rapidly equalized across other dimensions through collisions. However, even elliptical galaxies are non-collisional and these asymmetric stellar motions and the small rotational velocity show these galaxies are not relaxed.

The boring appearance of elliptical systems can conceal fascinating dynamical features. For example, some ellipticals show two kinematically distinct rotating populations: one population going in one direction and the other population is decoupled and rotating in a different plane. These features are thought to reflect a merger-based origin for such elliptical systems. The two populations reflect the original motions from the stars of the two galaxies that merged together to form the system.

Key Points

- The shapes of galaxies are set by the the orbits of stars inside the galactic potential. For example, the thickness of a stellar disk is set by how far stars move up and down out of the midplane.
- In disk galaxies, stars form on nearly circular orbits with minimal random motions.
- As stars get older, relaxation increases their random motions.
- Stellar orbits in a galaxy are not "closed" and individual stars move in and out radially and up and down vertically (Figure 6.3).
- Stellar bars develop in the centres of galaxies and the bar is a complicated pattern of stellar orbits (Figure 6.5) that keep the stars near each other.
- Spiral arms form from a small gravitational perturbation in the stellar disk, but appear significant because this triggers the formation of stars including luminous high-mass stars that highlight the region.
- Elliptical systems, spheroidal galaxies, and galactic bulges are characterized by strong random motions but the magnitudes of these motions are usually asymmetric since the systems are not relaxed.

6.5 *Is Too Much Relaxation a Bad Thing?*

The net indication of asymmetric drift is that older stars have been perturbed off the circular orbits on which they formed and are building up random motions. In terms of the velocity ellipsoid, estimates of the components in cylindrical coordinates give

$$\begin{aligned} \sigma_U &= 38 \text{ km s}^{-1} \\ \sigma_V &= 25 \text{ km s}^{-1} \\ \sigma_W &= 23 \text{ km s}^{-1}. \end{aligned}$$

If we assume that the stellar population is born with velocity dispersion approximately equal to the gas velocity dispersion of $\sigma_{\rm gas} = 10$ km s^{-1}, this implies that their random motions have grown significantly over the lifetime of the disk. We found in Equation 6.39 that the relaxation time in the solar circle is $\sim 3 \times 10^{13}$ yr whereas

the age of the disk is $\sim$ 10 Gyr so the disk should only be about 1/3000th of the way to relaxation. Instead, if the random velocities are 30 km s^{-1} and the circular velocity is 220 km s^{-1} the system is $(\sigma/v)^2 = 0.018 \sim 2\%$ relaxed. The random motions that have been built up should not be nearly as large as we observe and we conclude that the disk is being *heated* (i.e., made more random) by processes other than two body relaxation.

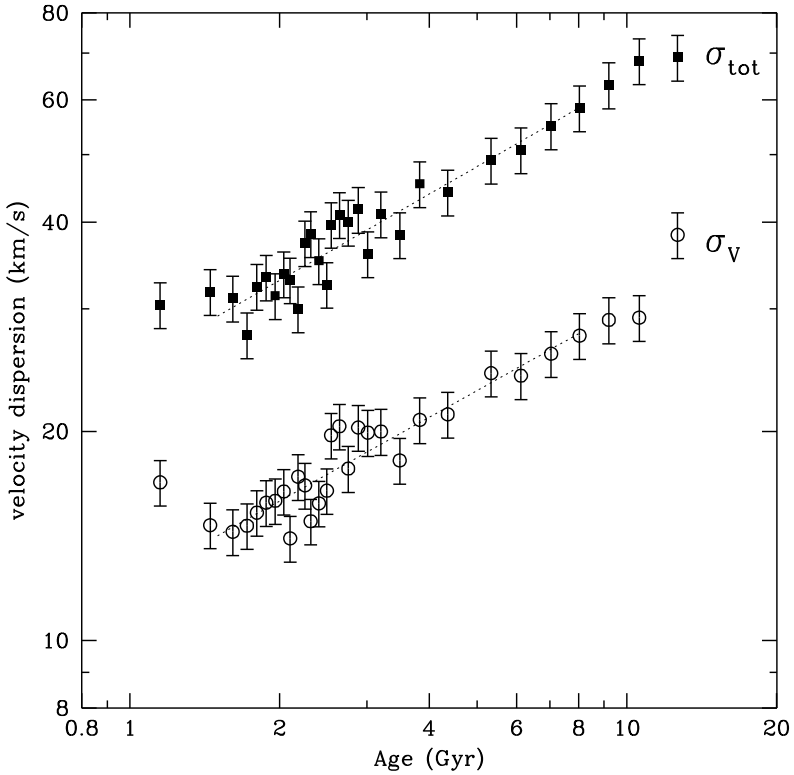


Figure 6.9: Stellar velocity dispersion vs. age for stars in the sample of Holmberg et al. [2009]. The degree of random motions rises clearly with the age of the stellar population under consideration showing that the disk heating process is *secular* rather than precipitated by a small number of events in the past.

This heating appears to be a secular process in that the amount of random motion grows with stellar age. As seen in Figure 6.9, the degree of random motions seen in a group of stars rises significantly and apparently smootly with age. We must rely on physical effects that act continuously over the age of the galaxy to drive up the stellar motions.

Fortunately, we look out and see that there is ample structure in the disk of the galaxy that can explain the excess of random motions observed in the stellar system. The most obvious of these is seen in our disk and in others: spiral instabilities. Thin stellar disk systems can develop large scale gravitational instabilities that perturb the beautiful symmetric potentials that we have been exploring. The

other clear candidate for stellar perturbations are molecular clouds. These star forming structures can be massive ($> 10^5\ M_\odot$) and relatively compact structures ($r < 50$ pc) in star forming galaxies which can tug on stars in the disk and increase their random motions.

Key Points

- Older stars show higher random motions. As stars orbit the disk, their random motions get amplified by more encounters.
- The random motions we observe are too large to be explained entirely by relaxation with other stars. Instead, we think that spiral arms and encounters with molecular clouds help increase the random motions.

6.6 Conclusions

As we wrap up our brief exposure to the underlying mechanics of stellar dynamics, we should emphasize why the physics we've ground out in this chapter is important. Critically, stellar dynamics establishes the shapes of galaxies. Since stars at the densities that typify galaxies are collisionless, the shape of the galaxy ultimately reflects the ensemble motions of a suite of stars in a self-consistently determined gravitational potential. We have, to now, largely neglected the influence of dark matter, but our knowledge of the dark matter points to it being well modelled by a collisionless fluid like stars. Thus, the orbits of stars are shaped by their momenta and positions within the galaxy's potential. Stars trace their orbits and emit light on these trajectories and we observe the light. Thus, the shapes of galaxies are set by the motions of stars inside the potential.

7
The Principles of Galaxy Evolution

With the essential ingredients developed, we conclude by exploring galaxy evolution, we an aim of understanding the questions we first posed about the galaxy population back in Chapter 5.

Galaxy Evolution Questions

- Why do galaxies have the shapes that they do?
- What processes shape the light we receive from galaxies?
- Why do galaxies have the contents that they do?
- How do the answers to these questions change as the Universe evolves?

Galaxy evolution is the combination of two families of physical effects: internal processes that shape how the galaxy evolves in isolation and then the interaction of the galaxy with others in groups and clusters.

7.1 Dark Matter Governs Galaxy Evolution

While we have largely ignored it and focused on the physics of baryons, the galaxy evolution is governed primarily by how the density distribution of dark matter is changing over time. As mentioned back at the beginning of the book, our current model of dark matter is that it is a Weakly Interacting Massive Particle (WIMP), which is a set of constraints developed by astronomical observations into the ΛCDM model for galaxy evolution.

By mass, there is approximately $8\times$ as much dark matter as there is baryonic matter in the Universe and the behaviour of the dark matter sets the conditions for the galaxy evolution that we trace. Galaxy evolution is thus a story about how dark matter forms the structures that it does and then the baryonic matter follows the gravitational

potential set by the dark matter. For our galaxy evolution model to yield the properties that it does, the dominant model of dark matter assumes that the dark matter is:

- Cold – This means that the random thermal motions of dark matter are relatively small so that the particles are non-relativistic. The velocity distribution of the dark matter should be comparable to the rotational speeds in the outskirts of galaxies, namely 100-300 km/s given the mass of galaxies. This is because the rotational speed of a galaxy it its outskirts is tracing the mass in the halo, which is dominated by the dark matter. Warm and hot dark matter particles would have substantially higher velocities and thus the particles would escape from the gravitational potential cause by the dark matter, smoothing out the dark matter distribution.
- Not Collisional – In cases where we evidence for dark matter halos passing into or through each other, we do not see evidence that the individual dark matter particles are colliding with each other. Instead, like stars, dark matter clouds will pass through each other, only interacting through their large scale average gravitational interactions and not through the particle-particle collisions that typify gas clouds interacting.
- Non-interacting – This quality is equivalent to saying that dark matter is dark. This means that we don't see evidence for it through the electromagnetic spectrum or other particle interactions. From the perspective of galaxy evolution, this means that the only influence that baryonic matter and radiation has on the dark matter comes from the gravitational influence of the regular matter and not on any other physical effects.

Given these two qualities, it is possible to model the evolution of dark matter since the implied physics is fairly simple: we start with a set of dark matter particles in a large Universe-scale box and let them start to collapse to form structures entirely under their own self gravity. The only relevant physics is gravity, though this gravitation must be explored in the context of the theory of General Relativity and not Newtonian gravitation since the expansion of the Universe plays a vital backdrop for this evolution. Then, we need to understand the initial conditions for the dark matter since these do come out of the Big Bang.

THE BULLET CLUSTER – Figure 7.1 presents important empirical results about the nature of dark matter, reinforcing our model that it is non-interacting and not collisional. This image shows a multiwaveband composite of the "Bullet" cluster (i.e., 1E 0657-56). The cluster

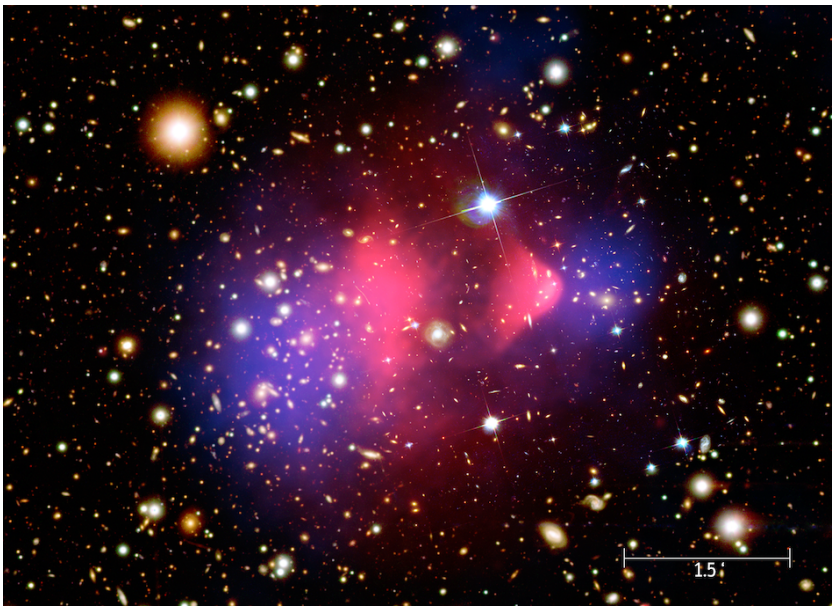


Figure 7.1: Image Credit: NASA/CXC/M. Weiss - Chandra X-Ray Observatory

is actually two clusters that have passed through each other in the process of merging. The cluster on the left is moving to the left and the cluster on the right is moving to the right. The pink/red gas in the middle of the image is traced by X-ray emission from the hot gas that was originally in both clusters. This gas is collisional and the two clouds have collided leaving the in the middle of the collision. In a galaxy cluster, most of the baryonic mass is in this X-ray emitting gas and not in the stars in the galaxies. The clusters of galaxies themselves have kept moving on since these systems are made of collisionless stars. What is more interesting is what has happened to the dark matter. In this case, the dark matter is traced in the blue background and it follows the collisionless stars. The dark matter is being detected using a technique called *weak lensing*. This technique uses images of background galaxies and measures their shapes. The statistical distribution of shapes is compared with the expected distribution of shapes. Any deviations from the expected distribution of shapes is attributed to the presence of dark matter along the line of sight, allowing us to map out the distribution of dark matter.

They key result from the Bullet cluster and similar observations is that the dark matter follows the stars even though they are not the most massive component of the galaxy cluster. If the dark matter arose from us misunderstanding the long-range nature of gravity, i.e., in theories like modified Newtonian dynamics (MOND), then the signatures of dark matter should trace the presence of most of the baryonic matter: the hot gas. Instead, the dark matter follows

the collisionless stars indicating that this is separate, collisionless mass component. Moreover, it also shows that the cross-section for self interaction of dark matter particles (i.e., a DM+DM collision) is also small or else the particles would collide with each other like the gas. This one cluster tells us a huge amount about the nature of dark matter and gives us an indication that we are largely on the correct course.

A Brief Summary of Cosmology – Without stealing too much of the themes from ASTRO 430, we need to summarize the Big Bang cosmology a bit since it forms the basis for our model for the expansion of the Universe. The fundamental assumptions of the model are that the Universe is *isotropic* and *homogeneous*. Isotropic means that there is no preferred orientation to the structures of the Universe and homogeneous means that the structure of the Universe is the same at different locations throughout. These statements are statistical in nature and refer to the large scales in the Universe (> 100 Mpc); on small scales neither are true since the interior of galaxies is clearly not the same conditions compared to a void.

The past two decades has seen a strong consensus emerge that the Universe is also geometrically *flat* meaning that the total amount of mass and energy in the Universe match the quantity needed so there is no curvature of space-time over large scales. with the field equations of general relativity and the assumptions of isotropy, homogeneity and the measured flatness, we can predict the evolution of the scale of the Universe over time using only the current measurement of its rate as an input. This current rate of expansion measurement comes from the Hubble constant (Chapter 5). A corollary of this state under the theory of relativity is that the time coordinate then behaves linearly. This relatively simple physical state does not necessarily have to be true and the reasons for geometry being so precisely flat are mysterious and ultimately trace back to the ideas in the theory of *inflation*.

Given these inputs, we can make a prediction for the scale factor of the Universe as a function of cosmic time. We define this scale factor now $a(t_0) = 1$ at the current time (t_0) so that $a(t) < a(t_0)$ for all $t < t_0$. This factor is the linear scale of distances between objects. If two objects are separated by a distance d_0 at the current time t_0, at earlier times they are separated by a distance

$$d(t) = d_0 \frac{a(t)}{at_0} = a(t) d_0 \tag{7.1}$$

because $a(t_0) = 1$. We refer to d_0 as the *comoving distance* and frequently present results over different redshifts in terms of a comoving distance. This scaling is only valid for things that are not under

the influence of their mutual self-gravitation. In that case, the local gravitational field will take over from the cosmological expansion and will govern the local motion of objects. This is only valid if the separation between the two objects is governed by the average expansion of the Universe. Once case where the points are not experiencing self gravitation is when those points are amplitude peaks of a light wave for which the cosmological expansion induces the aforementioned cosmological redshift. We can relate the redshift to the scale factor:

$$a(t) = \frac{1}{1+z}. \tag{7.2}$$

We usually just write the values of the scale factor in terms of the redshift z instead of a though the two are equivalent. This notation is also convenient since

$$\frac{\dot{a}(t_0)}{a(t_0)} = H_0(t_0) \tag{7.3}$$

where H_0 is the Hubble constant and $\dot{a}(t_0)$ refers to the time derivative of the expansion rate.

Example: *If two galaxies are separated by a comoving distance of* $d_0 = 200$ Mpc, *how far apart are these galaxies at* $z = 3$*?*
At $z = 3$, $a(t) = 1/(1+z) = 1/4$ so that $d(t) = d_0 a(t) = 200$ Mpc$/4 = 50$ Mpc.

Figure 7.2 shows the evolution of the scale factor of the Universe over the course of time and the corresponding redshift. This is derived for the concordance cosmology values: $\Omega_\Lambda = 0.7$, $\Omega_m = 0.3$, $H_0 = 70$ km/s/Mpc. The variable Ω is the mass-energy density of a component of the Universe relative to the value that would make the Universe's geometry flat. In this case Ω_Λ is the mass-energy density of the dark energy term and Ω_m is the corresponding term for the matter (dark matter and baryonic): $\Omega_\Lambda + \Omega_m = 1$ implies the Universe is flat. These values refer to the densities of these components at the current time t_0, since their relatively densities vary over time (though the sum of their contributions remains equal to 1.0 over time).

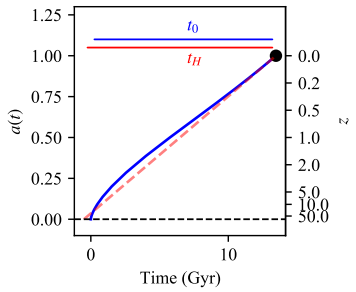


Figure 7.2: Scale factor of the Universe over time (blue solid line) and linear extrapolation of the expansion (red dashed line).

The variation of the scale factor over time is nearly linear for $a(t) > 0.2$. As such, we can make an extrapolation to the expansion of the Universe based on the Hubble constant. This is illustrated in Figure 7.2 as the red dashed line, with a slope equal to the Hubble constant. This extrapolation is pretty good, but the red dashed line hits the $y = 0$ line slightly before $t = 0$. The *Hubble time* is just the linearly extrapolated age of the Universe $t_H = 1/H_0$, which under this cosmological model is $t_H = 13.96$ Gyr. Figure 7.2 shows that this

time is a slight overestimate of the true age of the Universe, which is $t_0 = 13.47$ Gyr. We generically refer to the age of the Universe as the "Hubble time" though this figure shows that these are slightly different.

The shape of the curve in Figure 7.2 results from the contents of the Universe and largely represents two main factors. The initial trajectory of the expansion is set by the Big Bang. We don't know the causes of the Big Bang, so we accept them as initial conditions. The scale factor of the Universe is increasing but this expansion is initially decelerating. This deceleration is from the mutual attraction of all the mass-energy in the Universe. A reasonable analogy here is a ball thrown upward with some initial speed. The mutual gravity of Earth and the ball slows the initial speed down. Later in the Universe's evolution, the influence of the cosmomological constant Λ starts to become significant, which acts to re-accelerate the Universe outward. Here, our analogy with a ball thrown upward fails: this would be tantamount to the ball continuing upward with its speed accelerating. Not something we see in our usual kinematics experiments.

RECOMBINATION – In galaxy evolution, we take this trajectory for the scale factor of the Universe as a background and consider what happens to the matter (both dark and baryonic) as well as the energy. The main consequence of the scale factor of the Universe getting smaller as we look backwards in time is that it implies that the Universe was globally hotter and denser at an earlier state. The mean density of the Universe will just scale like $\rho(t) = \rho_0 a^{-3}$ and the energy density of radiation scales like $u(t) = u_0 a^{-4}$ where the extra factor of $a(t)$ comes from the expansion both making the photons lower density as well as redshifting the photons. Examining early times in this scaling implies that there was a point when the Universe was hot and dense throughout with temperatures and densities being comparable to the outer layers of stars (i.e., $T = 3000$ K and $n_{\rm H} \approx 2 \times 10^9$ m^{-3} at $z \approx 1100$). Under these conditions, matter is ionized and opaque to radiation passing through it, just like the outer layers of a star. As the Universe evolved forward from $z = 1100$, it cooled off, which prompted recombination (e.g., Equation 4.4) from ionized plasma into neutral hydrogen atoms.

Note that the term "recombination" is misleading here since the atoms of the Universe would have never been combined in the first place. However, we the analogy with our study of H II region is pretty good so we use the same word.

Once recombined, matter and light were no longer connected together through collisions. Indeed, one of the major points of quantum mechanics is that light only interacts with atoms at specific wavelengths corresponding to the energy levels of the atoms. If photons don't have the matching wavelengths, then the matter is essentially transparent to the light. At this point, radiation and matter decoupled from each other and the light propagated through

the Universe. We observe this relic radiation today after it has been redshifted from its original spectrum. Originally the radiation had a thermal (blackbody) spectrum at the temperature corresponding to the temperature of the Universe ($T \approx 3000$ K) at the time. After being redshifted, the temperature of the blackbody transforms into a blackbody with a temperature T_0:

$$T(z) = (1+z)T_0 \tag{7.4}$$

where T_0 is the current observed temperature of this radiation field: $T_0 = 2.725$ K. This temperature of blackbody generates most of its emission in the microwave section of the spectrum, following the Wien law, so the emission is referred to as the *cosmic microwave background* or CMB.

This emission is one of the defining predictions of the Big Bang theory, namely that the Universe was hot and dense at an earlier time. The predicted blackbody shape was confirmed to high precision by the Cosmic Background Explorer satellite (COBE), which launched in 1989 and garnered a few Nobel Prizes for the measurement.

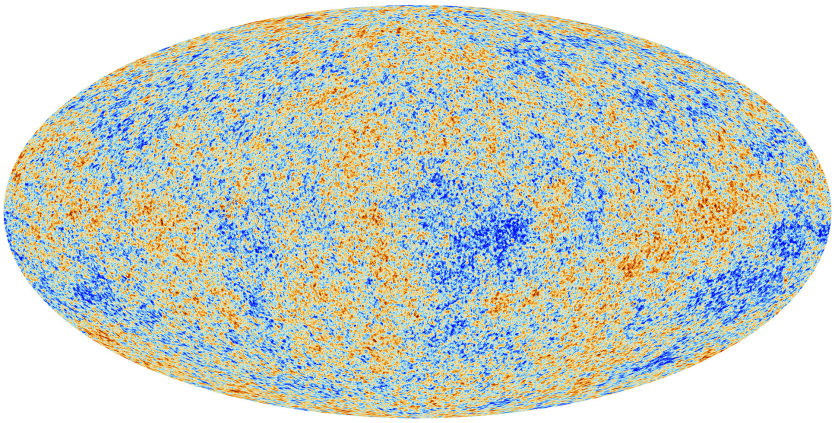


Figure 7.3: Cosmic Microwave Background fluctuations as observed by the *Planck* satellite. Image Credit: ESA and the Planck Collaboration.

In the context of galaxy evolution, the interesting thing isn't that the CMB is there or that it's 2.725 K, but rather that it isn't uniform on the sky. If you subtract off the emission from the Galaxy and other foreground objects, and correct for a Doppler shift of our solar system moving with respect to the CMB, we see that there are small fluctuations in brightness throughout the sky. These are illustrated in Figure 7.3, where the red deviations are slightly warmer than T_0 and the blue deviations are slightly cooler than T_0. These fluctuations have a typical amplitude on the μK scale and are tracing the density fluctuations in the early Universe. Regions that are slightly denser will undergo recombination slightly earlier in the Universe's

evolution (following the $n_{\rm H}^2$ density dependence on the recombination rate). Hence, the Universe will become transparent at a fixed temperature of $T = 3000$ K but this recombined part of the Universe will have its radiation redshifted slightly more than average so it will appear slightly cooler. Similarly the slightly warmer regions reflect regions where the density is lower and recombination happens a little later.

This is an amazing perspective: it quantifies how uneven the density field of the Universe is at a very early time: since the variations are μK scale on a 2.725 K blackbody, the answer is not very uneven. Even so, these density fluctuations are reflecting the underlying collapse of matter, which is driven by the collapse of dark matter and hence their amplitude at this early time provides an empirical measurement of how far along the collapse has proceeded at early time.

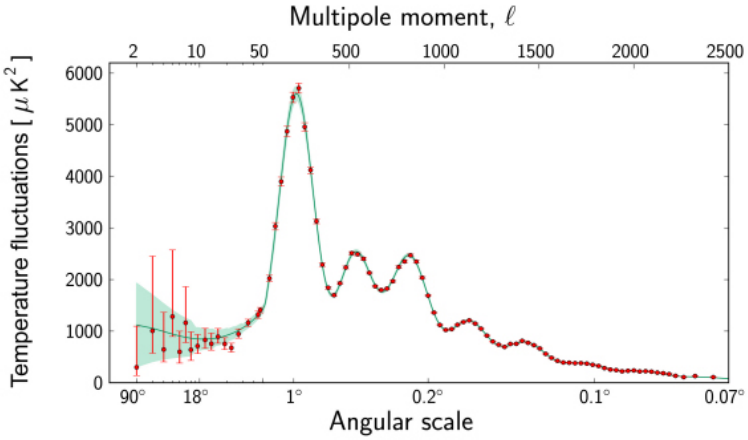


Figure 7.4: Angular power spectrum of the CMB fluctations measured by the Planck satellite. Image Credit: ESA and the Planck Collaboration.

We quantify this collapse with the *angular power spectrum* of the CMB. This is illustrated in Figure 7.4 which shows the magnitude of the fluctuations as measured by the *Planck* satellite. This "power spectrum" represents the coefficients in a spherical harmonic decomposition of the map shown in Figure 7.3. If you have seen the full quantum treatment of the hydrogen atom or some introductory chemistry, you've seen reference to these spherical harmonics. For our purposes, these a graph of the strength of the CMB signal on a range of different angular scales. The peak in the CMB power spectrum at a "multipole" number of about $\ell = 200$ which corresponds to angular scales of $\theta \sim 180^\circ/\ell$. This just means that the strongest peaks and valleys in Figure 7.3 have typical angular scale of $\sim 1^\circ$ and represent the largest scale of the Universe. The other bumps in the angular power spectrum correspond to other signatures of early Universe like sound waves in the plasma.

We can use these observations to characterize the initial fluctuations of the dark matter density distribution into analytic models and numerical simulations. The application of this information in analytic models is explored further in ASTRO 430 and later graduate-level courses. The results in numerical simulations give us less physical insight but a more correct answer where our mathematical treatment breaks down. Plus they make pretty pictures. These numerical simulations require the amplitude and physical scale of these fluctuations to seed small perturbations in a uniform density field. With these small perturbations in place, the physics is relatively simple: integrate the evolution of the density of billions of simulated dark matter particles, each representing $> 10^6\ M_\odot$ of mass under the effects of the following physics: (1) gravity and (2) the expansion of the Universe. There is no (3); the physics requires a lot of computation but is well understood.

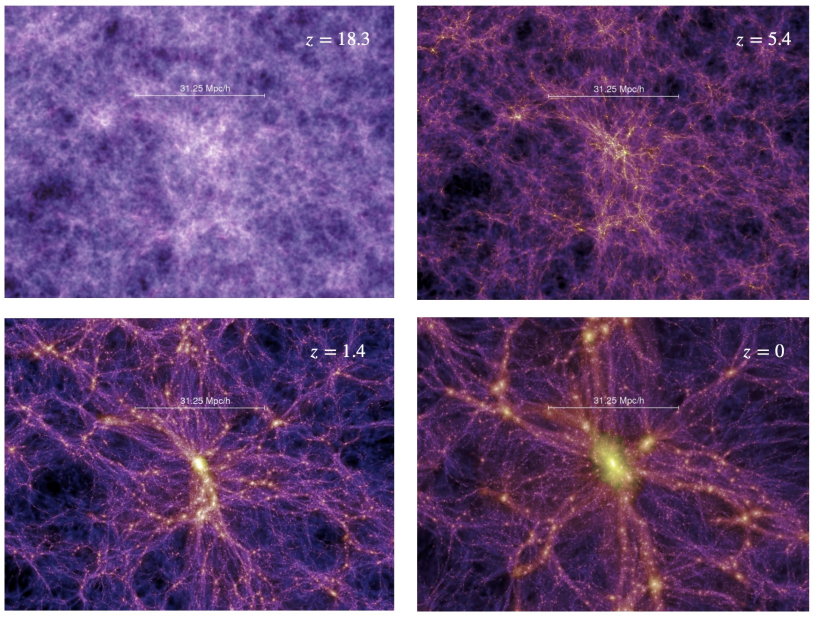


Figure 7.5: Evolution of the dark matter distribution with redshift. Data from the Millennium Simulation [Springel et al., 2005].

Figure 7.5 shows the results of one of the best dark-matter-only simulations called the *Millennium simulation*, which was published by Springel et al. [2005]. The figure shows four different snapshots at different redshifts, illustrating how dark matter gathers under its own self-gravity into progressively larger structures over time. Each of these structures is described as a dark matter *halo*. While we have tossed the term "halo" about with wild abandon earlier in the book, it has a theoretical definition as a dark matter structure that has decoupled from cosmological expansion through its own self-gravity.

Halos are the largest self-gravitating scales in the Universe but they can contain individual self-gravitating substructures called *subhalos*. The general action of dark matter halos under cosmology is that of hierarchical merging where small structures merge into progressively larger structures over time. This is clearly visible in Figure 7.5 where a single massive halo in the centre of the simulation builds up through hierarchical merging (smaller halos into larger halos) over the course of time.

Since dark matter is the dominant mass component of the Universe and is unaffected by physics that would prevent its collapse, the formation of halos drives the formation of galaxies. Baryonic matter falls under gravity into the potential wells created by dark matter, cools and collects into galaxies. Thus, it is natural to expect that the masses of galaxies will track the masses of the dark matter halos. Bigger halos will attract more of the matter making bigger galaxies. Hence, one of our main observables, namely the luminosities of galaxies and the Schechter luminosity function should be readily predicted through the mass distribution of halos.

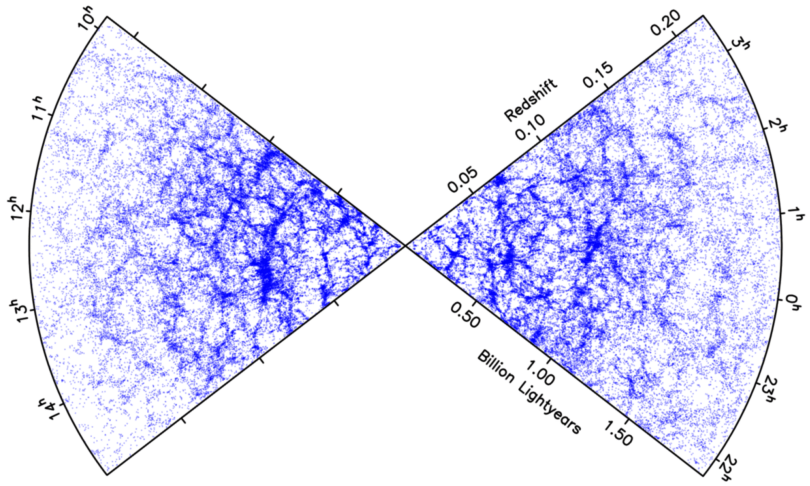


Figure 7.6: The large scale structure of the Universe as observed through the 2dF galaxy redshift survey from Colless et al. [2001].

Redshift Surveys and Large Scale Structure – Our model collapse of dark matter follows a characteristic filamentary pattern with large mass gathering through halo mergers where those filaments meet. We have an observational perspective on this structure through the hard work of galaxy redshift surveys like the SDSS and the 2dF galaxy redshift survey shown in Figure 7.6. This figure shows a large survey of galaxies made using the Anglo-Australian Telescope in Australia by Colless et al. [2001]. In that survey, they measured 80 000 individual galaxies across two narrow strips in declination but

large ranges in right ascension. Each point on the graph represents a single galaxy drawn in polar coordinates where the radial distance from the origin is the galaxy's redshift. Because of the Hubble law, the radial coordinate also maps linearly to distance (here plotted in light years rather than pc).

Figure 7.6 shows the large filamentary structure matching the same base structure as is seen in the dark matter simulations. Galaxies are gathered into filaments and sheets on large scales with vast *voids* with typical scales of 100 Mpc. The largest concentration of galaxies are seen in clusters where there are radial "fingers" pointing to the origin where the gravity of the cluster causes internal motions and thus leads to a redshift induced by galactic motion that adds to the cosmological redshift. This figure illustrates that galaxies are tracing large scale structure in the Universe. The actual origin of the science is in the opposite order of what has been presented here: the large scale structure was first measured in galaxy redshift surveys and then observations like these were used to motivate the need for the ΛCDM paradigm.

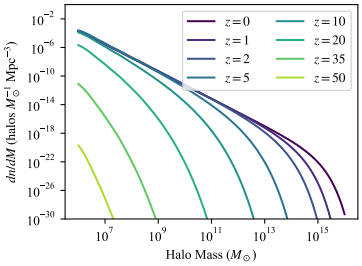


Figure 7.7: Mass distribution of halos as a function of redshift. Densities are expressed in comoving volume.

Figure 7.7 shows the mass distribution of halos with redshift. These functional forms are both predicted well by the analytical theory and measured in the dark matter only simulations. The mass distributions show three notable features. First, the abundance of halo masses grows over time (i.e., as $z \to 0$) representing the growth of structure seen in Figure 7.5. Halo growth is a "rich get richer" physical phenomenon where the largest mass halos have more gravitational attraction and can thus more matter in with higher effectiveness. The low mass end of the halo mass function follows a power law relationship with an index of $dn/dM \propto M^{-1.5}$ which is comparable to but slightly steeper than the field galaxy luminosity function. Finally, the halo mass function shows a turnover at the high mass end, just like the luminosity function shows a turnover at L_*. This characteristic feature is the propagation of the scales in the CMB fluctuation forward in time under the action of self-gravity. Thus, there is a characteristic mass scale in the Universe and it manifest in the halo mass function, and potentially also in the galaxy luminosity function.

Relevant for galaxy formation, halo structures are not spherically symmetric and thus they exert gravitational forces and torques on each other. These torques establish an angular momentum distribution for the halos in terms of the dimensionless spin parameter for halos

$$\lambda_J = \frac{JE^{1/2}}{GM^{5/2}} \sim \sqrt{\frac{E_{\rm rot}}{E_{\rm grav}}} \tag{7.5}$$

where J is the angular momentum, E is the total energy of the system and M is the mass of the system. The latter approximation is

approximate to factors of order unity. However, the dimensionality scaling is there to argue that small values of λ_J correspond to relatively weak rotation. Indeed, simulations of halos show that the spin parameter has a typical value of $\lambda_J = 0.037$ [Bullock et al., 2001] with a distribution function. The spin parameter indicates that there is a slight spin to halos but that the relatively importance of the angular momentum is small compared to the overall influence of gravity. The halo angular momentum the dictates gas angular momentum as the material falls down the halos into bottom of the potential well. Like all conservation of angular momentum problems, a small shear or rotation at large scales becomes quite significant in terms of orbital speeds once pulled down to small scales.

STAR FORMING GALAXY FORMATION – The dark matter halos set the initial conditions for galaxy formation and subsequent evolution. While we have focused on halos above, the Big Bang also sets the chemical composition of the Universe through Big Bang nucleosynthesis (BBNS) created the initial chemical composition of the Universe, in terms of mass fractions $X = 0.75, Y = 0.25, Z = 0$. This material, in the form of a gas, becomes the material that makes up the subsequent generations of stars and galaxies.

For now, we pick up the story qualitatively, having the dark matter collapse into halos and the gas cooling and falling into the potential wells created by those halos. After recombination at redshift $z \sim 1100$, this gas is neutral and relatively cool. Since gas is collisional (as opposed to stars or the dark matter), the gas will settle into rotationally supported disks given these initial input of angular momentum from the halo. Gas fluid elements with random motions radially or vertically will be damped out leaving thin disk with gas orbiting on circular orbits. It is in this circularly rotating disk that stars are able to form. These stars are forming from primordial material, which lacks any metals, leading to some important consequences for the first generation of stars. This rough model leads to a good explanation for understanding blue cloud / star forming main sequence galaxies. We can study this process in more detail by returning to the physics of gas that we studied in Chapter 4.

Key Points

- Galaxy evolution models require dark matter that is cold, non-collisional, and non-interacting.
- Cosmology assumes that the Universe is isotropic and homogeneous.
- The scale factor of the Universe, $a(t) = 1/(1+z)$, and is predicted by the contents of the Universe.
- Our Universe has good agreement between the Hubble time ($t_{\rm H} = 13.96$ Gyr and the actual age $t_0 = 13.47$ Gyr predicted by the full combination of physics.
- The Cosmic Microwave Background shows the Universe was once hot and dense. Its current temperature gives a good prediction for the redshift of recombination at $z = 1100$.
- The CMB shows fluctuations with a specific pattern of angular scales that show the initial scales and amplitudes for density fluctuations in the Universe.
- These fluctuations can set the initial conditions for dark matter density distribution, which then collapses under gravity through a process of hierarchical merging into progressively larger dark matter halos halos.
- Galaxy redshift surveys trace the large scale structure of the Universe showing it follows a filamentary network expected for dark matter collapse in the ΛCDM model.
- Dark matter halos have a mass spectrum is roughly a power law with a cutoff, like the Schechter luminosity function for a galaxy. The mass function means there are a few high mass halos and many low mass halos.
- The mass of all halos increase over time as dark matter falls into them. However, the high mass halos increase in mass relatively more.

7.2 *A Brief Return to Gas Physics*

The first thing to note is that galaxy evolution in the early universe is materially different from later times because of the ionization state of the universe. The first galactic structures are thought to form from

neutral gas, but this gas is quickly ionized by the first generation of stars, called *cosmic reionization*. This is though to occur between $z \sim 10$ and 20 and is a major science target for both long wavelength radio astronomy and the James Webb Space Telescope that launched in December 2021 and is currently undergoing commissioning. We should have much better ideas about this next year.

Note to self: add field-altering discoveries from JWST here for ASTRO 322 in Wi23.

It may seem amazing that a relatively small number of first generation stars could ionize literally Every Hydrogen Atom in the Universe, but the density of gas is quite low because it has not all fallen into dark matter halos. Moreover, the first generation of stars is unique since they will have $Z = 0$. Such stars lack bound-free opacity in their interiors and are more luminous for a given mass (Figure 2.14). There is a constant hunt for the lowest-metallicity stars but to date, no star has been found with zero metallicity. We have found stars with very low metallicity ($Z = 10^{-5}$) but nothing with $Z = 0$ as we expect. One reason for this could be that, without metals, only high-mass, short-lived stars will form. This can be seen from Jeans Mass arguments combined with the ISM cooling (Figure 7.8). To cool gas to $T < 100$ K requires molecular line and atomic transitions from metals (e.g., the [C I] line or the CO molecule). Without these metals, the temperature of star forming gas must be higher and then the characteristic mass of the forming stars (Equation 3.11) will be higher so nearly all stars will be high mass and short lived. The high mass stars have huge ultraviolet radiation outputs which serve to re-ionize the Universe. Recall Equation 4.4 which gives the recombination rate of ionized plasma has the form

$$\mathcal{N}_{\rm rec} = \alpha(T) n_p n_e \tag{7.6}$$

where $\alpha(T)$ is a coefficient and $\mathcal{N}_{\rm rec}$ is the number density of recombinations per volume per unit time. This value depends on the number density of the protons and electrons ($n_{\rm H}^2$). For redshift $z = 20$, the mean density of hydrogen in the Universe is $n \approx 5 \times 10^2$ m^{-3}. Even at $T = 10^4$ K, the recombination time for this gas is quite long: 0.3 Gyr, providing plenty of time for more stars to form and evolve allowing the gas to be readily ionized by ongoing star formation.

Once the gas is ionized, it needs to cool and fall into galaxies before it can be made into more stars and build up into larger galaxies. Following Chapter 4, the rate of thermal energy loss from a volume of gas is set by the cooling curves follow the form:

$$\dot{u}_T = \Lambda n^2 \tag{7.7}$$

where n is the particle density. Unless that thermal energy loss is balanced by a heating source, this will also establish the cooling time

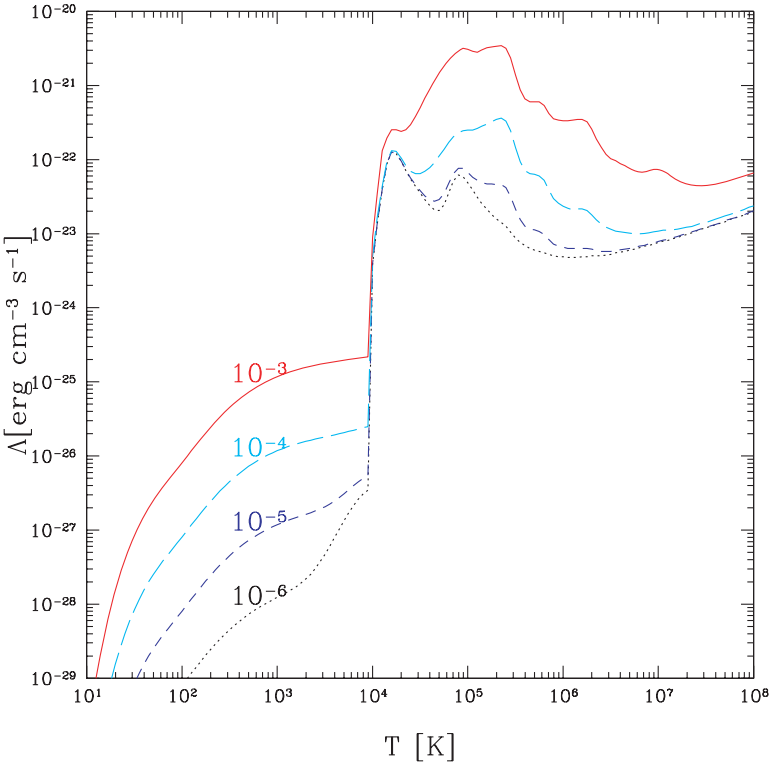


Figure 7.8: Cooling function for gas as a function of temperature. The different curves show the cooling function for different metallicities. Taken from Maio et al. [2007]. Note that the vertical axis units should be erg s^{-1} cm^3. Note that the typical state for hydrogen gas is ionized at $T > 10^4$ K and neutral for $T < 10^4$ K. Compare to the standard curves for the ISM in Figure 7.8.

Phase	Fraction
Galaxies (stars)	7%
Neutral Gas (atomic and molecular)	2%
Circumgalactic Gas (ionized)	5%
Intercluster Medium (ionized)	4%
Intergalactic Medium (ionized)	59%
Unknown (likely hot)	23%

Table 7.1: Approximate baryon budget at $z = 0$.

for gas:

$$\tau_{\rm cool} = \frac{3nkT}{2\dot{u}_T} = \frac{3kT}{2\Lambda n}. \tag{7.8}$$

The gas pressure $P/k = nT$ is roughly constant for gas in the ISM, meaning that hotter environments are also lower density. Given this, the cooling time is

$$\tau_{\rm cool} = \frac{3k^2T^2}{2P\Lambda}. \tag{7.9}$$

This leads to the conclusion that the cooling times for hot gas $T > 10^7$ K are quite long, but as $T \rightarrow 10^4$ K the cooling times will be short and the gas will quickly recombine into neutral and ultimately molecular gas. We usually study the ISM in a state of energy balance where the cooling functions in the atomic medium are balanced by heating mechanisms (stars). However, for the case of gas accreting onto galaxies, there are not strong heating mechanisms, so the gas will cool "unopposed" and the evolution timescale is set by this cooling.

The cooling function and the recombination rate are key because they establish the conditions in which galaxies are evolving at $z = 0$ and through the era when most of the mass in galaxies was built up. Stars form from the coldest molecular gas $T = 10$ K but the baryon reservoir in the universe is mostly hot and ionized. The build up of galaxies requires transforming that gas through the action of gravitation into stars. Observations of the cosmic microwave background and the trace elements that come from Big Bang Nucleosynthesis pin down the baryon content of the Universe on large scales. We [and by "we," I mean Shull et al., 2012] can then engage in the exercise of tallying up where that matter actually is found. The summary results of this work are shown in Table 7.1.

Perusing the table shows that most of the baryons in the $z = 0$ Universe are not in galaxies but are found in ionized gas. Most of the material we examined in this course is in the first two entries (Galaxies, Neutral Gas). Circumgalactic gas refers to material concentrated in individual galaxy halos. The intercluster medium refers to hot gas bound into the large potential of the galaxy cluster. The intergalactic medium is the material found between galaxy clusters and is distin-

guished from the other lines in the table by not being bound into a gravitational halo by dark matter.

Observationally, the hot material between galaxies is particularly difficult to observe. Recall that what we see of this material necessarily comes from light. Light is a cooling channel of the gas and when the cooling rates are so low (because the densities are low), then this material gives off very little light.

Key Points

- Gas is neutral after recombination with $Z = 0$, but the first generation of stars reionizes the Universe and seeds the cosmos with metals.
- The reionized gas remains ionized and transparent because it has very low densities so its recombiantion times are long compared to the lifespans of stars, allowing for continuous reionization.
- The cooling times for hot ionized gas are also very long because the densities of the intergalactic medium is very low.
- Most of the baryons in the $z = 0$ Universe are in the hot ionized gas between galaxy clusters.

7.3 *Building Galaxies*

Table 7.1 shows how relatively trivial galaxies are and also shows that the build up of material in galaxies requires moving material from the hot ionized phase into the cold neutral medium, which is the only gas from which new stars can form. Apart from our self-centred nature, we continue to care about galaxies since stars and AGN are what actually heat the Universe up and maintain its ionized state. We can now turn to how those galaxies are formed in more detail.

We have already briefly summarized how we build up star forming galaxies. Gas accretes into dark matter halos. The angular momentum of the individual halos gives a spin to the gas, which leads to a rotating flattened disk. In that disk of gas, stars form.

We can study the star formation rate of galaxies as a whole as a function of their parameters. When explored on the large scale, we find a relationship between the stellar mass of a galaxy ($M_\star$) and its ongoing star formation rate ($\dot{M}_\star$). Figure 7.9 shows this relationship directly. We see that the star forming main sequence is given empiri-

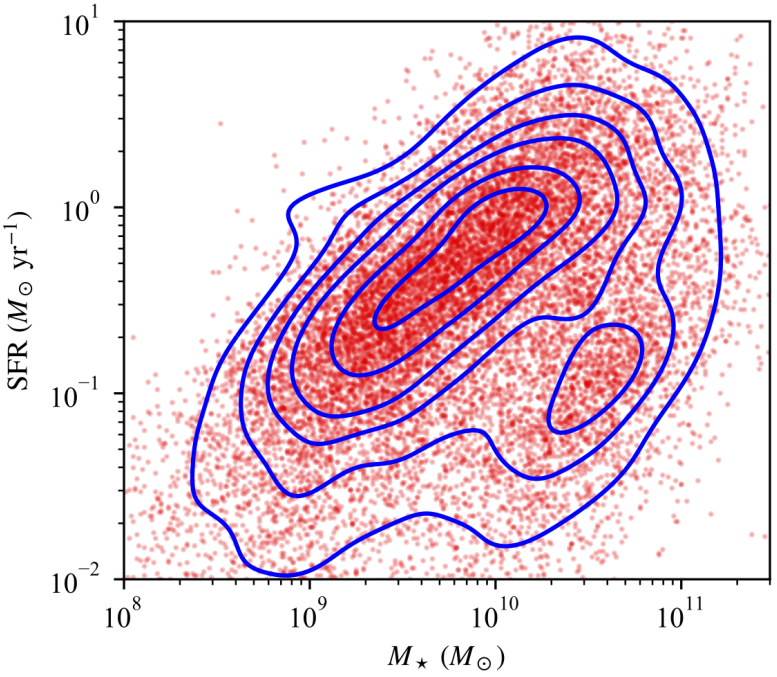


Figure 7.9: The star forming main sequence of galaxies as established in the zomgs [Leroy et al., 2019]. Galaxies separate into two distinct populations with respect to their star formation with respect to their stellar mass. This is a complementary perspective to the colour-magnitude diagram of galaxies (Figure 5.6).

cally as:

$$\frac{\dot{M}_\star}{M_\odot \text{ yr}^{-1}} = 0.67 \left(\frac{M_\star}{10^{10}\ M_\odot} \right)^{0.68} \tag{7.10}$$

- The star forming main sequence of galaxies has a declining *specific star formation rate* SSFR $\equiv \dot{M}_\star / M_\star$. High mass galaxies on the star forming main sequence are forming fewer stars per unit stellar mass than low mass galaxies.
- The non-star forming galaxies have a small but persistent level of star formation. These would correspond to the red sequence galaxies for which star formation could be detected.
- There is a large gap in SFR between the star forming sequence and the non-star forming sequence, nearly two orders of magnitude in SSFR.

Galaxies with $M_\star / \dot{M}_\star > 10^{10}$ yr had to come from higher star formation rates in the past, otherwise they couldn't achieve their current stellar mass at their observed star formation rate. This phenomenon is described as *downsizing*, meaning that the most massive galaxies were assembled earlier in the Universe.

The Star Formation History of the Universe – The star form-

ing main sequence picture and the consistent with the results seen in Figure 7.10. This figure shows the star formation rate per unit volume inferred from wide-area surveys of galaxies at different redshifts and then heavily corrected for selection effects and observational bias. After thinking about these hard issues very carefully, this census of observations shows roughly when the stellar mass of galaxies was assembled. We are well past "peak star formation" in the Universe at $z \sim 2$ when the biggest galaxies were being assembled. Big galaxies are found in big dark matter halos that collapsed early in the Universe.

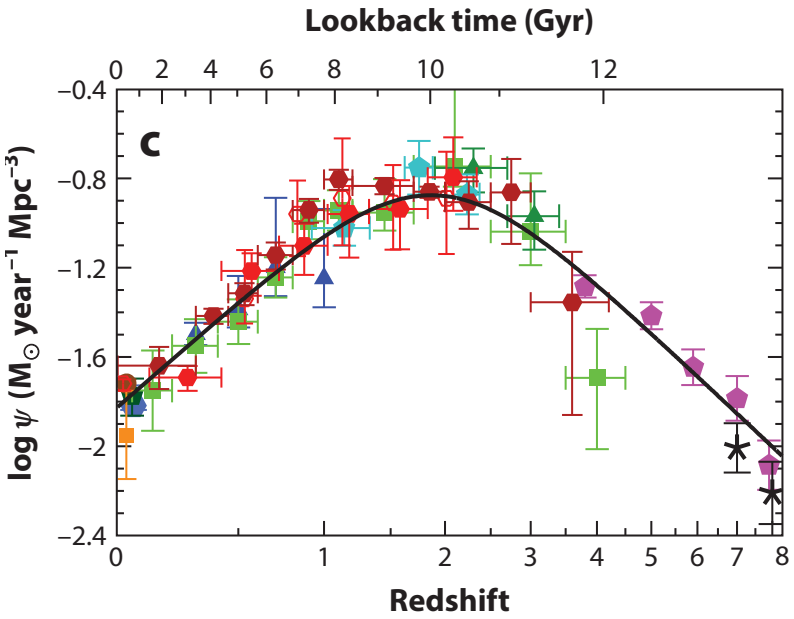


Figure 7.10: Star formation history of the Universe. The figure plots the star formation rate per unit volume inferred globally for the Universe. The star formation rate per volume was significantly higher in the past, peaking near $z \sim 2$ (10 Gyr ago). Figure from Madau and Dickinson [2014].

For context on the scale of the vertical axis, consider that the Local Group of galaxies has a star formation rate of about 5 $M_\odot$ yr^{-1} at spans a volume of $\sim$ 1 Mpc^3 placing our region $>$ 1.5 orders of magnitude up from the $z = 0$ data on the vertical scale. However, the Local Group is overdense by about the same factor with respect to the Universe as a whole, so this accounts for the discrepancy.

More recent work has shown that the star forming main sequence largely persists with the higher star formation history of the universe in the past. The sense of this is that the star formation rates of systems are globally higher for a given stellar mass of galaxies rather than the star formation happening in higher mass galaxies. The slope of the star forming main sequence also appears to steepen slightly in

the past.

The star formation main sequence and the cosmic star formation history support our basic picture how galaxies form and evolve in the view of the ΛCDM cosmology. The colour-magnitude diagram and the star formation main sequence both aim to discuss three main parts of the galaxy population: the actively star forming galaxies (blue cloud), the non-star forming galaxies (red sequence) and the transitional quenching systems (green valley) galaxies that lie between these two. Our picture, however, really only explains what is happening star forming galaxies. The quenching of galaxies and even the buildup of bulges inside galaxies like our own remain mysterious.

GALAXY-SCALE STAR FORMATION – The study of how stars form from gas is interesting in detail, which is why I sink a lot of my research life into it. However, the broadest estimates of how gas forms into stars are given by *star formation laws* which parameterize the star formation rate in terms of local variables $\dot{M}_\star = f(\rho, T, Z, \cdots)$, where interesting local variables may be the mass density of the gas (ρ), the temperature of the gas (T), the metallicity (Z) and many others. As noted in Chapter 3, the dominant model in the field is converging to a constant efficiency of star formation per free-fall time: $\dot{M}_\star = \epsilon_{\rm ff} M_{\rm cloud}/t_{\rm ff}$. Empirically, we view star formation through measurements of the surface density of gas and stars and find that

$$\dot{\Sigma}_\star \propto \Sigma_{\rm gas}^n \tag{7.11}$$

where $n \in [1, 2]$ depending on the exact nature of the measurement being made. The best established star formation laws follow $\dot{\Sigma}_\star \propto \Sigma_{\rm mol}^{1.0}$, meaning that once gas gets into the molecular phase it forms stars at a rate proportional to the mass reservoir of the gas. The details of *why* such a law emerges are a topic of current investigation since, like the studies of the IMF, we are interested in regions where the relationship breaks down. There is some evidence that the star formation law takes on different forms at high redshift or low metallicity.

Kennicutt and Evans [2012] present a good review of such star formation laws and the observational methods that lead to their development.

Like the IMF, the mean relationship of the star formation law appears surprisingly robust and gives a simple prescription of how galaxies build up their mass into stars: higher densities of gas into galaxies translates into more stars. This relationship holds over large scales in galaxies and can be thought of as an ensemble average of the properties of gas. On smaller scales, the relationship breaks down a bit and appears to depend sensitively on local conditions.

A key point is the constant of proportionality in these relationships. The molecular gas depletion time is defined as $\tau_{\rm dep} = \Sigma_{\rm gas}/\dot{\Sigma}_\star$

and represents the time it would take for a galaxy to deplete its reservoir of gas at its current star formation rate. For an index on the star formation law of $n = 1.0$, the depletion time is constant and has a value of 2 Gyr over most galaxies on the star forming main sequence. This depletion time can be compared to two other timescales.

- In comparing to the assembly timescale implied by the specific star formation rate, the depletion times are comparable to the assembly times for low mass systems on the star forming main sequence, but are significantly shorter for the high mass systems on that relationship. This means that galaxies will deplete their gas reservoirs into stars several times over the course of their evolution and this reservoir will need to be refilled. This inflow is thought to come from cooling gas of those hot baryons in the circumgalactic and intergalactic medium.
- The gas depletion times are long compared to the free fall times of the molecular gas in which the stars are formed ($\tau_{\rm ff} \sim 3$ Myr). Since stars form by the gravitational collapse of the gas, this means that the star formation process is globally inefficient. Only $\sim 1\%$ of the molecular gas mass forms into stars every free-fall time. The action of turbulence and stellar feedback is thought to make star formation inefficient, which is tantamount to saying the $\epsilon_{\rm ff} \ll 1$.

The star formation law is local and holds within galaxies. The net action of a star formation event is to convert a few percent of the gas mass into stars at a given location and disperse the rest of the gas into the interstellar medium. In addition to having the output of the star formation process as the IMF and the binary process, the star formation process also establishes the fraction of stars found in clusters and small groups of stars. Clustering arises because most star formation occurs in the densest parts of the molecular gas, which have a small filling fraction. Invoking the deep poetry of astrophysics, such dense regions are called "clumps" and host the formation of clusters and small groups of stars. This is just a byproduct of gravitation: high density regions tend to accrete more matter, making denser more massive regions and the process accelerates until the act of star formation dissipates the clump.

Many of these clusters formed in clusters will be gravitationally bound and persist for long times to be seen as open clusters in the sky. Others will be unbound and dissipate quickly into the broader stellar milieu. The fraction of stars formed in clusters appears to be significant and depends on local properties. Averaged over galaxies, this is typically 30%.

In addition to actually making the stellar mass of galaxies, the star formation process also sets the initial motions of stars. As soon as

stars form, they transition from a collisional to a collisionless fluid and are free from the effects of feedback, gas motions, and the radiation of other stars. They are cast loose into the broader galactic potential to start orbiting the system and are subject to the formalism developed around galaxy dynamics. Stars thus form with the motions of the molecular gas disk, namely on circular rotation with relatively small vertical velocity dispersions, and then stellar dynamics processes the stars.

It seems like we have all the ingredients we need to explain galaxies at this point: dark matter accretes gas, it cools into a rotating disk, which then forms stars inefficiently leading to a galaxy in the blue cloud on the star forming main sequence. However, there are a few wrinkles in the picture that keep this picture from being complete with just this.

Key Points

- Star forming galaxies mostly form on the star forming main sequence which relates the current stellar mass to the star formation rate (Equation 7.10.
- Low mass galaxies are forming stars at a faster rate relative to their stellar mass than high mass galaxies.
- The star formation rate of the Universe was $10\times$ higher at $z = 2$ than it is now. At earlier times, the star formation rate was lower.
- Galaxies form stars from their molecular gas reservoirs with a depletion time of approximately 2 Gyr. Thus, gas reservoirs must be replenished to sustain star formation over the duration of the Universe.

7.4 *Feedback*

Our ΛCDM model for halo formation runs into issues that need explaining on both ends of the mass distribution of galaxies.

Figure 7.11 shows the distribution of halo masses expected from our simulations in the $z = 0$ Universe (see also 7.7) compared to the masses of galaxies. There is clearly a changing ratio of halo mass compared to galaxy mass for different mass scales of structures. We need to explain why there are relatively few galaxies at both the low mass end and the high mass ends of the galaxy population. From a naïve model, there are too few dwarf galaxies and too few giant galaxies compared to the mass distributions of halos. Galaxies with

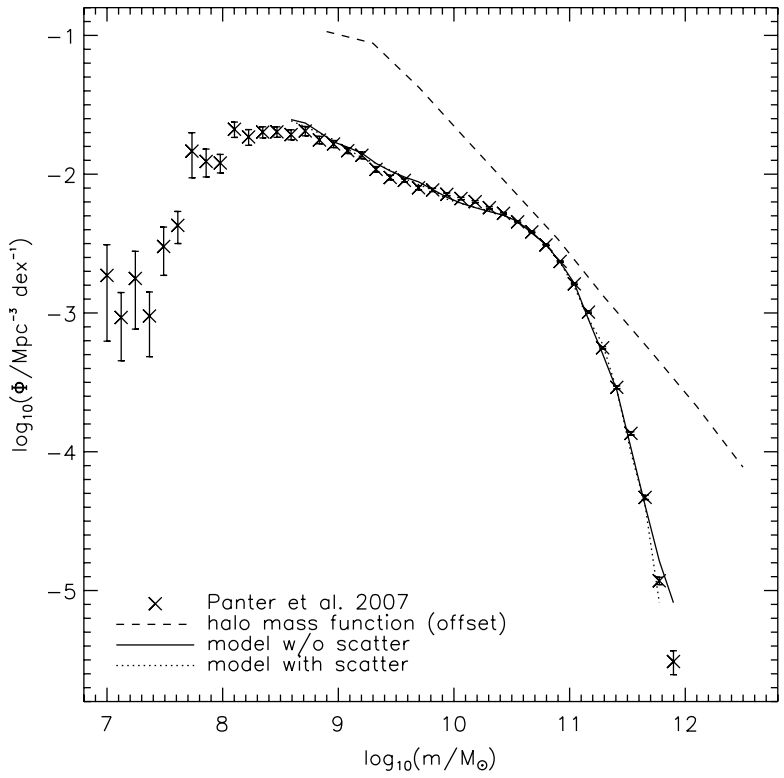


Figure 7.11: Comparison of the halo mass distribution (dashed line) to the observed galaxy mass function at $z = 0$ from [Moster et al., 2010].

$M_\star \approx 5 \times 10^{10}\ M_\odot$ which corresponds roughly to the L_* value in the local galaxy luminosity function (Equation 5.8). There should be a huge number of low mass dark matter halos but there are fewer dwarf galaxies than we would expect if every halo had a dwarf galaxy inside it. Where then are these missing dwarf galaxies, or equivalently, why are dwarf galaxies relatively inefficient at forming stars over a long time. Similarly, some process has set a maximum mass for galaxy at $\approx 10^{12}\ M_\odot$ even though the highest mass halos in the universe are almost $1000\times$ larger.

The answer to both of these questions appears to be the process of *feedback*. For the low-mass galaxies, the feedback is thought to be driven by the process of star formation. The energy and momentum injection from supernova explosions (see Chapter 4) and stellar winds is sufficiently strong to disrupt the star forming interstellar medium and drive it completely out of the dark matter halo. As such, material has a much lower efficiency at forming stars over the long term since the star forming gas reservoir is disrupted.

At the high mass end, the discrepancy is also from feedback, but this time from black hole accretion onto active galactic nuclei.

7.4.1 *Active Galactic Nuclei*

The evidence appears to be leaning toward every galaxy having at least one supermassive black hole in its centre. When this supermassive black hole accretes, it shows up as an active galactic nucleus (AGN). The light from accretion onto this black hole can be a significant, even dominant, part of the emitted light from a galaxy. There are many different types of AGN, mostly driven by the usual problem of not realizing these are the same type of phenomenon when originally observed. Figure 7.12 shows the spectra of these AGN classes.

There is an absurd number of AGN classifications and Figure 7.12 only summarize the major classifications as seen in the optical. When emission from the high and low frequency portions of the spectrum is included in the classification, the number balloons to $\sim$ 50 as carefully tabulated in Padovani et al. [2017] ranging from the historical "quasar" to the silly XBONGs (X-ray Bright, Optically Normal Galaxy). In the current context, we mostly care because AGN can emit a huge amount of radiation across the electromagnetic spectrum.

The radiation from the accretion process depends strongly depends on where the energy being observed is coming from. Figure 7.13 shows the standard model of the structure in an AGN engine. The figure shows the central back hole with a thin accretion disk, a

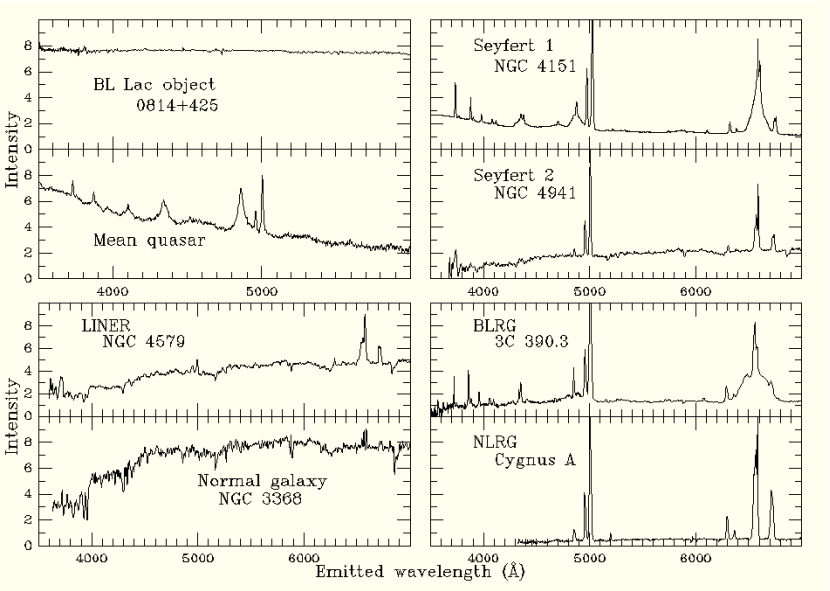


Figure 7.12: Spectra of the different classes of Active Galactic Nuclei. From Bill Keel's AGN teaching resources https://pages.astronomy.ua.edu/keel/agn/.

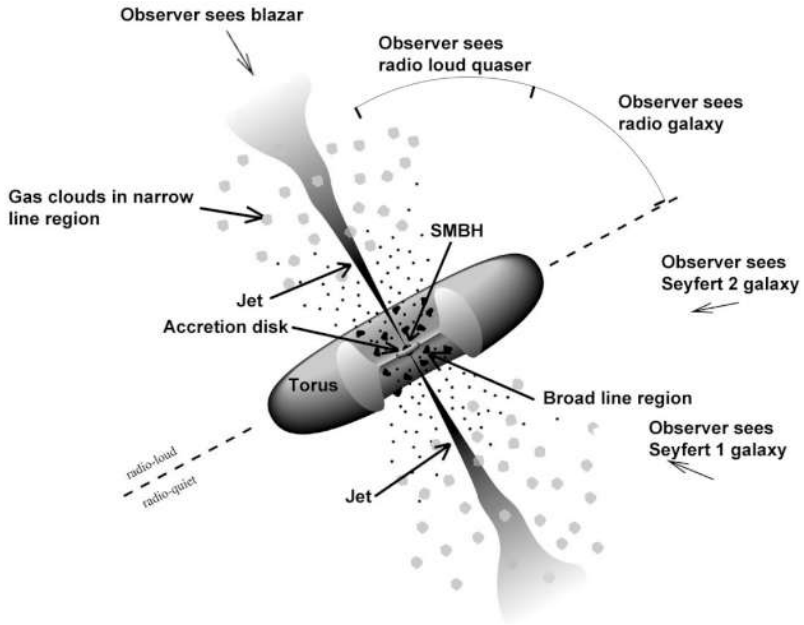


Figure 7.13: Schematic diagram of the standard AGN anatomy showing the inclination based unification scheme. Taken from the NASA Fermi website: https://fermi.gsfc.nasa.gov/science/eteu/agn/.

jet of high energy particles, and a gas torus surrounding the accerection disk. In general (though with many exceptions):

- Radio emission is usually associated with synchrotron emission in relativistic particles in the relativistic jets.
- Infrared emission comes from the dusty torus and large size scales in the nuclear region and is commonly associated with star formation occurring in the dense gas surrounding the black hole.
- Optical and ultraviolet emission is usually from the accretion disk.
- X-ray emission is from a *coronoa* of high temperature plasma around the nuclear region driven as powered as a byproduct of the accretion process.
- Gamma rays are associated with jet emission.

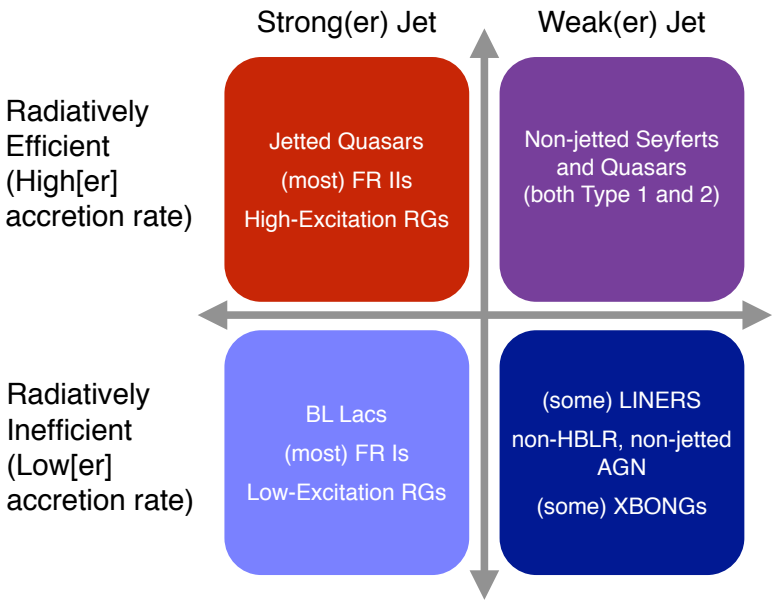


Figure 7.14: Four-quadrant classification of intrinsic AGN properties. Figure from Padovani et al. [2017] after scheme presented by Phil Hopkins. The sundry acronyms presented in this figure refer to the different classifications of AGN and aren't important at this point.

The consensus view has driven the perspective on these systems toward a synthesis model where the variations in the classification are driven by the inclination angle with respect to the angular momentum vector of the black hole. Unfortunately, the inclination-based classification can only explain part of the span of properties seen in the whole of the AGN population. Even controlling based on inclination angle, not all AGN appear to be the same which has brought up additional axes for variability as laid out in Figure 7.14. This scheme recognizes that the other two axes from AGN properties that need to

be considered are whether the AGN is radiatively efficient or not and whether the jet is strong or not. When an AGN is described in terms of its radiative efficiency, that refers to how much light is emitted by the AGN (i.e., its luminosity L) based on the amount of mass that is accreted ($\dot{M}$):

$$\epsilon = \frac{L}{\dot{M}c^2}. \tag{7.12}$$

ACCRETION POWER AND THE EDDINGTON LUMINOSITY – Here, we need to understand a bit more about the nature of black hole accretion. While a 10^7 $M_\odot$ may seem terrifyingly huge, on galactic scales it is relatively small. One consequence of our study of dynamics is that it can be challenging to force close approaches between objects on galactic scales. Actually accreting into a black hole requires the dissipation of vast amounts of angular momentum to get the radius of the orbit small enough that the material will fall through the event horizon. The *accretion disk* mentioned previously is the engine for dissipating this angular momentum, transferring the angular momentum outward while moving mass inward into the central object (accretion disks show up around protostars, neutron stars, white dwarfs as well as black holes). The material is disk-shaped for the same reason that the gas in galaxies is in a disk: the motions of the material are dominated by ordered motions carrying a net angular momentum. Random motions vertically with respect to the disk are rapidly dissipated by viscosity and non-circular orbits also lead to friction and dissipation of the motion through viscosity.

Unlike galaxies, the motions in a black hole accretion disk are dominated by the central black hole for which the circular velocity curve follow a Keplerian rotation curve:

$$V_c = \sqrt{\frac{GM_\bullet}{r}} \tag{7.13}$$

but in this case the mass is just a constant. For a gas parcel of mass m the total energy at any given radius is thus:

$$E_{\rm tot} = U_g + K = -\frac{GM_\bullet m}{r} + \frac{1}{2}mV_c^2 = -\frac{GM_\bullet m}{2r} \tag{7.14}$$

This gas follows the virial theorem where $\langle K \rangle = -\langle U_g \rangle$. As a gas particle moves inward ($r \to 0$), the gas particle loses energy and half of the gravitational potential energy liberated goes into the kinetic energy of the gas (it orbits around black hole faster). The other half must be removed from the gas somehow, typical through viscosity heating the gas disk and causing it to radiate. Similarly the angular momentum of the gas must also decrease as $r \to 0$:

$$J = mV_c r = m\sqrt{\frac{GM_\bullet}{r}}r = \sqrt{Gm^2 M_\bullet r} \tag{7.15}$$

This angular momentum is transferred outward to other gas in the outer disk as well as by twisting up the magnetic field threading the disk. This magnetic field twisting results in the jet seen from accreting systems and the launching material also carries away energy and angular momentum.

Because the radius of the black hole is small, accretion can produce vast amount of energy. For a gas parcel of mass m falling from $r = \infty$ down to the Schwarzschild radius of the black, the total energy released is:

$$E_{\rm acc} = \frac{GM_\bullet m}{2R_\bullet} = \frac{GM_\bullet m}{2(2GM_\bullet/c^2)} = \frac{mc^2}{4} \tag{7.16}$$

or a full 25% of the mass energy of the original matter. For reference, nuclear fusion of hydrogen to helium releases 0.7% of the mass energy of the material (Figure 2.3). This stage of fusion is the most efficient form of mass energy conversion available through fusion and accretion is 36× more efficient than it. Only matter-antimatter energy generation is more efficient in terms of mass energy.

It thus seems that the accretion of enough material could easily produce nearly infinite amounts of energy. However, there is a practical limit to the amount of energy that can be produced by an object before the radiation that is being created disrupts the flow of material onto the accreting object. We can find the luminosity by comparing the force of gravitational attraction required to bind the material to the object compared to the force that the radiation is pressing on the material. Here, we will consider a hydrogen ion as our test material. The electrostatic force keeps a plasma electrically neutral since it is far stronger than the other forces involved. The gravitational force acts primarily on the proton in the ion with a magnitude of force $F_g = GMm_{\rm H}/r^2$ in the radially inward direction. The radiation pressure primarily acts on the electron in the pair pushing it radially outward with a magnitude $F_{\rm rad} = f\sigma_T/c$ where f is the radiative flux of light and σ_T is the Thomson cross section of light scattering off electrons. The measured value $\sigma_T = 6.65 \times 10^{-29}$ m^2 (sorry, another variable named σ). Equating these two yields:

$$F_g = F_{\rm rad} \tag{7.17}$$

$$\frac{GM_\bullet m_{\rm H}}{r^2} = \frac{L}{4\pi r^2}\frac{\sigma_T}{c} \tag{7.18}$$

$$L = \frac{4\pi GM_\bullet m_{\rm H} c}{\sigma_T} \tag{7.19}$$

$$L_{\rm Edd} = 3.29 \times 10^4\ L_\odot \left(\frac{M_\bullet}{M_\odot}\right). \tag{7.20}$$

In the last equation, we have added the subscript "Edd", since this

luminosity is given the name *Eddington luminosity*. This is the maximum luminosity at which a black hole could accrete ionized material.

ACCRETION EFFICIENCY – Taking a step back, the act of accreting into a black hole turns gravitational potential energy into kinetic energy. In the case of a standard optically thick accretion disk, the kinetic energy of the gas can be coupled out into the radiation field effectively though radiative emission. However, if the emission is optically thin and the gas diffuse, the two body interactions required to produce electromagnetic radiation in a plasma are rare and the gas can be accreted into the black hole carrying its kinetic energy with it. This latter kind of accretion flow is called and *advection dominated accretion flow* (ADAF). Advection refers to how fluid flows move gas around and the associated properties of the gas are also moved (the energy, momentum, etc). Thus, if a fluid element in a flow is carried to a new location all the mass, momentum, energy, entropy of that fluid is also carried with the fluid element.

The exact division between all these different flow classifications aren't deeply important, but you should appreciate that radiatively efficient accretion can produce a great deal of electromagnetic radiation from accretion, but not all super-massive black holes are accreting with high radiative efficiency. Because the efficiency is related to the density of the accreting material, the radiative efficiency is also thought to be related to the accretion rate. Systems with high accretion rates are typically more efficient radiators. We also typically think of higher accretion rates leading to stronger jets, but these are not necessarily well coupled for supermassive black holes in AGN because the jet launching mechanism also varies in efficiency (for example lower magnetic fields).

Thus, the observed properties of AGN are not just a function of the inclination with which the central engine is being viewed but also the accretion rate directly into the central region and how effectively that black hole launches a jet. This two-axis classification scheme has, at its heart, the idea that AGN vary in their properties over long time periods depending of how they are accreting from their broader environment. This means that they are coupled to their broader nuclear environment and the amount of the material in the nuclear region of the galaxy. It is critical to keep in mind the relative scales of the regions involved. The Schwarzschild radius of the black hole is

$$R_{\rm Sch} = \frac{2GM_\bullet}{c^2} = 10^{-7}\ {\rm pc}\left(\frac{M_\bullet}{10^6\ M_\odot}\right) \tag{7.21}$$

Accretion disks are typically 10^{-4} pc to 10^{-2} pc. The nuclear ring or torus around the central engine has inner scales connecting the

accretion disk to the broader nuclear gas reservoir and has a range of 10^{-1} pc to $10^{2.5}$ pc. The boundary between accretion disk and the broader fuel reservoir (the "torus") is largely driven by where the orbital motion of the material is governed by the black hole and is thus Keplerian and a thin disk and where it is governed by the broader galactic potential. Much of the current research in black hole accretion asks the questions of how to channel this nuclear gas reservoir into the accretion disk, after which the behaviour of the gas is (marginally) better understood.

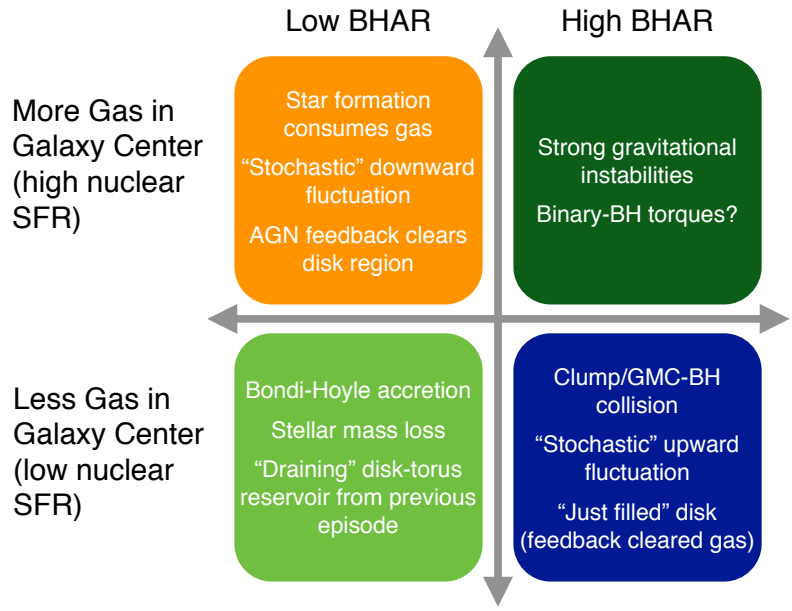


Figure 7.15: Two parameters describing the broader AGN environment, again from Padovani et al. [2017].

The broader connection to the galaxy environment raises the question of how to get the material into the central region. Figure 7.15 showcases the differences between these two scales. The nuclear gas content indicates how much gas could be accessed by the black hole over galaxy dynamical times (10^8 yr). In general, high gas content tends to leads to high star formation rate, but could also dump a lot of that gas into the nuclear engine leading to large accretion rates and AGN activity. The competition between consumption and expulsion of the gas from star formation and its associated feedback and the rate at which gas moves into the nuclear engine leads to the decoupling between nuclear gas content and AGN activity. For context, nuclear regions (central kpc) can contain a large fraction of the entire gas budget in galaxies (a gas rich nucleus can contain $\sim$ 10% but this fraction can be higher).

Similarly, low overall gas content in the nuclear region can still see

high accretion rates if a single cloud passes near the nuclear region because accretion efficiencies can be quite large. A relatively small $10^5\ M_\odot$ GMC cloud accreting per 100 Myr accreting with modest ($\epsilon = 10^{-2}$) radiative efficiency can produce a hefty $10^8\ L_\odot$.

In summary, AGN are a key component of the light emitted by galaxies. The actual radiation emitted depends on a wealth of factors that all describe the amount properties of gas in the nuclear regions of galaxies. The exact processes by which gas is moved from 10^2 pc scales to actually accrete onto the black hole starting from scales of 10^{-3} remain poorly understood and a variety of different physical mechanisms are conjectured. It is further thought that many of these mechanisms are in play, with different observational consequences.

Key Points

- Accretion onto supermassive black holes at the centres of galaxies drives *active galactic nuclei*, which can emit at high luminosity across the EM spectrum. An actively accreting AGN can dominate the light of a galaxy.
- There are a vast number of AGN classifications. The details of these classifications are *not* important for this class, but they are determined by:
  - The inclination angle between the AGN jet and the line of sight.
  - The rate of accretion onto the black hole.
  - Whether the accretion is radiatively efficient (high density, so cools effectively) or inefficient (low density, so ineffective cooling).
- Accretion can release up to 25% of the mass energy of material falling into a black hole.
- The maximum accretion rates are set by the Eddington luminosity (Equation 7.20). If an AGN is emitting above the Eddington limit, the radiation will disrupt the accretion onto the black hole.
- One of the biggest unsolved questions in AGN feeding is how to move material from moderate scales near the black hole (300 pc) down to the scale of the accretion disk (0.1 pc).

7.4.2 *Black Holes and Stellar Bulges*

A major clue about the formation and evolution of spheroidal stellar populations comes from the observed relationship between the mass of the central supermassive black holes ($M_\bullet$) in galaxies and the velocity dispersion of the hosting bulge / spheroidal system (σ_v). Specifically, the observed relationship is that $M_\bullet \propto \sigma_v^4$. In disk galaxies with significant bulges and elliptical galaxies, it appears that $M_\bullet = M_{\rm bulge}/200$. These relationships are surprising because of the vast difference in scales between the two systems: compared to the host system the black hole is $\sim$ 3 orders of magnitude smaller in mass and $\sim$ 8 orders of magnitude smaller in size. What physical process drives this correlation?

The latest answer to this question appears to be feedback from accretion onto the supermassive black hole.

The rough scaling can be established by assuming the radiation pressure from a black hole accreting at the Eddington limit sweeps the gas out to the edge of the galaxy and then balances the radiation pressure with gravity at that point. If the gas mass $M_{\rm gas}$ is a fraction f of the mass of the galaxy $M_{\rm gal}$,

$$P_{\rm rad} = \frac{L_{Edd}}{c} = \frac{4\pi G M_\bullet m_p}{\sigma_{\rm T}} = \frac{G M_{\rm gal} M_{\rm gas}}{r^2} \tag{7.22}$$

$$= \frac{f G M_{\rm gal^2}}{r^2} \tag{7.23}$$

$$= \frac{fG}{r^2}\left(\frac{3\sigma^2 r}{\beta G}\right)^2, \tag{7.24}$$

where the final equality holds by applying the virial theorem leading to the conclusion that

$$\frac{4\pi G M_\bullet m_p}{\sigma_{\rm T}} = \frac{9 f \sigma_v^4}{\beta^2 G} \text{ or} \tag{7.25}$$

$$M_\bullet \propto \sigma_v^4. \tag{7.26}$$

This explains part of the observed correlation between the black hole and the velocity dispersion of the region, but it does little to explain the roughly constant mass ratio between spheroid mass and black hole mass. For this, the influence of dust is typically invoked. Since $\sigma_{\rm dust}/\sigma_{\rm T} \sim 10^3$, the relationship is thought to be established by black hole accretion being locally (i.e., right around the black hole) at the Eddington limit for electron scattering since the material is more transparent, but at larger radii, the Eddington limit for dust is relevant since dust grains can exist. This ratio in cross sections thus sets the ratio in the maximum amount of mass that can be retained before the black hole dissipates the material. Given these results, the galaxy

properties are thought to be strongly regulated by feedback, and this relationship is likely to establish the link between star formation history and dynamical processes.

Key Points

- The stellar velocity dispersion of the stellar bulge and the mass of the central black hole are correlated, implying a physical connection between these two.
- The mass of the central black hole is $\approx$ 1/200 the mass of the stellar bulge that hosts it.
- This correlation is attributed to AGN feedback disrupting accretion onto the black hole. More massive galaxies can resist feedback more, allowing the black hole to grow in response to the growing stellar bulge.

7.5 Galaxy Collisions

While simple model of halo-hosted galaxy formation explains why galaxies settle into disks it is less clear on how to produce red sequence or quenched galaxies. The orbits of stars in red sequence galaxies have significantly higher random motions compared to their circular velocities. The key point about the evolution of a stellar system from our exploration of dynamics is that it will randomize itself on timescales like the relaxation time, which are extremely long for galaxies. Thus, the orbits of stars in ellipticals cannot reflect a disk galaxy that has randomized itself. Similarly, the orbits are unlikely to be primordial, since the gas that formed those stars would have had to be on random orbits where it would have collided with itself and settled into a disk.

The dominant hypothesis is that the red sequence of galaxies reflects the product of merging halos and collisions. When two dark matter halos collide, the large scale gravitational potential will dynamically "heat" the stellar systems inside, randomizing their orbits. The effects of stellar feedback are thought to heat the gas and drive it out to low densities where it will have long cooling times and low ionization rates. Without cool gas, there will be no ongoing star formation and the stellar population will naturally age to the red sequence.

Galaxy collisions provide a quick way to randomize angular momentum of orbits into essentially spheroidal orbits. When galaxies collide, the net angular momentum of the system would be the vector sum of the angular momentum of the two galaxies. This will

add stochastically, so that the net specific angular momentum will be lower than it would be compared to relatively undisturbed disk galaxies. Collisions also provide a means of disrupting the disk motions of the gas in galaxies, prompting a rapid consumption of the cool gas reservoir in a starburst. Starbursts would provide a lot of feedback in the form of energy and momentum injection which can remove the remainder of the cold gas reservoir, quenching the galaxy. Hence, collisions have a lot of the properties we would look for in something that creates a red sequence galaxy.

THE RATE OF GALACTIC COLLISIONS – Back in the Chapter 6, we argued that stellar collisions are rare, but what about galaxies? Given the rhetorical setup for this chapter, it's not surprising that galaxy collisions are a lot more frequent. We can use the same collision time argument for galaxies, but the size, density, and speeds of galaxies all have different fiducial scales. From Equation 6.8, we have:

$$t_c = \frac{v^3}{4\pi G^2 \langle M \rangle^2 n_\star} \tag{7.27}$$

$$= 4.2 \text{ Gyr} \left(\frac{v}{100 \text{ km s}^{-1}} \right)^3 \left(\frac{\langle M \rangle}{10^{12}\ M_\odot} \right)^{-2} \left(\frac{n_{\rm gal}}{1 \text{ Mpc}^{-3}} \right)^{-1} \tag{7.28}$$

where we have used the scalings from the Local Group of galaxies where the velocity dispersion refers to the galaxy-galaxy velocity dispersion, the density refers the characteristic separations and the galaxy mass has been chosen to be approximately that of the Milky Way or Andromeda systems. For context, the scale for close interactions between galaxies is also large given their velocity dispersions: > 500 kpc.

DYNAMICAL FRICTION – Galaxies actually colliding at impact parameters comparable to their size scales (10s of kpc) still take a long time (3×10^4 Gyr), which suggests that clusters will evolve and relax without there being many gravitational collisions. However, the process of *dynamical friction* affects galaxies, facilitating their direct collisions with each other. In summary, dynamical friction occurs when a large object (a galaxy with mass M) has gravitational encounters with many objects of much smaller mass (stars of mass m in the other galaxy). If the large mass object passes through a field of lower density objects, it accelerates them toward it. The resulting concentration of matter following the more massive object then pulls backward on it, slowing it down, much like friction would act. The key physics here is the asymmetry of mass scales requires modifying our simple relaxation and collision time arguments to imagine what happens when a galaxy of mass M passes by a single star of mass m.

In this case, we consider ourselves to be in the large impact parameter limit with respect to a strong encounter with the individual star. We consider the kinetic energies of the star and galaxy perpendicular to the direction of travel:

$$\Delta K = \frac{1}{2}M\left(\frac{2Gm}{bV}\right)^2 + \frac{1}{2}m\left(\frac{2GM}{bV}\right)^2 \quad (7.29)$$

$$= \frac{2G^2mM(M+m)}{bV^2}. \quad (7.30)$$

The total energy in the system must remain unchanged before and after the collision and the potential energy contribution is small at large separations. This also leads to a small change in the forward speed of the galaxy δV so we can calculate the full kinetic energy.

$$K_i = K_f \quad (7.31)$$

$$\frac{1}{2}MV^2 = \Delta K + \frac{1}{2}M(V+\delta V)^2 + \frac{1}{2}m(\frac{M}{m}\delta V)^2 \text{ thus} \quad (7.32)$$

$$\delta V \approx -\frac{\Delta K}{MV} = -\frac{2G^2m(M+m)}{bV^3}. \quad (7.33)$$

If we consider the passage of the large mass through a field of density of objects with volume number density of n, the we can calculate the decceleration:

$$\frac{\delta V}{\delta t} = \int_{b_{\rm min}}^{b_{\rm max}} 2\pi b db\, nV\delta V \quad (7.34)$$

$$= -\int_{b_{\rm min}}^{b_{\rm max}} 2\pi b db\, nV\frac{2G^2m(M+m)}{bV^3} \quad (7.35)$$

$$= -\frac{4\pi G^2(M+m)}{V^2}mn\ln\Lambda \quad (7.36)$$

$$\approx -\frac{4\pi G^2M}{V^2}\rho\ln\Lambda \quad (7.37)$$

In the last step, we have used $M \gg m$ and that mn is a mass density. This highlights how it doesn't really matter what the scale of the individual small targets are, so while we worked this out in the case of small stars, it could also be dark matter particles. The key difference here compared to the two-body relaxation derivation is the big differences in the mass scales. Note that mass scale matters: there is also more friction for larger mass objects in a given mass density field. This difference means that local dark matter does not affect stellar motion much, only decelerating the Sun by $\sim -10^{-4}$ km s^{-1} Gyr^{-1}. In contrast, the deceleration of a galaxy in the Local Group is $\sim -10^2$ km s^{-1} Gyr^{-1}. This friction facilitates direct collisions between galaxies, causing them to merge on small number of crossing times through the group.

GROUPS AND CLUSTERS – Galaxies mergers are thus common in clusters and groups. Collisions are generally classified as either *wet* or *dry* based on whether or not the galaxies contain a large fraction of gas. Since gas is a fluid and is collisional, during a galaxy collision, the gas will be pressed into a smaller volume through the impact. In general, increasing the density of a fluid will decrease its cooling time, prompting the formation of a molecular medium and the aforementioned burst of star formation. In contrast, dry mergers are a dynamical process where the phase distributions of stars mixing through the collision are the only processes. Mergers are further broken down by comparing the mass ratio of the two colliders with a *major* merger referring to comparable mass colliders and a *minor* merger occurring when $M_1 \lesssim 0.1 M_2$. Given the shape of the galaxy mass distribution is decreasing ($dN/dM \propto M^{-2}$), there are more small galaxies than large galaxies and minor mergers are more common than major mergers. Indeed, the Milky Way is undergoing a minor merger right now with the Sagittarius dwarf galaxy but will have to wait several Gyr before pending major merger with M31.

The Antennae galaxy merger (Figure 7.16) is the classic case of a wet galaxy merger, where the gas in the galaxy is force into cold dense clouds which are then undergoing a starburst. Figure 7.16 also shows some of the consequences of dry mergers through the stellar shells surrounding the elliptical galaxy NGC 474. The shells highlight regions of phase space that are populated after the collision.

Galaxy collisions are thus the contrasting mode of galaxy evolution with respect to secular evolution. Appealing to collisions explains the changing demography of galaxies as they move into groups and clusters. Cosmology holds that new galaxies and halos are constantly falling deeper into the potential wells of groups and clusters over cosmic time, so gravity forces ever more systems closer together.

There is no clear dividing line between the definition of a group and a cluster except clusters have more members and larger masses. At the upper end of this continuum, the most massive clusters have several characteristic features that make them interesting laboratories for galaxy evolution. The mass scales of clusters make them the largest coherent objects in the Universe, with masses of $M_{\rm cl} \gtrsim 10^{14}\ M_\odot$ and core size scales of ~ 1 Mpc. These masses are almost entirely dark matter dominated with $M/L > 10^2\ M_\odot/L_\odot$ in solar units (stellar populations usually have $M/L < 10\ M_\odot/L_\odot$ in most bands). These large mass-to-light ratios were calculated by determining the virial mass of the cluster and comparing to the luminous mass, which provided one of the first lines of evidence for dark matter. However, counting only the mass found in stars does

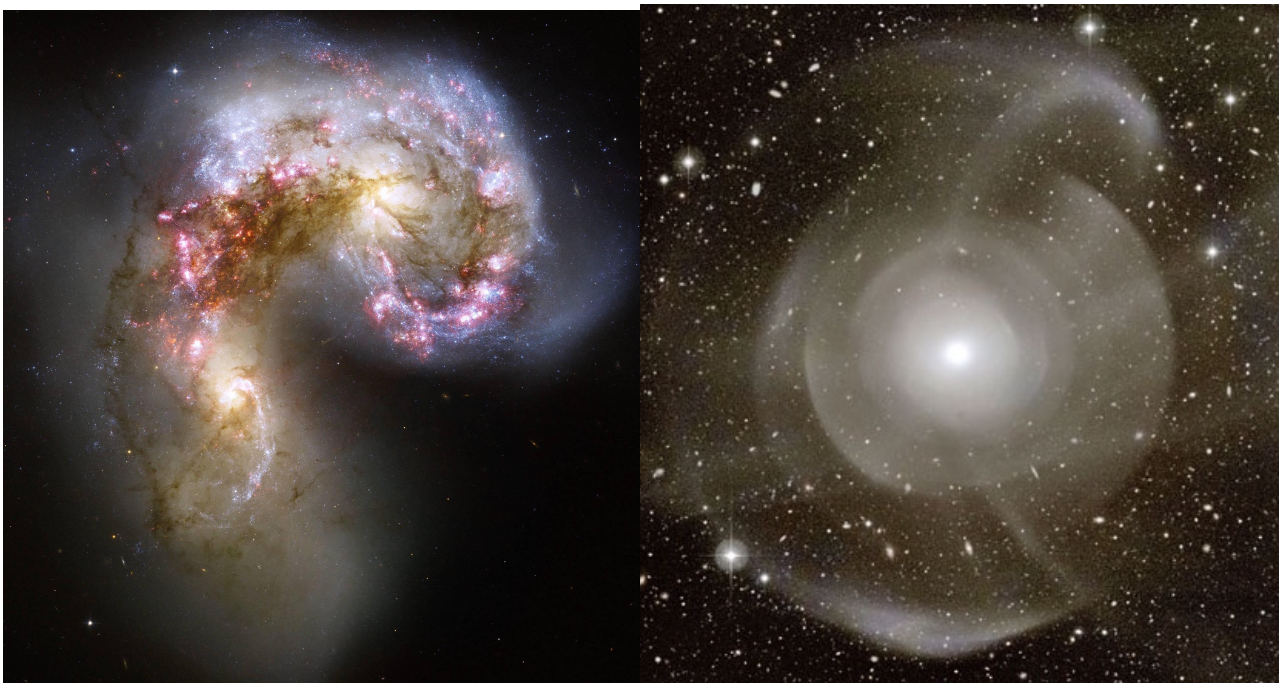


Figure 7.16: Two contrasting types of galaxy merger. (left) The Antennae Galaxy Merger as imaged in three colours by the Hubble Space Telescope and (right) CFHT imaging of NGC 474, a red sequence galaxy that has undergone several minor dry mergers, which showcase the presence of stellar streams.

not capture the baryonic mass budget of clusters particularly well. Most (90%) of the visible emission in the cluster is actually found in x-ray emitting plasma as depicted in Figure 7.17.

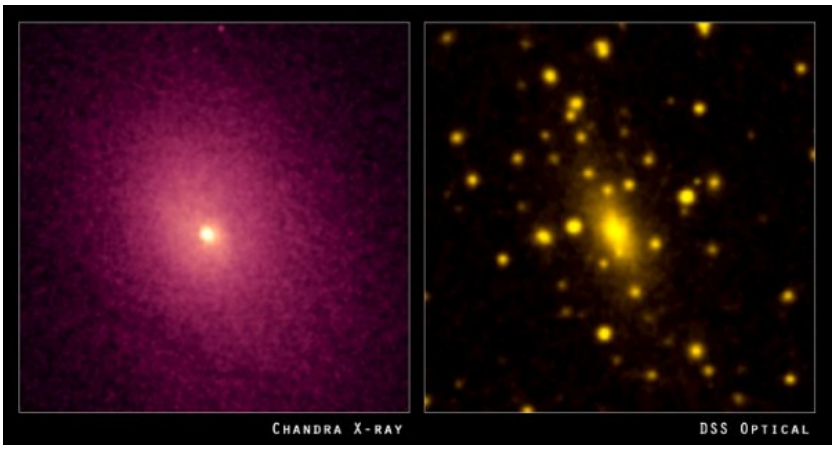


Figure 7.17: Comparing x-ray emission to optical light in a galaxy cluster. The intracluster medium is filled with relatively high density, high temperature plasma.

RAM PRESSURE STRIPPING – This plasma has $T \gtrsim 10^7$ K and densities of $n \sim 10^{-3}$ cm^{-3}. Note the equivalent pressure of such gas $P/k = nT = 10^4$ K cm^{-3}, which is comparable to the ISM pressure inside the typical galaxy. This high value of the pressure shows that

galaxies will have their ISM shaped by the intracluster medium as they enter into the regions. Since galaxies enter galaxies with speeds comparable to the virial velocity of the cluster ($10^{2.5}$ km s^{-1}), they crash into the hot plasma at high speed. Note that the plasma is relatively static in these clusters since the sound speed $c_s \sim \sqrt{kT/m_{\rm H}}$ is comparable to the speed at which galaxies fall in, so any inhomogeneities get smoothed out by sounds waves. In the frame of the galaxy, the cluster medium amounts to a high speed wind passing through the galaxy stripping out any gaseous material. The highest pressure regions of the ISM will also get compressed further. A particularly dramatic example of the *ram pressure stripping* is shown in Figure 7.18.

Figure 7.18: Ram pressure stripping of spiral galaxy entering a stellar cluster for the first time. Note the trail of young, blue stars coming off the galaxy associated with the dense gas clumps being stripped out of the system.

Rich clusters show two other phenomenological features that fit nicely with our broad themes of galaxy evolution. First, we see a difference in the specific star forming rates and the relative numbers of different galaxies in clusters. Clusters are dominated by non-star forming galaxies that are on the red sequence. These tend to be elliptical galaxies, but include a substantial number of So galaxies by morphological type. Examining the behaviour of the different mass distributions of galaxies in clusters vs. in small groups, we see that the fraction of red sequence galaxies not only increases, but it is also the location where one finds the most massive red galaxies.

Key Points

- Galaxy collisions are a primary channel to create galaxies dominated by random motions.
- Galaxy collisions are much more common than stellar collisions and they are further enhanced by the process of dynamical friction. Dynamical friction occurs when a massive object pulls on a low density field of objects and slows down as a result (Equation 7.37).
- A wet galaxy collision happens when one or more of the galaxies has cold gas in the system, so the collision triggers a starburst. A dry galaxy collision happens between two gas-poor systems and leads just to disrupted stellar structures.
- Galaxy groups and clusters have high rates of collisions and have a much higher fraction of red-sequence / quenched galaxies compared to the field because of the higher rates of galaxy collision.
- Galaxy clusters are filled with hot ($T > 10^7$ K) gas that can remove cold gas from galaxies falling into the cluster through ram pressure stripping.

7.6 Summarizing Galaxy Evolution

In the preceding chapter, we focused on the shapes of galaxies as governed by dynamics. In this and previous chapters, we have broadly discussed the light emitted from galaxies, which has three main origins: stars, accretion into the central supermassive black hole, and cooling processes in the interstellar medium. The mix between these types of light varies with galaxy type. Most galaxies have their emission dominated by the emission from stars either directly (photospheric emission) or after reprocessing through dust. Nuclear gas rich regions have most of their emission from AGN. The hot ISM of some elliptical galaxies can also provide a substantial contribution to the overall SED, especially since their starlight is relatively weak.

It now comes to a question of where these shapes and light have come from, and how those interpret those results in terms of the intrinsic properties of galaxies like mass and metallicity. Thus, we now turn to the long term evolution of galaxies. Viewed through the lens of galaxy evolution, AGN would be little more than an observational novelty were it not for the notion of *feedback*. Feedback refers

to energy and momentum deposited in the galaxy, specifically in the gas, by AGN radiation and jets. Since AGN can produce such large amounts of radiation in relatively small volumes, this radiation directly affects the surrounding ISM through dissociating, ionizing and heating. Particle outflows from the AGN (referred to as *winds*) also directly deposit momentum into the ISM.

Similarly, star formation produces high mass stars in the IMF that also inject feedback into the local ISM and regulate the properties of the surrounding gas. Star formation also materially changes galaxies by (wait for it) forming stars, and this increases the stellar mass and metal content of the system. In both AGN and star formation, feedback can eject a lot of the material from the galaxy, potentially quenching star formation by starving the system of its gas reservoir. These questions focus the community's recent interest on the idea of feedback and the quenching of star formation however these processes play out in the broader context of galaxy evolution.

In Figure 7.19, we attempt to summarize the different physical effects that lead to the observed populations of galaxies. This summary captures the primary physical effects that lead to a specific type of galaxy but they cannot capture all the subtleties involved. The study of galaxies is difficult primarily because it synthesizes such a wide array of physical effects. From stellar dynamics and evolution to optical properties of dust grains, the range of concepts that could be important is so vast that it can seem like an intractable web of material. Thus, it makes sense to revisit the answers to questions we asked back in Chapter 5 which we now have a better perspective on.

WHY DO GALAXIES HAVE THE SHAPES THAT THEY DO? – The shapes of galaxies are set by the shapes of orbits of stars inside their gravitational potential. Their potential is mostly set by the the dark matter and the stars in the galaxy. Stars are at different points in their orbits and fill the volume of a galaxy so we see the light in the shape of the galaxy. The shapes of galaxies indicate how well organized those orbits are with disk galaxies being "dynamically cold" with little random motion and elliptical galaxies being much hotter. These dynamics are perturbed through galaxy collisions and mergers which heat up the stars.

- How are stars and gas moving through galaxies? – Stellar orbits are collisionless and stars interact with each other primarily through long range gravitational interactions. Over time, these interactions slowly increase the amount of random motions in stellar population. Gas is a collisional fluid and vertical motions rapidly dissipate leading to rotating disks of material.

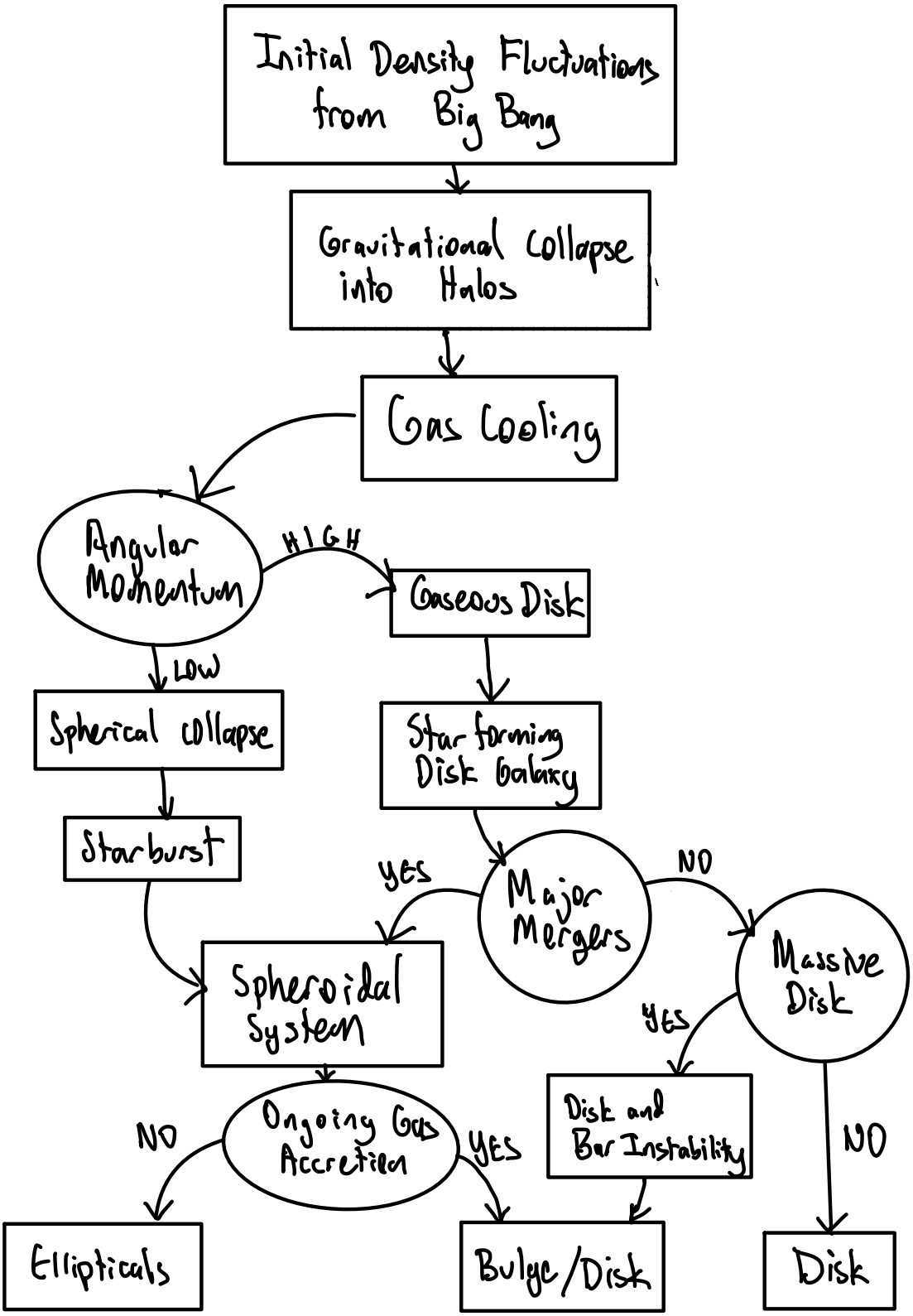


Figure 7.19: Evolutionary Flowchart for Galaxies. Adapted from Mo et al. [2010].

- *What is different about the history of the bulge and disk of a galaxy?* – The bulge of the galaxy has an older stellar population with a higher degree of random motions. The bulge formed earlier in the galaxy's evolution from low angular momentum material.
- *Why are the masses of the black holes in the centres of galaxies correlated with their bulges?* – While we don't know for sure, this seems to be established by feedback from the active galactic nucleus in the centre of galaxies controlling the amount of material that can accrete onto the black hole. As the bulge mass increases, the strength of the feedback (which increases with mass of the black hole) can be higher without disrupting the system.

WHAT PROCESSES SHAPE THE LIGHT WE RECEIVE FROM GALAXIES? – The light from a stellar population is primarily shaped by its age, with relatively blue light indicating recent star formation. There is substantial reprocessing of light by dust and nebular emission from gas, as well as some contributions from AGN. However, the dominant colour and amount of light emitted from galaxies is set by how much star formation happened and when.

- *What distinguishes the red sequence from the blue cloud galaxies?* – The blue cloud of galaxies (also known as the the star forming main sequence) is characterized by ongoing recent star formation. In the absence of such star formation, the stellar population ages and becomes red.
- *Are these galaxy populations connected through evolution?* – Yes. A galaxy can become quenched if it no longer has cold gas, which leads to it stopping star formation and becoming red. If the system is disk like, it will appear as an So or lenticular system. If the system is spheroidal, it will become an elliptical galaxy. Galaxy collisions can drive quenching through a starburst system as can a galaxy entering a cluster, where the hot intercluster medium disrupts the cold ISM in galaxies falling into the cluster.

WHY DO GALAXIES HAVE THE CONTENTS THAT THEY DO? – The baryonic component of the Universe is only $1/8$ of the matter content. The properties of dark matter, namely that it is cold and self-gravitating seem to set where galaxies form, their relative distribution of masses, and where the galaxies are relative to each other.

- *Why are galaxies mostly stars with only a little bit of gas?* – This wasn't always true. Galaxies start out gas rich and then convert that gas into stars. In the early universe, the gas rich galaxies were actively star formation with a peak around $z = 2$.

- *What physics makes the masses galaxies change over time?* – The accretion of gas from the hot intergalactic medium into halos. This material cools and enters galaxy disks or spheroids and becomes stars through the process of star formation.
- *What physics changes the metallicity of galaxies and the enrichment of the material in a galaxy with the products of stellar evolution?* – We did not cover this is detail because I went slowly. However, the processes of stellar nucleosynthesis shows that most star forming populations require feedback and an influx of new matter to explain the metallicities of stars that we see.
- *What makes a galaxy build up mass faster or slower?* – The amount of material falling in from the halo and the angular momentum of that material. Material with high angular momentum settles into a disk which forms stars continuously and slowly. Material with low angular momentum forms all the stars at once leading to spheroidal system.

How do the answers to these questions change as the Universe evolves? The particular routes for galaxy evolution are controlled by gas consumption. Galaxy formation and evolution is fed by the accretion and consumption of gas, since this is the channel for forming stars and altering the stellar populations. The gas reservoir in galaxies is increased through accretion of material from other systems or the hot gas in the intergalactic medium. However, gas is highly responsive to feedback, either through AGN or bursts of star formation.

7.7 The Beginning

At this point, we have presented a high level view of galaxy evolution with an attention on the subjects most tractable at the third-year level. However, galaxy evolution is a major area of ongoing research and one of the main reasons I presented these notes is that we are in the middle of developing a greatly improved understanding of how all these effects are playing out. The previous standard books were excellent for the time that they were written but our picture has changed a lot in the past 20 years and it will definitely develop further over the next 20.

8
Data Exercises

8.1 The Rotation Curve of M33

In this exercise, we will explore the rotation curve of the nearby galaxy M33. We will use data from Koch et al. [2018] to measure the rotation speed of the galaxy as a function of distance from the centre. Then, using estimates of the galaxy's mass, we can infer the presence of dark matter in the system based on the galaxy's rotation.

The method that we will use is to observe the Doppler shift from atomic hydrogen gas in the galaxy's interstellar medium. In the main text, we focused on the electronic transitions of hydrogen which give rise to the Lyman, Balmer, etc. series of transitions. Here, we are going to focus on the *hyperfine transition* of the ground state of atomic hydrogen.

A hydrogen atom consists of a proton and an electron. The main electronic energy levels come from the Coulomb interaction between these two particles and when formulated under quantum mechanics, can be used to explain the electronic energy levels of the atom. These give rise to transitions of order a few eV. There are also smaller scale energy levels within the hydrogen atom. These come from magnetic interactions between the particles, where the angular momentum of the different particles makes them behave like little magnets. Each of the particles, the proton and the electron, have an intrinsic "spin" which means they behave like small magnets. Moreover, if the electron is in a state with angular momentum, then it effectively orbiting the proton and thus acts like a little electromagnet as a current-carrying loop. Interactions between the proton's spin and the magnetic moment introduced by the electron's orbit lead to splitting of the electronic energy levels into *fine structure*. However, we're going to focus on the interaction between the spins of the two particles, which gives rise to the hyperfine structure.

As depicted in the figure, the ground state of hydrogen can have the spins of the two particles align (high energy) or antialigned (low

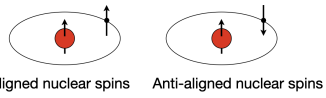


Figure 8.1: Spins of particles in a hydrogen atom.

energy). There is a quantum transition between these two states with an energy difference of 6 μeV corresponding to a photon with wavelength $\lambda = 21$ cm. This photon is in the radio portion of the spectrum and can be observed by telescopes like the VLA. Figure 8.2 shows the results of mapping M33 using the Very Large Array in data presented in Koch et al. [2018]. The left-hand panel shows the brightness of the emission in the galaxy, which illustrates the structure of the neutral atomic ISM. The right-hand panel shows the line-of-sight velocity at each point in the map, which has been derived from the Doppler shift for this spectral line. The colours of the map on the right correspond to the blue- and redshifted sides of the galaxy. There is a gradient of the velocity because M33 is a disk galaxy that is rotating. The top side of the galaxy is coming toward us and the bottom side is rotating away from us.

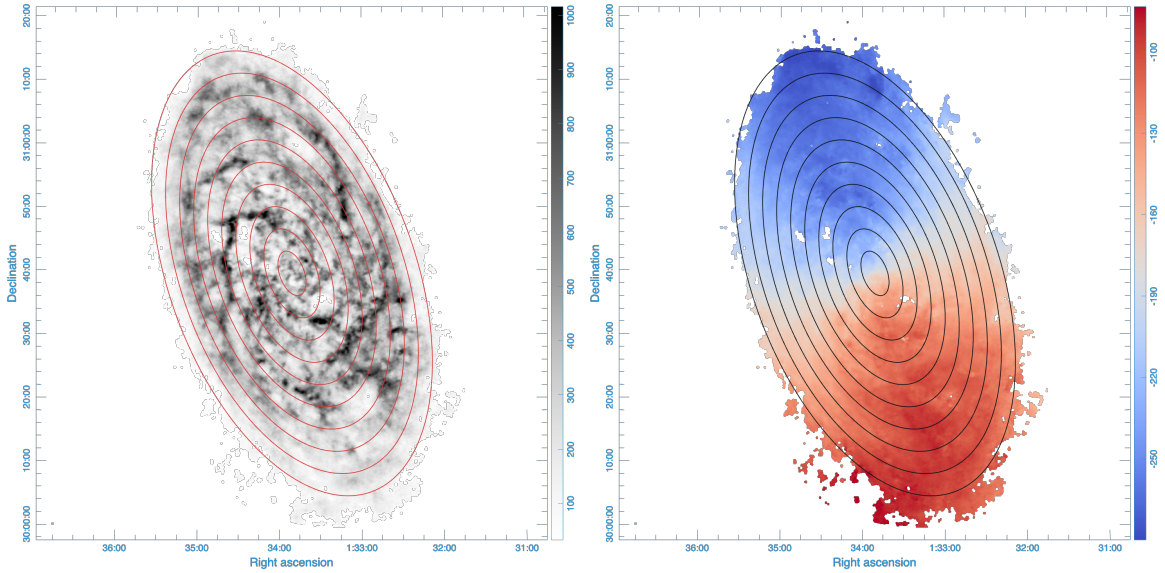


Figure 8.2: The brightness (left) and line of sight (right) velocity of atomic hydrogen gas in M33. Data taken from Koch et al. [2018].

The ellipses in the figure are lines of constant galactocentric radius, $R_{\rm gal}$ running from $R_{\rm gal} = 1, 2, \cdots 10$ kpc. These lines appear as ellipses rather than circles because M33 is inclined with respect to our line of sight. If we could see the galaxy face-on then these would appear as circles.

Our goal is to use the data in the right-hand panel to measure the *rotation curve* of the galaxy, which is a measure of the circular rotation speed around the centre of the galaxy as a function of radius: $V(R_{\rm gal})$. Galaxies do not rotate as solid disks. Instead, they are more like our solar system where the planets are moving around the Sun at different orbital speeds / angular frequencies. It is worth noting here that the analogy with our solar system isn't perfect: in the solar system, nearly all of the mass is in the Sun and the planets really do

orbit around the Sun. In a galaxy, the mass is not centrally concentrated and the rotation speed depends on how the mass is distributed in the galaxy.

We can track this motion through the Doppler shift of the gas, but the galaxy is really far away, so we cannot measure the proper motion of the gas. Hence, our knowledge of the velocity is restricted to the velocity projected along the line of sight. We can still use this to derive the rotation curve by assuming that the gas is on circular orbits, which is fairly close to reality. We can then use the geometry of an inclined disk to determine the component of the velocity vector that is projected along the line of sight. In astronomy, we describe the orientation of a disk galaxy in terms of two angles: the inclination i and the position angle PA. The inclination is defined as the angle between a galaxy's angular momentum vector and the line of sight. For a face-on galaxy, $i = 0°$ and for an edge on galaxy $i = 90°$. The position angle PA is defined as the angle measured counterclockwise ("east of north") on the sky measured from the line heading to the north celestial pole around to the redshifted side of the major axis of the galaxy. For M33, $i = 55°$ and $PA = 201°$. We want to use this model of galaxy orientation to determine what the line-of-sight component of the velocity vector will be.

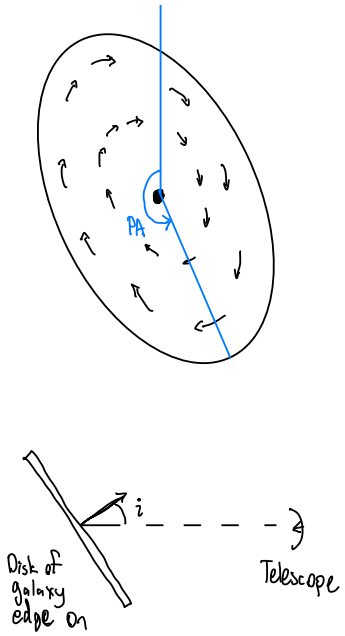


Figure 8.3: The geometry of an inclined disk galaxy.

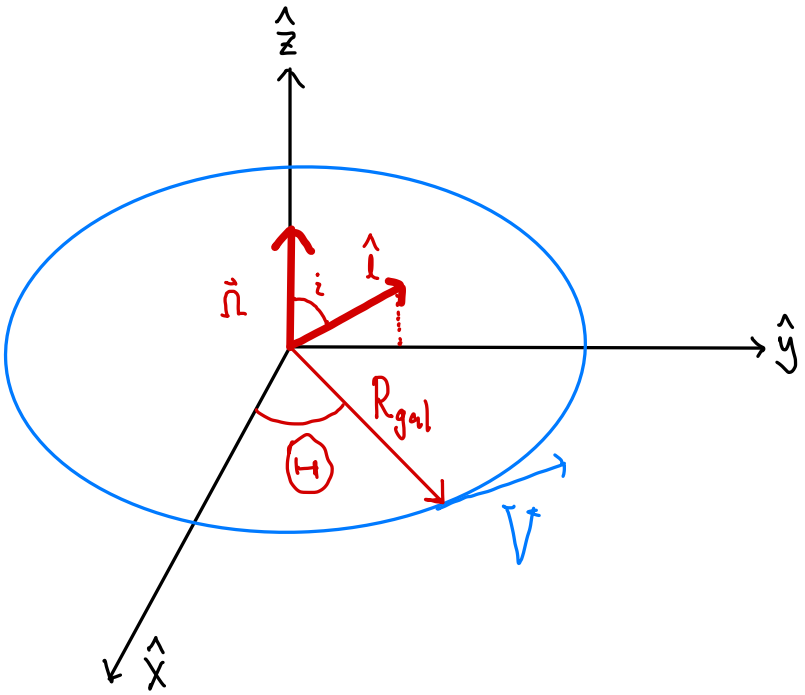


Figure 8.4: Vectors used to derive the line-of-sight component of the velocity in a rotating disk viewed at an angle.

Figure 8.4 illustrates a rotating ring of gas (blue) in the plane of

a galaxy (the $\hat{x}$-$\hat{y}$ plane. The ring has a radius defined by the vector $\vec{R}_{\rm gal}$, which can be decomposed into its components along the coordinate axes and expressed as:

$$\vec{R}_{\rm gal} = R_{\rm gal}\cos\Theta\hat{x} + R_{\rm gal}\sin\Theta\hat{y} \tag{8.1}$$

where Θ is the angle measured in the disk of the galaxy from the $\hat{x}$ axis. The $\hat{x}$ axis is the major axis of the galaxy. We will also assume that the motion the rotating ring is defined by an angular velocity vector $\vec{\Omega}$ parallel to the $\hat{z}$ axis. In this case the velocity vector $\vec{V} = \vec{\Omega} \times \vec{R}_{\rm gal}$. We want to determine the component of this velocity vector along the line of sight vector, here illustrated as the vector $\hat{\ell}$ which is in the $\hat{y}$-$\hat{z}$ plane and the angle between the vector and the $\hat{z}$ axis is the inclination angle i.

Given this geometry, we want to find $V_r = \hat{\ell} \cdot \vec{V}$, which is the magnitude of the velocity vector along the line of sight. This is then $\hat{\ell} \cdot \vec{\Omega} \times \vec{R}_{\rm gal}$, a vector identity once decomposed into the Cartesian coordinate system. Noting that $\hat{\ell} = \ell_y\hat{y} + \ell_z\hat{z} = \sin i\hat{y} + \cos i\hat{z}$ and $\vec{\Omega} = \Omega\hat{z}$,

$$\hat{\ell} \cdot \vec{\Omega} \times \vec{R}_{\rm gal} = \begin{vmatrix} 0 & 0 & R_{\rm gal}\cos\Theta \\ \sin i & 0 & R_{\rm gal}\sin\Theta \\ \cos i & \Omega & 0 \end{vmatrix} \tag{8.2}$$

The $|\cdots|$ is the determinant operator. This expression then reduces to $V_r = \Omega R_{\rm gal}\cos\Theta\sin i = V\cos\Theta\sin i$.

There is one final amendment to this expression, namely the entire galaxy is moving with respect to the observer at a *systemic velocity*, denoted $v_{\rm sys}$. We can just add that onto the velocity due to rotation to achieve the observed velocity.

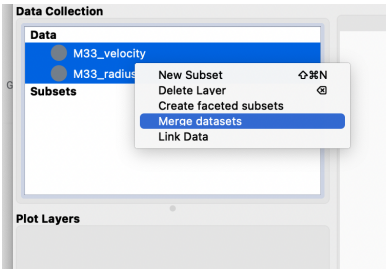
$$v_r = v_{\rm sys} + V\cos\Theta\sin i \tag{8.3}$$

We now can turn the data in Figure 8.2 (right) into a rotation curve, measuring V as a function of $R_{\rm gal}$. We'll accomplish this in GLUE using the two FITS images of the galaxy, `M33_velocity.fits` and `M33_radius.fits`.

FITS=Flexible Image Transport System, a common astronomical format for images

1. Load the two images into GLUE. When you drag the second image in, you will get a pop-up window asking if you want to use the Auto-Linking plugin. Select "Apply" to link these two images automatically.

2. Use shift-select to select both of the images in the Data Collection. Right click on the selected pair of images and choose "Merge datasets" from the context menu. This will combine the two data sets into a single data.



3. Examine the images by dragging the combined image and make a 2D image. This shows what's actually in the image files with nice coordinates. You can explore the visualization features to show off different features in the two images.

4. The main analysis comes from plotting the data set as a 2D scatter plot. In making your plot, you should select `PRIMARY [M33_radius]` as the x-axis and `PRIMARY [M33_velocity]` as the y-axis. This should yield a plot like what's below. The radius is measured in kpc and the velocity axis is measured in km/s. The figure has

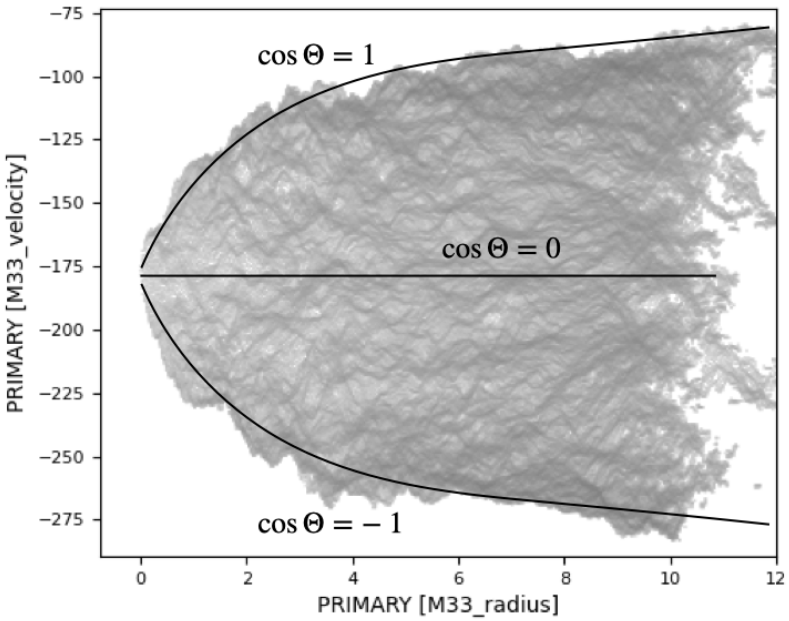


Figure 8.5: Sample Glue visualization of the 21-cm velocity data.

been annotated with where the data are coming from in the data set. First, the data in a roughly symmetric distribution around the velocity $v_{\rm sys} \approx -180$ km/s. The points are bounded by an envelope. Points at a fixed radius will range between $v = v_{\rm sys} + V$ (for $\cos\Theta = 1$) and $v = v_{\rm sys} - V$ for ($\cos\Theta = -1$). These bounds form curves illustrated in the figure. By examining the difference between the envelope points and $v_{\rm sys}$, we can infer V.

5. GLUE makes this easy for us by allowing us to calculate an arithmetic attribute that gives us the difference between the points and the systemic velocity. We can set this difference as

```
np.abs({PRIMARY [M33_velocity]} + 180)
```

which is $\Delta V = |v_r - v_{\rm sys}|$.

6. From here, we just need to read off some values of Δv as a function of radius. This gives us $V \sin i$ for that radius. Do this for a few points and we have a rotation curve!

8.2 *The Energetics of Stellar Clusters*

In this exercise, we will study the motions of stars and spatial distribution in an open stellar cluster to understand its gravitational potential energy and its kinetic energy. In this exercise, we will convert the Gaia data from sky coordinates to a three dimensional coordinate system as we have done before. Here, we will now incorporate the proper motion information of stars so that we can measure the velocity distribution and size of the clusters. This will allow us, with the virial theorem, to estimate the mass of the cluster.

Here, we'll rely on the coordinate transforms from Chapter 1. The only wrinkle here is that we are going to work in RA/Dec coordinates instead of Galactic coordinates, mostly because the proper motion information is given in this coordinate frame and we don't want to transform. Instead, we are going to work in a coordinate system that is centred on the cluster and oriented along the line of sight. This will make our coding a bit easier, and because we will assume that clusters are approximately spherically symmetric, this won't affect our physical interpretation.

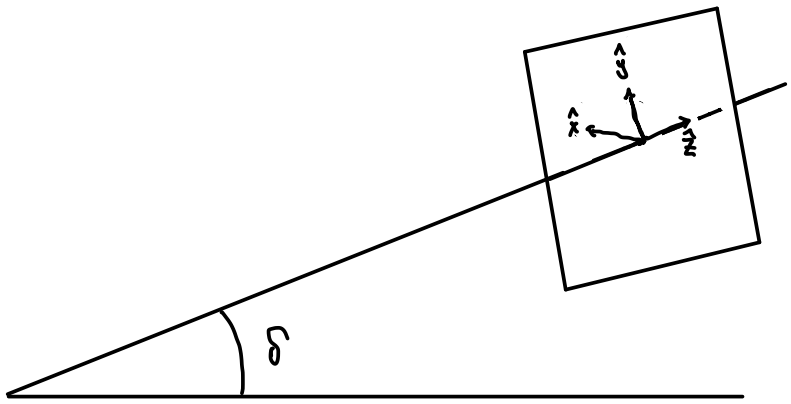


Figure 8.6: Coordinate System centred on a cluster at declination δ. The square illustrates a tangent plane to the celestial sphere which is perpendicular to the line of sight.

In this case, we can consider coordinates from the central position of the cluster (α_0, δ_0) at a mean distance of d_0.

$$x = d(\alpha - \alpha_0)\cos\delta_0 \tag{8.4}$$

$$y = d(\delta - \delta_0) \tag{8.5}$$

$$z = d - d_0. \tag{8.6}$$

This formulation is using small angle formulas to measure displacements along the plane of the sky as we did a lot in Chapter 1. Thus, the angles in this formula must be in radians and not degrees to use the small angle formula.

Practically, when we are working with Gaia data, we get our distances from parallax so that, in Glue's Arithmetic attributes syntax `d = 1000 / {parallax}` since the parallax is given in milliarcseconds. The values of α and δ are given in decimal degrees, so in using trig functions, we will need to convert to radians. Finally, to calculate α_0, δ_0, d_0, we will need to use the terminal. If we drag the data into the terminal and give it a name like `p`, we can calculate these mean values.

```
import numpy as np
In [1]: np.mean(p['ra'])
Out[1]: 56.585552742658436

In [2]: np.mean(p['dec'])
Out[2]: 24.090429213350326

In [3]: np.mean(1000/p['parallax'])
Out[3]: 136.03376603026103
```

This allows us to calculate the three dimensional coordinates of the cluster directly in Glue as new Mathematical Attributes. We can open these up and insert the code for these, for example, defining a new variable `x` as

```
1000/{parallax} * ({ra} - 56.58555) *np.pi / 180 * np.cos(24.0904 * np.pi/180)
```

This cosine term here is the correction for the convergence of lines of right ascension as they approach the celestial pole. The factors of $\pi/180^\circ$ are converting the decimal degrees to radians. Similarly, `y` is

```
1000 / {parallax} * ({dec} - 24.0904) * np.pi/180
```

and the `z` distance along the line of sight is

```
1000 / {parallax} - 136.03376
```

Then, we can also project the proper motions to get a velocity in this coordinate system:

$$v_x = \mu_\alpha d \cos\delta_0 \tag{8.7}$$

$$v_y = \mu_\delta d \tag{8.8}$$

$$v_z = v_r \tag{8.9}$$

It was a long time ago, but we wrote down that we could determine the velocity from proper motion units of

$$\left(\frac{v}{\text{km s}^{-1}}\right) = 4.74 \left(\frac{d}{1\text{ pc}}\right) \left(\frac{\mu}{1\,''/\text{yr}}\right). \tag{8.10}$$

We will convert these velocity values to km/s. Note that the Gaia data give the proper motions in the RA and Dec directions as `pmra` and `pmdec` respectively and, critically, the `pmra` includes the $\cos\delta_0$ term listed in the equation, i.e., `pmra` $\equiv \mu_\alpha \cos\delta$! These values are in milliarcseconds per year. Thus, our velocities in the x- and y-directions are

```
vx = 4.74 * {pmra} / {parallax}
vy = 4.74 * {pmdec} / {parallax}
```

where the factors of 1000 from milliarcseconds have cancelled out. The v_z coordinate is just given as `radial_velocity` in the structure. There are more stars in Gaia with good proper motions than values with radial velocities and those with radial velocities tend to be bright (to get a good spectroscopic detection). Hence, we will tend to focus on v_x and v_y values.

Feel free to explore your cluster here. You can show the shape of the cluster in 3D, add vectors to the velocities and colour the points by stellar properties. All quite nifty.

CALCULATING THE VELOCITY DISPERSION – We want to calculate the velocity dispersion of the cluster. In this case, we will use our previously calculated values of v_x and v_y and estimate the velocity dispersion simply by taking the standard deviation of the values in the terminal. You can drag the data collection or a subset into the IPython terminal and that will pop up a window asking you to name it. In this example, I use the variable name `p`. In the IPython terminal in Glue, calculate

```
np.std(p['vx'])
```

assuming that you named your variable `p` and the velocity in the x-direction is `vx`. Note that occasionally there are some bad data in the structure which are represented as `nan` values which just means Not-A-Number. In this case, you can calculate the standard deviation ignoring bad data using

```
np.nanstd(p['vx'])
```

This is the velocity disperion in the x-direction, $\sigma_{v,x}$. You can then do a similar exercise to calculate $\sigma_{v,y}$. The average of these two values is a good estimate for the 1D velocity dispersion of the cluster.

Figure 8.7 shows the velocity distribution of stars in a cluster. The mean value of this distribution (around -29.5 km/s) shows the average velocity of the cluster moving through space. The width (about 1 km/s) is the velocity dispersion in the x-direction.

CALCULATING THE CLUSTER RADIUS – Here, we can calculate the value of $r = \sqrt{x^2 + y^2 + z^2}$. All that hard work at the beginning makes this a little easier. We can use our IPython terminal to calculate the values of r and the estimate the cluster radius R as the mean value of all the radii: $R = \langle r \rangle$.

In this case, we can calculate this in the IPython terminal:

```
r = np.sqrt(p['x']**2 + p['y']**2 + p['z']**2)
print(np.mean(r))
```

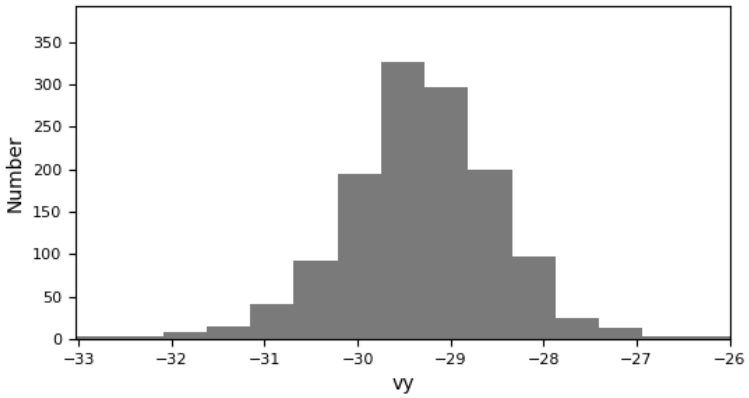


Figure 8.7: Velocity distribution of the stars in a cluster.

.

This will print out the mean value of the stellar radius, which we will take as the cluster radius, R.

9
Appendix

9.1 Essential Constants

The Essential Physical Constants

Name	Canonical Variable	Value
The speed of light	c	3.00×10^{8} m/s
The gravitational constant	G	6.67×10^{-11} m^3 kg^{-1} s^{-2}
Planck's constant	h	6.63×10^{-34} J s
Boltzmann's constant	k	1.38×10^{-23} J/K
The Stefan-Boltzmann Constant	σ_{SB}	5.67×10^{-8} W m^{-2} K^{-4}
Characteristic Scales		
Solar mass	$M_\odot$	1.99×10^{30} kg
Solar radius	$R_\odot$	6.96×10^{8} m
Solar luminosity	$L_\odot$	3.83×10^{26} W
Astronomical unit	AU	1.49×10^{11} m
Parsec	pc	3.09×10^{16} m
electron-Volt	eV	1.60×10^{-19} J
Hydrogen mass	m_{H}	1.67×10^{-27} kg
Electron mass	m_e	9.11×10^{-31} kg

9.2 *Gaia Observation Keywords*

A limited set of column names for the Gaia database are given below, but for the full set of keywords for Gaia DR2 is available from the Gaia archive: https://gea.esac.esa.int/archive/documentation/GDR2/Gaia_archive/chap_datamodel/sec_dm_main_tables/ssec_dm_gaia_source.html

- `source_id` — Gaia source identifier
- `mg` — Absolute magnitude in the Gaia G band
- `phot_g_mean_mag` — Apparent magnitude in the Gaia G band
- `bp_rp` — Gaia Blue-Red colour ($BP - RP$)
- `parallax` — Parallax angle in milliarcseconds
- `parallax_error` — Uncertainty in the parallax
- `ra`— Right Ascension
- `dec` — Declination
- `pmra` — Proper motion in the RA direction
- `pmdec` — Proper motion in the Dec direction
- `pmra_error` — Uncertainty in `pmra`
- `pmdec_error` — Uncertainty in `pmdec`
- `radial_velocity` — Radial velocity of star in km/s
- `radial_velocity_error` — Uncertainty in radial velocity
- `e_bp_min_rp_val` — line-of-sight reddening $E(B_P - R_P)$, a measure of dust
- `a_g_val` — Dust extinction in the G band
- `l` — Galactic longitude
- `b` — Galactic latitude
- `ecl_lat` — Ecliptic latitude
- `ecl_lon` — Ecliptic longitude

Mass ($M_\odot$)	$\log(L/L_\odot)$	$\log_{10}(T/\mathrm{K})$	M_G (mag)	G_{BP} (mag)	G_{RP} (mag)
0.10	-3.00	3.46	13.77	16.08	12.48
0.20	-2.32	3.51	11.57	13.04	10.43
0.30	-1.98	3.53	10.52	11.77	9.46
0.40	-1.69	3.56	9.63	10.70	8.63
0.50	-1.41	3.58	8.78	9.71	7.85
0.60	-1.13	3.61	7.95	8.74	7.10
0.70	-0.84	3.65	7.02	7.63	6.29
0.80	-0.57	3.69	6.20	6.67	5.57
0.90	-0.33	3.73	5.52	5.91	4.97
1.00	-0.13	3.76	4.98	5.30	4.48
1.50	0.64	3.83	3.06	3.25	2.74
2.00	1.25	3.97	1.73	1.74	1.71
2.50	1.61	4.04	1.15	1.12	1.20
3.00	1.91	4.09	0.68	0.62	0.77
3.50	2.17	4.14	0.30	0.21	0.41
4.00	2.38	4.17	-0.02	-0.12	0.11
5.00	2.73	4.24	-0.53	-0.66	-0.37
6.00	3.05	4.30	-0.98	-1.14	-0.78
7.00	3.30	4.34	-1.36	-1.53	-1.14
8.00	3.48	4.37	-1.65	-1.84	-1.42
9.00	3.66	4.40	-1.94	-2.14	-1.69
10.00	3.81	4.42	-2.19	-2.39	-1.93
12.00	4.05	4.46	-2.59	-2.81	-2.31
15.00	4.32	4.50	-3.06	-3.29	-2.76
18.00	4.54	4.54	-3.43	-3.67	-3.13
20.00	4.66	4.55	-3.63	-3.87	-3.32
25.00	4.91	4.59	-4.02	-4.27	-3.71
30.00	5.07	4.61	-4.30	-4.54	-3.98
40.00	5.35	4.65	-4.73	-4.98	-4.41
50.00	5.56	4.67	-5.06	-5.31	-4.73
100.00	6.10	4.72	-6.12	-6.38	-5.79
200.00	6.57	4.76	-7.10	-7.36	-6.79

Table 9.1: Properties of Zero Age Main Sequence stars in the Gaia passbands (G, G_{BP}, G_{RP}) used in Gaia DR2.

9.3 *Properties of Stars in Gaia*

9.4 *The Spitzer Layer*

Here, we provide some supplemental material that helps us understand the properties of a disk of stars. The structure of galaxies is set by the self-gravitation of the material in the disk. Thus, the determining the structure requires something of a circular argument. We need to know the structure to find the gravitational potential, but the gravitational potential is determined by the structure of the disk. Hence, the idea of *self-consistency* is a vital part of understanding how galaxies form.

The simplest (really) problem here is to solve for the vertical structure of the local Galactic disk using the *Jeans equations* which define the motion of stars and follow a lot of the heritage of fluid dynamics. We can state, without proof, that the Jeans equation describing vertical motion becomes

$$\frac{\partial}{\partial z} n \langle v_z^2 \rangle + n \frac{\partial \Phi}{\partial z} = 0. \tag{9.1}$$

where n is the number density of particles, Φ is the gravitational potential per unit mass, and z is the vertical distance from the midplane. We used $Z_{\rm gal}$ in the main text to avoid confusion with the redshift, but here we are safe from that particular variable choice and we stick with our regular variables. If we divide through by n and take another partial derivative with respect to z, we arrive at

$$\frac{\partial}{\partial z} \frac{1}{n} \frac{\partial}{\partial z} n \langle v_z^2 \rangle = - \frac{\partial^2 \Phi}{\partial z^2}. \tag{9.2}$$

In this approximation of the disk being thin, we can then further take $\nabla^2 \Phi \approx \partial^2 \Phi / \partial z^2$ and employ Poisson's equation and replace the right hand side with $-4\pi G \rho$. This points to a way to use the stellar motions to assess the mass in the local solar neighbourhood. Observationally, this would require constructing the second derivative of the velocity dispersion of the kinetic energy density of stars with respect to vertical displacement. Since stars a discrete and subject to completeness errors making this estimate of the mass density to the smoothness required to execute two numerical derivatives is tricky. Instead, we integrate vertically over both sides of the equation, considering the material that is found at a distance less than a total value Z from the disk.

$$\frac{1}{n} \frac{\partial}{\partial z} n \langle v_z^2 \rangle = -4\pi G \int_{-Z}^{Z} dz \, \rho(z) = -4\pi G \Sigma(|z| < Z) \tag{9.3}$$

This can be measured at a height of Z to give

$$-4\pi G \Sigma(|z| < Z) = \frac{1}{n(Z)} \left. \frac{\partial}{\partial z} n \langle v_z^2 \rangle \right|_{z=Z} \tag{9.4}$$

Making these measurements for stellar velocity dispersions at heights of $Z \sim 1$ kpc implies a local surface mass density of the galactic disk of $\Sigma \sim 60 M_\odot$ pc^{-2}. We can make an observational inventory of the stars and gas that fill the local volume. In general, the mass of the ISM is about $\Sigma_{\rm ISM} = 10\ M_\odot$ pc^{-2} with about 80% of that contribution in neutral (atomic and molecular) gas and the remaining 20% in ionized gas. The contributions of stars are thought to be about $\Sigma_\star = 40 - 50\ M_\odot$ pc^{-2} where there is substantial uncertainty given the difficulty in finding the faintest stars (the M dwarfs) as well in counting the population of stellar remnants, notably the white dwarfs which contribute a lot of mass and fade relatively quickly beyond the limits of detection. Regardless, the mass estimate of the local disk leaves little to room for a large contribution from dark matter in the solar neightbourhood. We think there is some contribution from dark matter, but this implies that the distribution is a thick spheroid and there is little room for a dark matter component that is distributed in the form of a disk.

Utilizing the integrals of motion in the thin disk approximation that we have established here allows for us to write down a self-consistent model of the Galaxy disk that provides a useful point of departure for understanding the stellar density profiles. In this case, let us consider only energy as an integral of motion such that the phase space distribution can be written as

$$f(\mathbf{x}, \mathbf{p}) = \frac{n_0}{\sqrt{2\pi\sigma_z^2}} \exp\left(-\frac{\epsilon}{\sigma_z^2}\right) \tag{9.5}$$

with $\epsilon = v_z^2/2 + \Phi(z)$ (energy per unit mass). If we integrate the phase space density over v_z, we get

$$n(z) = n_0 \exp\left[-\Phi(z)/\sigma_z^2\right], \tag{9.6}$$

which is basically an atmosphere model: the density drops off with distance from the midplane of the disk. We can further make the case that the density of stars provides all the gravitation in the disk of the galaxy to self-consistently identify the density distribution and gravitational potential that lead to this form. We will work with density distributions that are symmetric around the midplane and assume that the potential has an extremum in the disk such that $d\Phi/dz = 0$. Pushing on this further, let's also assume that the average mass of the individual stars is m such that

$$\rho(z) = m n_0 \exp\left[-\Phi(z)/\sigma_z^2\right]. \tag{9.7}$$

We then have

$$\frac{d^2\Phi}{dz^2} = 4\pi G m n_0 \exp\left[-\frac{\Phi(z)}{\sigma_z^2}\right]. \tag{9.8}$$

To solve this easily, we proceed by non-dimensionalizing the differential equation. Let's set

$$H^2 = \frac{\sigma^2}{8\pi G m n_0}, \zeta = \frac{z}{H}, \text{ and } \phi = \frac{\Phi}{\sigma^2} \tag{9.9}$$

to arrive at

$$2\frac{d^2\phi}{d\zeta^2} = e^{-\phi}. \tag{9.10}$$

As with all good differential equations, the best solution is to know where we are going. So let's multiply both sides of the equation by $d\phi/d\zeta'$ and integrate from o to ζ

$$\begin{aligned}
2\int_0^\zeta d\zeta' \frac{d^2\phi}{d\zeta'^2}\frac{d\phi}{d\zeta'} &= \int_0^{\phi(\zeta)} d\phi' e^{-\phi'}\frac{d\phi}{d\zeta'} && (9.11)\\
\int_0^\zeta d\zeta' \frac{d}{d\zeta'}\left(\frac{d\phi}{d\zeta'}\right)^2 &= e^{-\phi(0)} - e^{-\phi(\zeta)} && (9.12)\\
\left(\frac{d\phi(\zeta)}{d\zeta}\right)^2 - \left(\frac{d\phi(0)}{d\zeta}\right)^2 &= e^{\phi(0)} - e^{-\phi(\zeta)} && (9.13)\\
\left(\frac{d\phi(\zeta)}{d\zeta}\right)^2 &= 1 - e^{-\phi(\zeta)} && (9.14)\\
\frac{d\phi(\zeta)}{d\zeta} &= \sqrt{1 - e^{-\phi(\zeta)}} && (9.15)\\
& && (9.16)
\end{aligned}$$

To get to Equation 9.14, we have set $\phi(0) = d\phi/d\zeta|_{\zeta=0} = 0$ as the boundary condition on the potential. The first condition we are free to select from our contstant offset and the second condition asserts that the potential is at an extremum at the midplane. At this point, the equation is separable:

$$\int_0^\phi \frac{d\phi'}{\sqrt{1 - e^{-\phi'}}} = \int_0^{\zeta'} d\zeta'. \tag{9.17}$$

The left-hand integral can be executed directly, but it's smoother to make a substitution that $u = e^{-\phi'/2}$ such that $du = -e^{-\phi'/2}d\phi'/2$ or

that $d\phi' = -2du/u$ and then

$$\int_1^t \frac{-2du}{u\sqrt{1-u^2}} = \zeta \tag{9.18}$$

$$2\ln\left(\frac{\sqrt{1-t^2}+1}{t}\right) = \zeta \tag{9.19}$$

$$\frac{\sqrt{1-t^2}+1}{t} = e^{\zeta/2} \tag{9.20}$$

$$\frac{\sqrt{1-t^2}+1}{t} + \frac{t}{\sqrt{1-t^2}+1} = e^{\zeta/2} + e^{-\zeta/2} \tag{9.21}$$

$$\frac{2(1+\sqrt{1-t^2})}{t\sqrt{1-t^2}+1} = e^{\zeta/2} + e^{-\zeta/2} \tag{9.22}$$

$$\frac{2}{t} = e^{\zeta/2} + e^{-\zeta/2} \tag{9.23}$$

$$t = \frac{2}{e^{\zeta/2} + e^{-\zeta/2}} \tag{9.24}$$

$$t = \mathrm{sech}(-\zeta/2) \tag{9.25}$$

Equation 9.22 is substituted to simplify the radicals in the equation. The final leap of insight here is to realize that the variable $t = \exp(-\phi/2)$ is just $\sqrt{n/n_0}$ given Poisson's equation and then we can find the number density of stars as

$$n = n_0\,\mathrm{sech}^2[-z/(2H)] \tag{9.26}$$

with the scale height $H = \sigma/\sqrt{8\pi G m n_0}$.

The sech^2 profile forms a nicely behaved density distribution that is smooth near $z = 0$ and then limits to an exponential form on either side of the disk with a scale height determined by the velocity dispersion of the stars and the local number density. Note that this is a single form of a density distribution that relied on isothermality (i.e., constant midplane velocity dispersion) and a conjectured form of the distribution function f. It is not necessary to have this simple of a form and the reality of our Galaxy points to multiple independent populations with different properties.

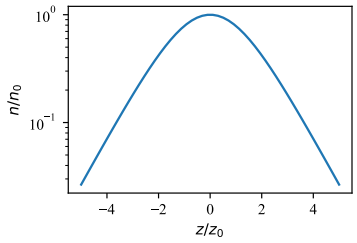


Figure 9.1: The sech^2 density profile for stars in a thin isothermal disk.

Bibliography

S. M. Adams, C. S. Kochanek, J. R. Gerke, K. Z. Stanek, and X. Dai. The search for failed supernovae with the Large Binocular Telescope: confirmation of a disappearing star. *Mon. Not. Roy. Astron. Soc.*, 468(4):4968–4981, July 2017. DOI: 10.1093/mnras/stx816.

Nate Bastian, Kevin R. Covey, and Michael R. Meyer. A Universal Stellar Initial Mass Function? A Critical Look at Variations. *Annu. Rev. Astron. Astrophys.*, 48:339–389, September 2010. DOI: 10.1146/annurev-astro-082708-101642.

Michael S. Bessell. Standard Photometric Systems. *Annu. Rev. Astron. Astrophys.*, 43(1):293–336, September 2005. DOI: 10.1146/annurev.astro.41.082801.100251.

Joss Bland-Hawthorn and Ortwin Gerhard. The Galaxy in Context: Structural, Kinematic, and Integrated Properties. *Annu. Rev. Astron. Astrophys.*, 54:529–596, September 2016. DOI: 10.1146/annurev-astro-081915-023441.

J. S. Bullock, A. Dekel, T. S. Kolatt, A. V. Kravtsov, A. A. Klypin, C. Porciani, and J. R. Primack. A Universal Angular Momentum Profile for Galactic Halos. *Astrophys. J.*, 555:240–257, July 2001. DOI: 10.1086/321477.

Matthew Colless, Gavin Dalton, Steve Maddox, Will Sutherland, Peder Norberg, Shaun Cole, Joss Bland-Hawthorn, Terry Bridges, Russell Cannon, Chris Collins, Warrick Couch, Nicholas Cross, Kathryn Deeley, Roberto De Propris, Simon P. Driver, George Efstathiou, Richard S. Ellis, Carlos S. Frenk, Karl Glazebrook, Carole Jackson, Ofer Lahav, Ian Lewis, Stuart Lumsden, Darren Madgwick, John A. Peacock, Bruce A. Peterson, Ian Price, Mark Seaborne, and Keith Taylor. The 2dF Galaxy Redshift Survey: spectra and redshifts. *Mon. Not. Roy. Astron. Soc.*, 328(4):1039–1063, December 2001. DOI: 10.1046/j.1365-8711.2001.04902.x.

C. Conroy and J. E. Gunn. The Propagation of Uncertainties in Stellar Population Synthesis Modeling. III. Model Calibration, Com-

parison, and Evaluation. *Astrophys. J.*, 712:833–857, April 2010. DOI: 10.1088/0004-637X/712/2/833.

Gerard de Vaucouleurs, Antoinette de Vaucouleurs, Jr. Corwin, Herold G., Ronald J. Buta, Georges Paturel, and Pascal Fouque. *Third Reference Catalogue of Bright Galaxies*. 1991.

Aaron Dotter. MESA Isochrones and Stellar Tracks (MIST) 0: Methods for the Construction of Stellar Isochrones. *Astrophys. J. Supp.*, 222(1):8, January 2016. DOI: 10.3847/0067-0049/222/1/8.

Bruce T. Draine. *Physics of the Interstellar and Intergalactic Medium*. 2011.

A. Einstein. Die Grundlage der allgemeinen Relativitätstheorie. *Annalen der Physik*, 354(7):769–822, January 1916. DOI: 10.1002/andp.19163540702.

Eric Emsellem, Eva Schinnerer, Francesco Santoro, Francesco Belfiore, Ismael Pessa, Rebecca McElroy, Guillermo A. Blanc, Enrico Congiu, Brent Groves, I-Ting Ho, Kathryn Kreckel, Alessandro Razza, Patricia Sanchez-Blazquez, Oleg Egorov, Chris Faesi, Ralf S. Klessen, Adam K. Leroy, Sharon Meidt, Miguel Querejeta, Erik Rosolowsky, Fabian Scheuermann, Gagandeep S. Anand, Ashley T. Barnes, Ivana Bešlić, Frank Bigiel, Médéric Boquien, Yixian Cao, Mélanie Chevance, Daniel A. Dale, Cosima Eibensteiner, Simon C. O. Glover, Kathryn Grasha, Jonathan D. Henshaw, Annie Hughes, Eric W. Koch, J. M. Diederik Kruijssen, Janice Lee, Daizhong Liu, Hsi-An Pan, Jérôme Pety, Toshiki Saito, Karin M. Sandstrom, Andreas Schruba, Jiayi Sun, David A. Thilker, Antonio Usero, Elizabeth J. Watkins, and Thomas G. Williams. The PHANGS-MUSE survey – Probing the chemo-dynamical evolution of disc galaxies. *arXiv e-prints*, art. arXiv:2110.03708, October 2021.

Gaia Collaboration, C. Babusiaux, F. van Leeuwen, M. A. Barstow, C. Jordi, A. Vallenari, D. Bossini, A. Bressan, T. Cantat-Gaudin, M. van Leeuwen, A. G. A. Brown, T. Prusti, J. H. J. de Bruijne, et al. Gaia Data Release 2. Observational Hertzsprung-Russell diagrams. *Astron. & Astrophys.*, 616:A10, August 2018. DOI: 10.1051/0004-6361/201832843.

Frédéric Galliano. A nearby galaxy perspective on interstellar dust properties and their evolution. *arXiv e-prints*, art. arXiv:2202.01868, February 2022.

Philipp Girichidis, Thorsten Naab, Stefanie Walch, and Thomas Berlok. The in situ formation of molecular and warm ionized gas

triggered by hot galactic outflows. *Mon. Not. Roy. Astron. Soc.*, 505 (1):1083–1104, July 2021. DOI: 10.1093/mnras/stab1203.

Karl D. Gordon, Geoffrey C. Clayton, K. A. Misselt, Arlo U. Landolt, and Michael J. Wolff. A Quantitative Comparison of the Small Magellanic Cloud, Large Magellanic Cloud, and Milky Way Ultraviolet to Near-Infrared Extinction Curves. *Astrophys. J.*, 594(1):279–293, September 2003. DOI: 10.1086/376774.

Gravity Collaboration, R. Abuter, A. Amorim, M. Bauböck, J. P. Berger, H. Bonnet, W. Brandner, Y. Clénet, V. Coudé Du Foresto, P. T. de Zeeuw, J. Dexter, G. Duvert, A. Eckart, F. Eisenhauer, N. M. Förster Schreiber, P. Garcia, F. Gao, E. Gendron, R. Genzel, O. Gerhard, S. Gillessen, M. Habibi, X. Haubois, T. Henning, S. Hippler, M. Horrobin, A. Jiménez-Rosales, L. Jocou, P. Kervella, S. Lacour, V. Lapeyrère, J. B. Le Bouquin, P. Léna, T. Ott, T. Paumard, K. Perraut, G. Perrin, O. Pfuhl, S. Rabien, G. Rodriguez Coira, G. Rousset, S. Scheithauer, A. Sternberg, O. Straub, C. Straubmeier, E. Sturm, L. J. Tacconi, F. Vincent, S. von Fellenberg, I. Waisberg, F. Widmann, E. Wieprecht, E. Wiezorrek, J. Woillez, and S. Yazici. A geometric distance measurement to the Galactic center black hole with 0.3% uncertainty. *Astron. & Astrophys.*, 625:L10, May 2019. DOI: 10.1051/0004-6361/201935656.

A. Heger, C. L. Fryer, S. E. Woosley, N. Langer, and D. H. Hartmann. How Massive Single Stars End Their Life. *Astrophys. J.*, 591 (1):288–300, July 2003. DOI: 10.1086/375341.

Amina Helmi. The stellar halo of the Galaxy. *Astron.& Astrophys. Rev.*, 15(3):145–188, June 2008. DOI: 10.1007/s00159-008-0009-6.

J. Holmberg, B. Nordström, and J. Andersen. The Geneva-Copenhagen survey of the solar neighbourhood. III. Improved distances, ages, and kinematics. *Astron. & Astrophys.*, 501:941–947, July 2009. DOI: 10.1051/0004-6361/200811191.

R. C. Kennicutt and N. J. Evans. Star Formation in the Milky Way and Nearby Galaxies. *Annu. Rev. Astron. Astrophys.*, 50:531–608, September 2012. DOI: 10.1146/annurev-astro-081811-125610.

Eric W. Koch, Erik W. Rosolowsky, Felix J. Lockman, Amanda A. Kepley, Adam Leroy, Andreas Schruba, Jonathan Braine, Julianne Dalcanton, Megan C. Johnson, and Snežana Stanimirović. Kinematics of the atomic ISM ifigun M33 on 80 pc scales. *Mon. Not. Roy. Astron. Soc.*, 479(2):2505–2533, September 2018. DOI: 10.1093/mnras/sty1674.

Mark R. Krumholz. Notes on Star Formation. *arXiv e-prints*, art. arXiv:1511.03457, November 2015.

Adam K. Leroy, Karin M. Sandstrom, Dustin Lang, Alexia Lewis, Samir Salim, Erica A. Behrens, Jérémy Chastenet, I-Da Chiang, Molly J. Gallagher, Sarah Kessler, and Dyas Utomo. A z = 0 Multiwavelength Galaxy Synthesis. I. A WISE and GALEX Atlas of Local Galaxies. *Astrophys. J. Supp.*, 244(2):24, October 2019. DOI: 10.3847/1538-4365/ab3925.

Timothy C. Licquia and Jeffrey A. Newman. Sizing Up the Milky Way: A Bayesian Mixture Model Meta-analysis of Photometric Scale Length Measurements. *Astrophys. J.*, 831(1):71, November 2016. DOI: 10.3847/0004-637X/831/1/71.

P. Madau and M. Dickinson. Cosmic Star-Formation History. *Annu. Rev. Astron. Astrophys.*, 52:415–486, August 2014. DOI: 10.1146/annurev-astro-081811-125615.

U. Maio, K. Dolag, B. Ciardi, and L. Tornatore. Metal and molecule cooling in simulations of structure formation. *Mon. Not. Roy. Astron. Soc.*, 379:963–973, August 2007. DOI: 10.1111/j.1365-2966.2007.12016.x.

J. S. Mathis, W. Rumpl, and K. H. Nordsieck. The size distribution of interstellar grains. *Astrophys. J.*, 217:425–433, October 1977. DOI: 10.1086/155591.

Houjun Mo, Frank C. van den Bosch, and Simon White. *Galaxy Formation and Evolution*. 2010.

M. Moe and R. Di Stefano. Mind Your Ps and Qs: The Interrelation between Period (P) and Mass-ratio (Q) Distributions of Binary Stars. *Astrophys. J. Supp.*, 230:15, June 2017. DOI: 10.3847/1538-4365/aa6fb6.

Benjamin P. Moster, Rachel S. Somerville, Christian Maulbetsch, Frank C. van den Bosch, Andrea V. Macciò, Thorsten Naab, and Ludwig Oser. Constraints on the Relationship between Stellar Mass and Halo Mass at Low and High Redshift. *Astrophys. J.*, 710(2): 903–923, February 2010. DOI: 10.1088/0004-637X/710/2/903.

Donald E. Osterbrock and Gary J. Ferland. *Astrophysics of gaseous nebulae and active galactic nuclei*. 2006.

P. Padovani, D. M. Alexander, R. J. Assef, B. De Marco, P. Giommi, R. C. Hickox, G. T. Richards, V. Smolčić, E. Hatziminaoglou, V. Mainieri, and M. Salvato. Active galactic nuclei: what's in

a name? *Astron.& Astrophys. Rev.*, 25:2, August 2017. DOI: 10.1007/s00159-017-0102-9.

A. J. Pickles. A Stellar Spectral Flux Library: 1150-25000 Å. *Proc. Astron. Soc. Pac.*, 110(749):863–878, July 1998. DOI: 10.1086/316197.

A. R. Riedel, S. C. Blunt, E. L. Lambrides, E. L. Rice, K. L. Cruz, and J. K. Faherty. LACEwING: A New Moving Group Analysis Code. *Astron. J.*, 153:95, March 2017. DOI: 10.3847/1538-3881/153/3/95.

F. R. N. Schneider, H. Sana, C. J. Evans, J. M. Bestenlehner, N. Castro, L. Fossati, G. Gräfener, N. Langer, O. H. Ramírez-Agudelo, C. Sabín-Sanjulián, S. Simón-Díaz, F. Tramper, P. A. Crowther, A. de Koter, S. E. de Mink, P. L. Dufton, M. Garcia, M. Gieles, V. Hénault-Brunet, A. Herrero, R. G. Izzard, V. Kalari, D. J. Lennon, J. Maíz Apellániz, N. Markova, F. Najarro, P. Podsiadlowski, J. Puls, W. D. Taylor, J. T. van Loon, J. S. Vink, and C. Norman. An excess of massive stars in the local 30 Doradus starburst. *Science*, 359:69–71, January 2018. DOI: 10.1126/science.aan0106.

J. A. Sellwood. Secular evolution in disk galaxies. *Reviews of Modern Physics*, 86:1–46, January 2014. DOI: 10.1103/RevModPhys.86.1.

J. M. Shull, B. D. Smith, and C. W. Danforth. The Baryon Census in a Multiphase Intergalactic Medium: 30% of the Baryons May Still be Missing. *Astrophys. J.*, 759:23, November 2012. DOI: 10.1088/0004-637X/759/1/23.

Nathan Smith. Mass Loss: Its Effect on the Evolution and Fate of High-Mass Stars. *Annu. Rev. Astron. Astrophys.*, 52:487–528, August 2014. DOI: 10.1146/annurev-astro-081913-040025.

Volker Springel, Simon D. M. White, Adrian Jenkins, Carlos S. Frenk, Naoki Yoshida, Liang Gao, Julio Navarro, Robert Thacker, Darren Croton, John Helly, John A. Peacock, Shaun Cole, Peter Thomas, Hugh Couchman, August Evrard, Jörg Colberg, and Frazer Pearce. Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, 435(7042):629–636, June 2005. DOI: 10.1038/nature03597.

P. G. van Dokkum and C. Conroy. A substantial population of low-mass stars in luminous elliptical galaxies. *Nature*, 468:940–942, December 2010. DOI: 10.1038/nature09578.

Benjamin F. Williams, Andrew E. Dolphin, Julianne J. Dalcanton, Daniel R. Weisz, Eric F. Bell, Alexia R. Lewis, Philip Rosenfield, Yumi Choi, Evan Skillman, and Antonela Monachesi. PHAT. XIX. The Ancient Star Formation History of the M31 Disk. *Astrophys. J.*, 846(2):145, September 2017. DOI: 10.3847/1538-4357/aa862a.

Index