Last modified November 22, 2018

General Qualifying Exam Solutions: Galactic Astronomy

Jessica Campbell, Dunlap Institute for Astronomy & Astrophysics (UofT)

Contents

1 Galactic Astronomy **2**
1.1 Question 1 . . . 2
1.2 Question 2 . . . 15
1.3 Question 3 . . . 25
1.4 Question 4 . . . 26
1.5 Question 5 . . . 27
1.6 Question 6 . . . 32
1.7 Question 7 . . . 34
1.8 Question 8 . . . 36
1.9 Question 9 . . . 38
1.10 Question 10 . . . 46
1.11 Question 11 . . . 47
1.12 Question 12 . . . 51
1.13 Question 13 . . . 52
1.14 Question 14 . . . 62
1.15 Question 15 . . . 69
1.16 Question 16 . . . 77
1.17 Question 17 . . . 78
1.18 Question 18 . . . 84
1.19 Question 19 . . . 85
1.20 Resources . . . 87

1 Galactic Astronomy

1.1 Question 1

What is a stellar Initial Mass Function (IMF)? Sketch it. Give a couple of examples of simple parametric forms used to describe the IMF, such as the Chabrier, Kroupa, or Salpeter functions.

1.1.1 Short answer

Answer.

1.1.2 Additional context

The IMF is perhaps the single most important distribution in stellar and galactic astrophysics. Almost all inferences that go from light to physical properties for unresolved stellar populations rely on an assumed form of the IMF, as do almost all models of galaxy formation and the ISM.
If we sum over the stars formed in a large star-forming region (e.g., the Orion Nebula Cluster), we can discuss the distribution of initial stellar masses – the **initial mass function**, or IMF. Beginning with the pioneering work of Salpeter (1955), there have been many studies of the IMF in different regions of the Milky Way, and in other galaxies. There is no reason to think that the IMF should be universal, yet it shows remarkable uniformity from region to region. There may be systematic variations in the IMF depending on environmental conditions, but the variations are surprisingly small. It is difficult to determine the IMF at the high-mass end because massive stars are rare, and at the low-mass end because low-mass stars are faint. Nevertheless, for 0.01 to $50\,\mathrm{M}_\odot$, there is reasonable agreement between different studies.
Figure 1 shows two recent estimates (Kroupa 2001; Chabrier 2003) for the IMF in the disk of the MW. For $M \gtrsim 1\,\mathrm{M}_\odot$, the observations are consistent with a power law $\mathrm{d}N/\mathrm{d}M \propto M^{-2.3}$, very close to the slope $\mathrm{d}N/\mathrm{d}M \propto M^{-2.35}$ originally found in the pioneering study by Salpeter (1955). The Kroupa and Chabrier estimates for the IMF differ only in detail. While appreciable numbers of low-mass stars are formed, the mass per logarithmic mass interval peaks near $\sim 0.5 - 1\,\mathrm{M}_\odot$. Table 1 provides some useful integral properties of the IMF. For example, for a total star formation rate $\dot{M}$, the rate of formation of $M > 8\,M_\odot$ stars is $\dot{M} \times 0.2118/19.14\,\mathrm{M}_\odot$. If $M > 8\,M_\odot$ stars become Type II supernovae, then the Milky Way star formation rate $\sim 1.3\,\mathrm{M}_\odot\,\mathrm{yr}^{-1}$corresponds to a Type II SN rate $0.014\,\mathrm{yr}^{-1} = 1/70\,\mathrm{yr}$.

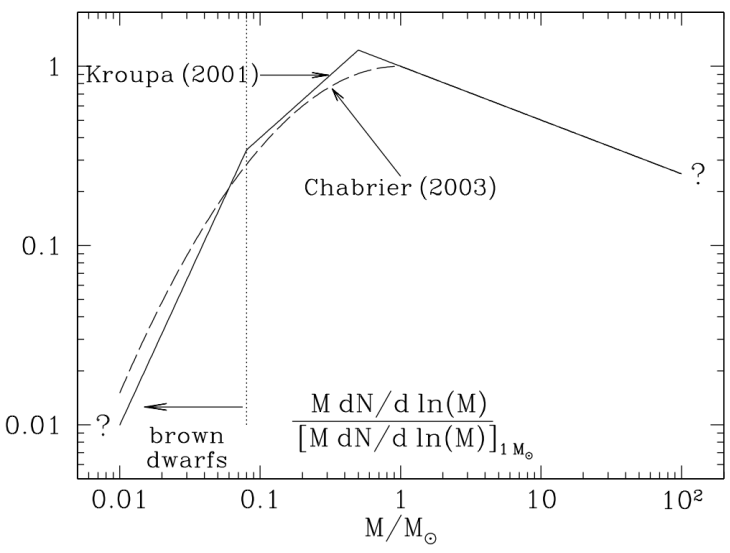

Figure 1: $M\mathrm{d}N/\mathrm{d}lnM$, the mass formed per logarithmic interval in stellar mass M, for IMFs of Kroupa (2001) and Chabrier (2003), normalized to the value at $M = 1\,M_\odot$. Figure taken from Draine (2011).

Since the observable signatures for star formation are obtained only from massive stars, their formation rate needs to be extrapolated to lower masses to obtain the full SFR by assuming an IMF. Typically, a Salpeter-IMF is chosen between $0.1\,M_\odot \leq M \leq 100\,M_\odot$. However, there are clear indications that the IMF may be flatter for $M \lesssim 1\,M_\odot$ than described by the Salpeter law, and several descriptions for such modified IMFs have been developed over the years, mainly based on observations and interpretation of star-forming regions in our MW or in nearby galaxies. The total stellar mass, obtained by integration over the IMF, is up to a factor of ~ 2 lower in these modified IMFs than for the Salpeter IMF. Thus, this factor provides a characteristic uncertainty in the determination of the SFR from observations; a similar, though somewhat smaller uncertainty applies to the stellar mass density whose estimation also

Mass range ($M_\odot$)	mass/total mass	$\langle M \rangle / M_\odot$
0.01–0.08	0.0482	0.0379
0.08–1	0.3950	0.2830
1–8	0.3452	2.156
8–16	0.0749	10.96
16–100	0.1369	32.31
8–100	0.2118	19.14
0.01–100	1.0000	0.3521

[a] For lower and upper cutoffs of 0.01 and 100 $M_\odot$.

Table 1: Some Properties of the Chabrier (2003) IMF[a]. Table taken from Draine (2011).

is mainly based on the more massive stars of a galaxy which dominate the luminosity. Furthermore, the IMF need not be universal, but may in principle vary between different environments, or depend on the metallicity of the gas from which stars are formed. Whereas there has not yet been unambiguous evidence for variations of the IMF, this possibility must always be taken into account.

As with theoretical models of the star formation rate, there is at present no completely satisfactory theory for the origin of the IMF, just different ideas that do better or worse at various aspects of the problem. To recall, the things we would really like to explain most are (1) the slope of the power-law at high masses, and (2) the location of the peak mass. We would also like to explain the little-to-zero variation in these quantities with galactic environment. Furthermore, we would like to explain the origin of the distribution of binary properties.

The power-law tail: Let us begin by considering the power-law tail at high masses, $\mathrm{d}N/\mathrm{d}M \propto M - \alpha$ with $\alpha \approx 2.3$. There are two main classes of theories for how this power-law tail is set: competitive accretion, and turbulence. Both are scale-free processes that could plausibly produce a power-law distribution of masses comparable to what is observed.

Competitive accretion: One hypothesis for how to produce a power-law mass distribution is to consider what will happen in a region where a bunch of small "seed" stars are formed, but then begin to accrete at a rate that is a function of their current mass. Quantitatively, and for simplicity, suppose that every star accretes at a rate proportional to some power of its current mass, i.e.,

$$\frac{dM}{dt} \propto M^\eta \ [\mathrm{M}_\odot \ \mathrm{yr}^{-1}].$$

If we start with a mass M_0 and accretion rate $\dot{M}_0$ at time t_0, this ODE is easy to solve for the mass at later times. We get

$$M(t) = M_o \begin{cases} [1-(\eta-1)\tau]^{1/(1-\eta)} \ [\mathrm{M}_\odot], & \text{if } \eta \neq 1 \\ \exp(\tau) \ [\mathrm{M}_\odot], & \text{if } \eta = 1 \end{cases},$$

where $\tau = t/(M_0/\dot{M}_0)$ is the time measured in units of the initial mass-doubling time. The case for $\eta = 1$ is the usual exponential growth, and the case for $\eta > 1$ is even faster, running away to infinite mass in a finite amount of time $\tau = 1/(\eta - 1)$.

Now suppose that we start with a collection of stars that all begin at mass M_0, but have slightly different values of τ at which they stop growing, corresponding either to growth stopping at different physical times from one star to another, to stars stopping at the same time but having slightly different initial accretion rates $\dot{M}_0$, or some combination of both. What will the mass distribution of the resulting population be? If $dN/d\tau$ is the distribution of stopping times, then we will have

$$\frac{dN}{d\tau} \propto \frac{dN/d\tau}{dM/\tau} M(\tau)^{-\eta} \frac{dN}{d\tau} \ [\text{dimensionless}].$$

Thus the final distribution of masses will be a power-law in mass, with index $-\eta$, going from $M(\tau_{\min})$ to $M(\tau_{\max})$. Thus a power-law distribution naturally results.

The index of this power-law will depend on the index of the accretion law, η. What should this be? In the case of a point mass accreting from a uniform, infinite medium at rest, the accretion rate onto a point mass was worked out by Hoyle; Bondi generalized to the case of a moving medium. In either case, the accretion rate scales as $\dot{M} \propto M^2$, so if this process describes how stars form, then the expected mass distribution should follow $dN/dM \propto M^{-2}$, not so far from the actual slope of -2.3 that we observe. A

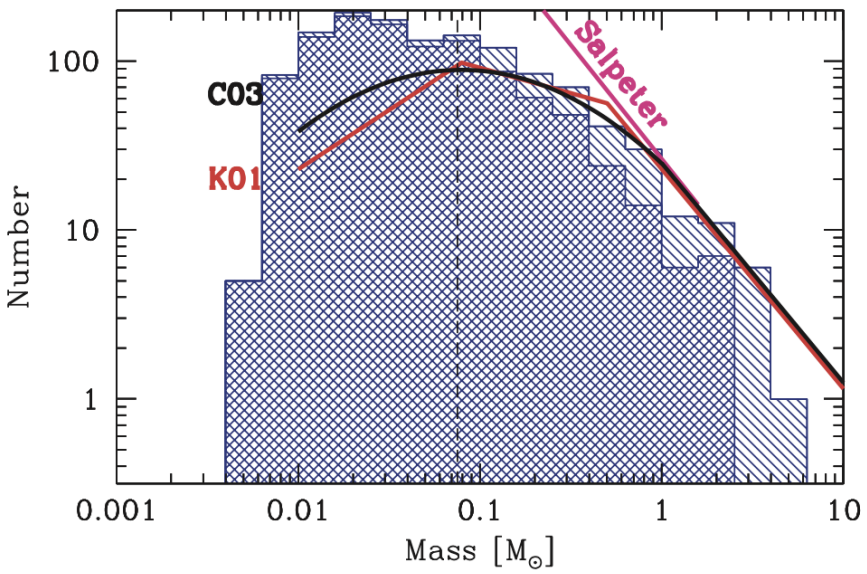

Figure 2: The IMF measured in a simulation of the collapse of a 500 $M_\odot$ initially uniform density cloud (Bate, 2009a). The single-hatched histogram shows all objects in the simulation, while the double-hatched one shows objects that have stopped accreting. Figure taken from Draine (2011).

number of authors have argued that this difference can be made up by considering the effects of a crowded environment, where the feeding regions of smaller stars get tidally truncated, and thus the growth law winds up begin somewhat steeper than $\dot{M} \propto M^2$.

This is an extremely simple model, requiring no physics but hydrodynamics and gravity, and thus it is easy to simulate. Simulations done based on this model do sometimes return a mass distribution that looks much like the IMF, as illustrated in Figure 2. However, this appears to depend on the choice of initial conditions. Generally speaking, one gets about the right IMF if one stars with something with a viral ratio $\alpha_{\rm vir} \sim 1$ and no initial density structure, just velocities. Simulations that start with either super-virial or sub-virial initial conditions, or that begin with turbulent density structures, do not appear to grow as predicted by competitive accretion.

Another potential problem with this model is that it only seems to work in environments where there is no substantial feedback to drive the turbulence or eject the gas. In simulations where this is not true, there appears to be no competitive accretion. The key issue is that competitive accretion seems to require a global collapse where all the stars fall together into a region where they can compete, and this is hard to accomplish in the presence of feedback.

Turbulent fragmentation: A second class of models for the origin of the power-law slope is based on the physics of turbulence. The first of these models was proposed by Padoan et al. (1997), and there have been numerous refinements since. The basic assumption in the turbulence models is that the process of shocks repeatedly passing through an isothermal medium leads to a broad range of density distributions, and that stars form wherever a local region happens to be pushed to the point where it becomes self-gravitating. We then proceed as follows. Suppose we consider the density field smoothed on some size scale ℓ. The mass of an object of density ρ in this smoothed field is

$$M \sim \rho \ell^3 \ [\mathrm{M}_\odot],$$

and the total mass of objects with characteristic density between ρ and $\rho + d\rho$ is

$$dM_{\rm tot} \sim \rho p(\rho) d\rho \ [\mathrm{M}_\odot],$$

where $p(\rho)$ is the density PDF. Then the total number of objects in the mass range from M to $M + dM$ on size scale ℓ can be obtained just by dividing the total mass of objects at a given density by the mass per object, and integrating over the density PDF on that size scale,

$$\frac{dN_\ell}{dM} = \frac{dM_{\rm tot}}{M} \sim \ell^{-3} \int p(\rho) d\rho \ [\text{dimensionless}].$$

Not all of these structures will be bound. To filter out the ones that are, we can impose a density threshold. We assert that an object will be bound only if its gravitational energy exceeds its kinetic energy, that is, only if the density exceeds a critical value given by

$$\frac{GM^2}{\ell} \sim M \sigma_v(\ell)^2 \ [\mathrm{J}] \qquad \Rightarrow \qquad \rho_{\rm crit} \sim \frac{\sigma_v(\ell)^2}{G\ell^2} \ [\mathrm{g\,cm^{-3}}],$$

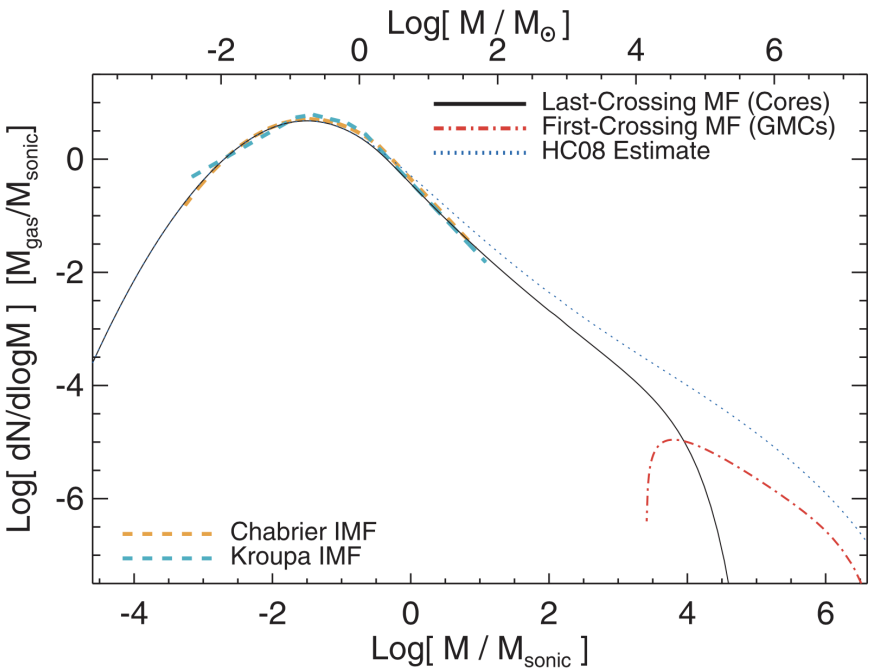

Figure 3: The IMF predicted by an analytic model of turbulent fragmentation by Hopkins (2012). Figure taken from Draine (2011).

where $\sigma_v(\ell)$ is the velocity dispersion on size scale ℓ, which we take from the linewidth-size relation, $\sigma_v(\ell) = c_s\sqrt{(\ell/\ell_s)}$. Thus we have a critical density

$$\rho_{\rm crit} \sim \frac{c_s^2}{G\ell_s\ell}\ [{\rm g\,cm^{-3}}],$$

and this forms a lower limit on the integral.

There are two more steps in the argument. One is simple: just integrate over all length scales to get the total number of objects. That is,

$$\frac{dN}{dM} \propto \frac{dN_\ell}{dM}d\ell\ [{\rm M_\odot^{-1}}].$$

The second is that we must know the functional form of $p(\rho)$ for the smoothed density PDF. One can get this in a couple of different ways, but there isn't a fully rigorous calculation. Hopkins get it by assuming that the PDF is log-normal smoothed on all scales with a dispersion that is an integral over the dispersions on smaller scales. Hennebelle & Chabrier, in their model, assume that the density power spectrum is a power-law, and derive the density PDF from that. The assumptions yield similar but not identical results.

At this point we will simply assert that one can evaluate all the integrals to get an IMF. The result clearly depends only on two dimensional quantities: the sound speed c_s and the sonic length ℓ_s. However, at masses much greater than the sonic mass $M_s \approx c_s^2\ell_s/G$, the result is close to a power-law with approximately the right index. Figure 3 shows an example prediction.

As with the competitive accretion model, this hypothesis encounters certain difficulties. First, there is the technical problem that the choice of smoothed density PDF estimate is not at all rigorous, and there are noticeable differences between on how the choice is made. Second, the dependence on the sonic length is potentially problematic, because real molecular clouds do not really have constant sonic lengths. Regions of massive star formation are observed to be systematically more turbulent.

Third, the theory does not address the question of why gravitationally-bound regions dont sub-fragment as they collapse. Finally, the model has trouble explaining the IMF peak, for the exact same reason as competitive accretion.

The peak of the IMF: A power-law is scale-free, but the peak has a definite mass scale. This mass scale is one basic observable that any theory of star formation must be able to predict. This immediately tells us something about the physical processes that must be involved. We have thus far thought of molecular clouds as consisting mostly of isothermal, turbulent, magnetized, self-gravitating gas. However, we can show that there must be additional processes beyond these at work in setting a peak mass.

We can see this in a few ways. First we'll demonstrate it in a more intuitive but not rigorous manner, and then we can demonstrate it rigorously. The intuitive arguments is as follows. In the system we have described, there are four energies in the problem: thermal energy, bulk kinetic energy, magnetic energy, and gravitational potential energy. From these energies we can define three dimensionless ratios, and the

behavior of the system will be determined by these three ratios. As an example, we might define

$$\mathcal{M} = \frac{\sigma_v}{c_s} \text{ [dimensionless]} \qquad \beta = \frac{8\pi\rho c_c^2}{b^2} \text{ [dimensionless]} \qquad n_J = \frac{\rho L^2}{c_s^3/\sqrt{G^3\rho}} \text{ [dimensionless]}.$$

The ratios describe the ratio of kinetic to thermal energy, the ratio of thermal to magnetic energy, and the ratio of thermal to gravitational energy. (This last quantity is called the Jeans number: it is the ratio of the cloud mass to the Jeans mass.) Other ratios can be derived from these, e.g., the Alfvénic Mach number $\mathcal{M} = \mathcal{M}\sqrt{\beta/2}$ is the ratio of the kinetic to magnetic energy.
Now notice the scalings of these numbers with density ρ, velocity dispersion σ_v, magnetic field strength B, and length scale L:

$$\mathcal{M} \propto \sigma_v \text{ [dimensionless]} \qquad \beta \propto \rho B^{-2} \text{ [dimensionless]} \qquad n_J \propto \rho^{3/2} L^3 \text{ [dimensionless]}.$$

Notice that if we scale the problem by $\rho \to x\rho$, $L \to x^{-1/2}L$, $B \to x^{1/2}B$, all of these dimensionless numbers remain fixed. Thus the behavior of two systems, one with density a factor of x times larger than the other one, length a factor of $x^{-1/2}$ smaller, and magnetic field a factor of $x^{1/2}$ stronger, are simply rescaled versions of one another. If the first system fragments to make a star out of a certain part of its gas, the second system will too. Notice, however, that the masses of those stars will not be the same! The first star will have a mass that scales as ρL^3, while the second will have a mass that scales as $(x\rho)(x^{-1/2}L)^3 = x^{-1/2}\rho L^3$.
We learn from this an important lesson: isothermal gas is scale-free. If we have a model involving only isothermal gas with turbulence, gravity, and magnetic fields, and this model produces stars of a given mass M_*, then we can rescale the system to obtain an arbitrarily different mass. Explaining the IMF peak requires appealing to some physics beyond that of isothermal, magnetized turbulence plus self-gravity. This immediately shows that the competitive accretion and turbulence theories we outlined to explain the power-law tail of the IMF cannot be adequate to explaining the IMF peak, at least not by themselves. Something must be added, and models for the origin of the IMF peak can be broadly classified based on what extra physics they choose to add.
The outer scale of turbulence: One option is hypothesize that the IMF is set at the outer scale of the turbulence, where the molecular clouds join to the atomic ISM (in a galaxy like the MW), or on sizes of the galactic scale-height (for a molecule-dominated galaxy). Something in this outer scale picks out the characteristic mass of stars at the IMF peak.
This hypothesis comes in two flavors. The simplest is that characteristic mass is simply set by the Jeans mass at the mean density of the cloud, so that

$$M_{\rm peak} \propto \frac{c_s^3}{\sqrt{G^3\bar{\rho}}} \text{ [M}_\odot\text{]}.$$

While simple, this hypothesis immediately encounters problems. Molecular clouds have about the same temperature everywhere, but they do not all have the same density – indeed, the density should vary with cloud mass as $M^{1/2}$. Thus at face value this hypothesis would seem to predict a factor of ~ 3 difference in characteristic peak mass between 10^4 and $10^6\,\mathrm{M}_\odot$ clouds in the MW. This is pretty hard to reconcile with observations. The problem is even worse if we think about other galaxies, where the range of density variation is much greater and thus the predicted IMF variation is too. One can hope for a convenient cancellation, whereby an increase in the density is balanced by an increase in temperature, but this seems to require a coincidence.
A somewhat more refined hypothesis, which is adopted by all the turbulence models, is that the IMF peak is set by the sound speed and the normalization of the linewidth-size relation. As discussed above, in the turbulence models the only dimensional free parameters are c_s and ℓ_s, and from them one can derive a mass in only one way:

$$M_{\rm peak} \propto \frac{c_s^2 \ell_s}{G} \text{ [M}_\odot\text{]}.$$

Hopkins calls this quantity the sonic mass, but its the same thing as the characteristic masses in the other models.
This value can be expressed in a few ways. Suppose that we have a cloud of characteristic mass M and radius R. We can write the velocity dispersion in terms of the virial parameter:

$$\alpha_{\rm vir} \sim \frac{\sigma_v R}{GM} \text{ [dimensionless]}.$$

This is the velocity dispersion on the outer scale of the cloud, so we can also define the Mach number on this scale as

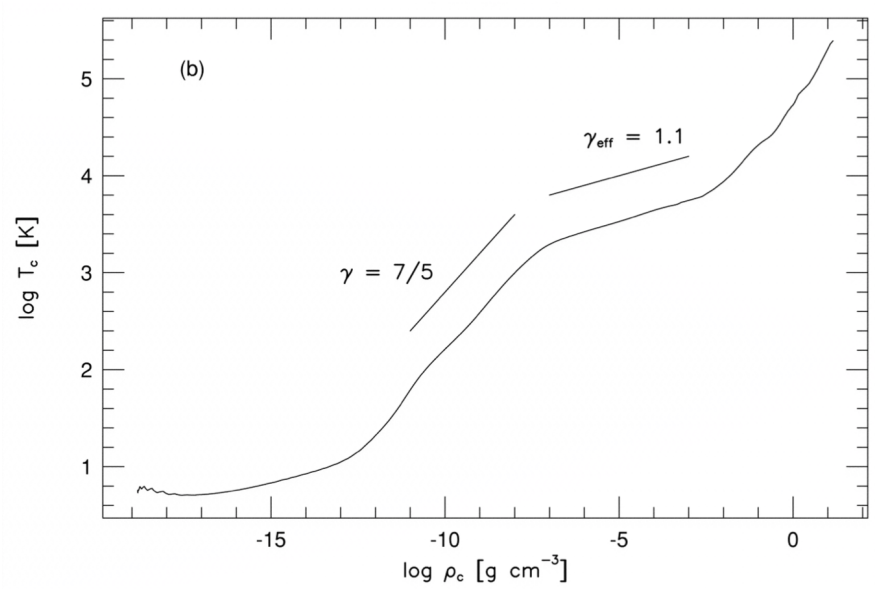

Figure 4: Temperature versus density found in a one-dimensional calculation of the collapse of a $1\,M_\odot$ gas cloud, at the moment immediately before a central protostar forms. Figure taken from Draine (2011).

$$\mathcal{M} = \frac{\sigma_v}{c_s} \sim \sqrt{\alpha_{\rm vir}\frac{GM}{Rc_s^2}}\ \text{[dimensionless]}.$$

The sonic length is just the length scale at which $\mathcal{M} \sim 1$, so if the velocity dispersion scales with $\ell^{1/2}$, then we have

$$\ell_s \sim \frac{R}{\mathcal{M}^2} \sim \frac{c_s^2}{\alpha_{\rm vir}G\Sigma}\ \text{[pc]}.$$

Substituting this in, we have

$$M_{\rm peak} \sim \frac{c_s^4}{\alpha_{\rm vir}G^2\Sigma}\ [\mathrm{M}_\odot],$$

and thus the peak mass simply depends on the surface density of the cloud. We can obtain another equivalent expression by noticing that

$$\frac{M_J}{\mathcal{M}} \sim \frac{c_s^3}{\sqrt{G^3\bar{\rho}}}\sqrt{\frac{Rc_s^2}{\alpha_{\rm vir}GM}} \sim \frac{c_s^4}{\alpha_{\rm vir}G^2\Sigma} \sim M_{\rm peak}\ [\mathrm{M}_\odot].$$

Thus, up to a factor of order unity, this hypothesis is also equivalent to assuming that the characteristic mass is simply the Jeans mass divided by the Mach number.

An appealing aspect of this argument is that it naturally explains why molecular clouds in the MW all make stars at about the same mass. A less appealing result is that it would seem to predict that the masses could be quite different in regions of different surface density, and we observe that there are star-forming regions where Σ is indeed much higher than the mean of the MW GMCs. This is doubly-true if we extend our range to extragalactic environments. One can hope that this will cancel because the temperature will be higher and thus c_s will increase, but this again seems to depend on a lucky cancellation, and there is no a priori reason why it should.

Non-isothermal fragmentation: The alternative to breaking the isothermality at the outer scale of the turbulence is to relax the assumption that the gas is isothermal on small scales. This has the advantage that it avoids any ambiguity about what constitutes the surface density or linewidth-size relation normalization for a "cloud".

The earliest versions of these models were proposed by Larson (2005), and followed up by Jappsen et al. (2005). The basic idea of these models is that the gas in star-forming clouds is only approximately isothermal. Instead, there are small deviations from isothermality, which can pick out preferred mass scales. There are two places where significant deviations from isothermality are expected (Figure 4).

At low density the main heating source is cosmic rays and UV photons, both of which produce a constant heating rate per nucleus if attenuation is not significant. This is because the flux of CRs and UV photons is about constant, and the rate of energy deposition is just proportional to the number of target atoms or dust grains for them to interact with. Cooling is primarily by lines, either of CO once the gas is mostly molecular, or of $\mathrm{C_{II}}$ or O where it is significantly atomic.

In both cases, at low density the gas is slightly below the critical density of the line, so the cooling rate per nucleus or per molecule is an increasing function of density. Since heating per nucleus is constant but cooling per nucleus increases, the equilibrium temperature decreases with density. As one goes to higher density and passes the CO critical density this effect ceases. At that point one generally starts to reach densities such that shielding against UV photons is significant, so the heating rate goes down and thus the temperature continues to drop with density.

This begins to change at a density of around $10^{18}\,g\,cm^{-3}$, $n \sim 10^5 - -10^6\,cm^{-3}$. By this point the gas and dust have been thermally well-coupled by collisions, and the molecular lines are extremely optically thick, so dust is the main thermostat. As long as the gas is optically thin to thermal dust emission, which it is at these densities, the dust cooling rate per molecule is fixed, since the cooling rate just depends on the number of dust grains. Heating at these densities comes primarily from compression as the gas collapses, i.e., it is just PdV work. If the compression were at a constant rate, the heating rate per molecule would be constant. However, the free-fall time decreases with density, so the collapse rate and thus the heating rate per molecule increase with density. The combination of fixed cooling rate and increasing heating rate causes the temperature to begin rising with density. At still higher densities, $10^{13}\,\mathrm{g\,cm^{-3}}$, the gas becomes optically thick to dust thermal emission. At this point the gas simply acts adiabatically, with all the PdV work being retained, so the heating rate with density rises again.

Larson (2005) pointed out that deviations from isothermality are particularly significant for filamentary structures, which dominate in turbulent flows. It is possible to show that a filament cannot go into runaway collapse if T varies with ρ to a positive number, while it can collapse if T varies as ρ to a negative number. This suggests that filaments will collapse indefinitely in the low-density regime, but that their collapse will then halt around $10^{18}\,\mathrm{g\,cm^{-3}}$, forcing them to break up into spheres in order to collapse further. The upshot of all these arguments is that the Jeans or Bonnor-Ebert mass one should be using to estimate the peak of the stellar mass spectrum is the one corresponding to the point where there is a changeover from sub-isothermal to super-isothermal.

In other words, the ρ and T that should be used to evaluate M_J or M_{BE} are the values at that transition point. Larson proposes an approximate equation of state to represent the first break in the EOS: Combining all these effects, Larson (2005) proposed a single simple equation of state

$$T = \begin{cases} 4.4\rho_{18}^{-0.27}\ [\mathrm{K}], & \text{if } \rho_{18} < 1 \\ 4.4\rho_{18}^{0.07}\ [\mathrm{K}], & \text{if } \rho_{18} \geq 1 \end{cases},$$

where $\rho_{18} = \rho/(10^{18}\,\mathrm{g\,cm^{-33}})$. Conveniently enough, the Bonnor-Ebert mass at the minimum temperature here is $M_{BE} = 0.067\,\mathrm{M_\odot}$, which is not too far off from the observed peak of the IMF at $M_{\mathrm{peak}} = 0.2\,\mathrm{M_\odot}$. (The mass at the second break is a bit less promising. At $\rho = 10^{13}\,\mathrm{g\,cm^{13}}$ and $T = 10\,\mathrm{K}$, we have $M_{BE} = 7 \times 10^{-4}\,\mathrm{M_\odot}$).

Simulations done adopting this proposed equation of state seem to verify the conjecture that the characteristic fragment mass does depend critically on the break on the EOS (Figure 5).

While this is a very interesting result, there are two problems. First, the proposed break in the EOS occurs at $n = 4 \times 10^5\,\mathrm{cm^{-3}}$. This is a fairly high density in a low mass star-forming region, but it is actually quite a low density in more typical, massive star-forming regions. For example, the Orion Nebula cluster (ONC) now consists of $4600\,M_\odot$ of stars in a radius of 0.8 pc, giving a mean density $n = 3.7 \times 10^4\,\mathrm{cm^{-3}}$. Since the SF efficiency was less than unity and the cluster is probably expanding due to mass loss, the mean density was almost certainly higher while the stars were still forming. Moreover, recall that, in a turbulent medium, the bulk of the mass is at densities above the volumetric mean density. The upshot of all this is that almost all the gas in Orion was probably over Larson (2005)'s break density while the stars were forming. Since Orion managed to form a normal IMF, it's not clear how the break temperature could be relevant.

A second problem is that, in dense regions like the ONC, the simple model proposed by Larson (2005) is a very bad representation of the true temperature structure, because it ignores the effects of radiative feedback from stars. In dense regions the stars that form will heat the gas around them, raising the temperature. Figure 6 shows the density-temperature distribution of gas in simulations that include radiative transfer, and that have conditions chosen to be similar to those of the ONC.

These two observations suggest that one can build a model for the IMF around radiative feedback. There are a few numerical and analytic papers that attempt to do so, including Bate (2009b, 2012), Krumholz (2011), and Krumholz et al. (2012b). The central idea for these models is that radiative feedback shuts off fragmentation at a characteristic mass scale that sets the peak of the IMF.

The basic idea is as follows. Suppose that we form a first, small protostellar that radiates at a rate L. The temperature of the material at a distance R from it, assuming the gas is optically thick, will be roughly

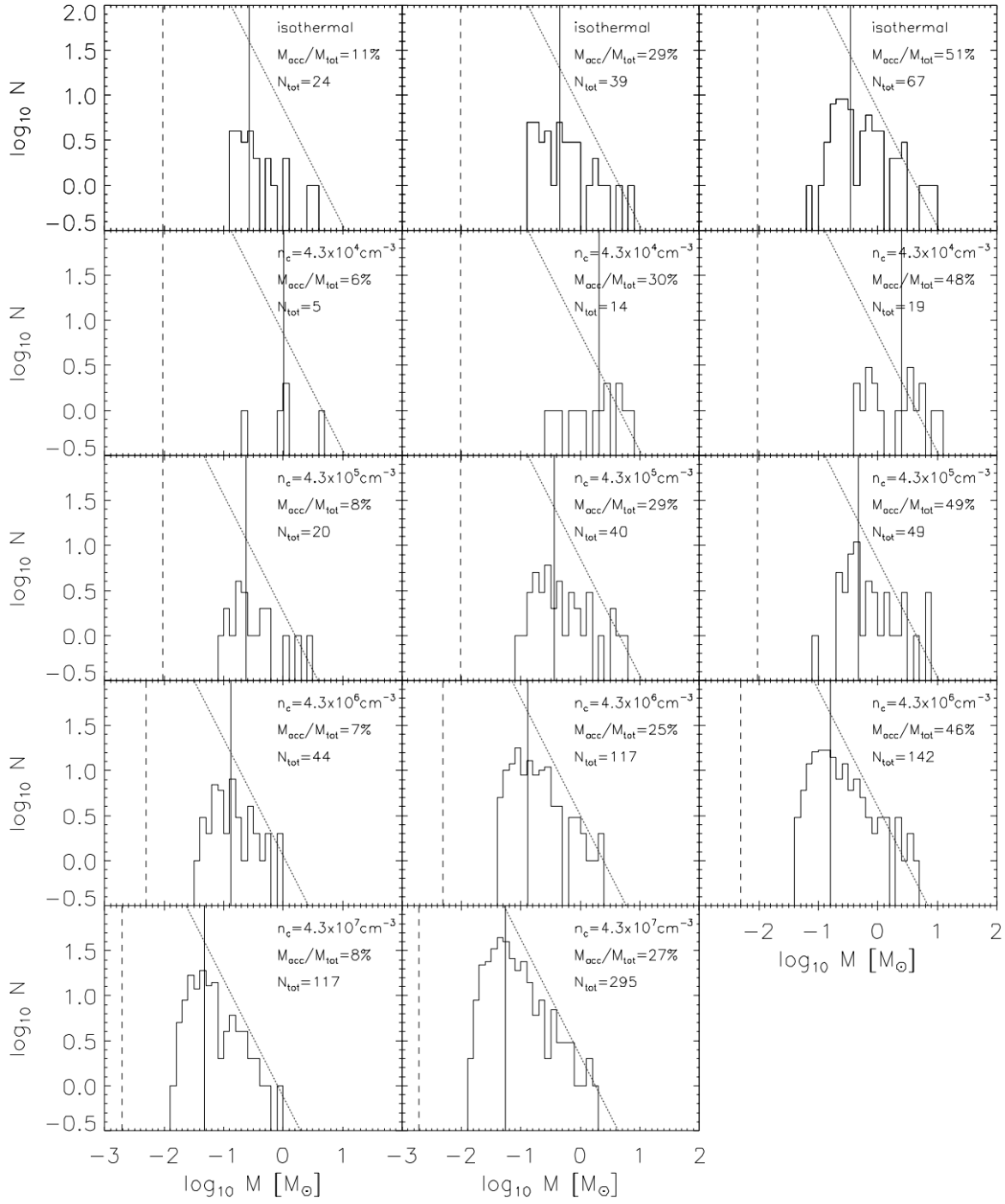

Figure 5: Measured stellar mass distributions in a series of simulations of turbulent fragmentation using non-isothermal EOSs. Each row shows a single simulation, measured at a series of times, characterized by a particular mass in stars as indicated in each panel. Different rows use different EOSs, with the vertical line in each panel indicating the Jeans mass evaluated at the temperature minimum of the equation of state. Histograms show the mass distributions measured for the stars. Figure taken from Draine (2011).

$$L \approx 4\pi R^2 \sigma T^4 \ [\mathrm{erg\,s^{-1}}].$$

Now let us compute the Bonnor-Ebert mass using the temperature T:

$$M_{BE} \approx \frac{c_s^3}{\sqrt{G^3\rho}} = \sqrt{\left(\frac{k_B T}{\mu m_H G}\right)^3 \frac{1}{\rho}},$$

where $\mu = 2.33$ is the mean particle mass, and we are omitting the factor of 1.18 for simplicity. Note that M_{BE} here is a function of R. At small R, T is large and thus M_{BE} is large, while at larger distances the

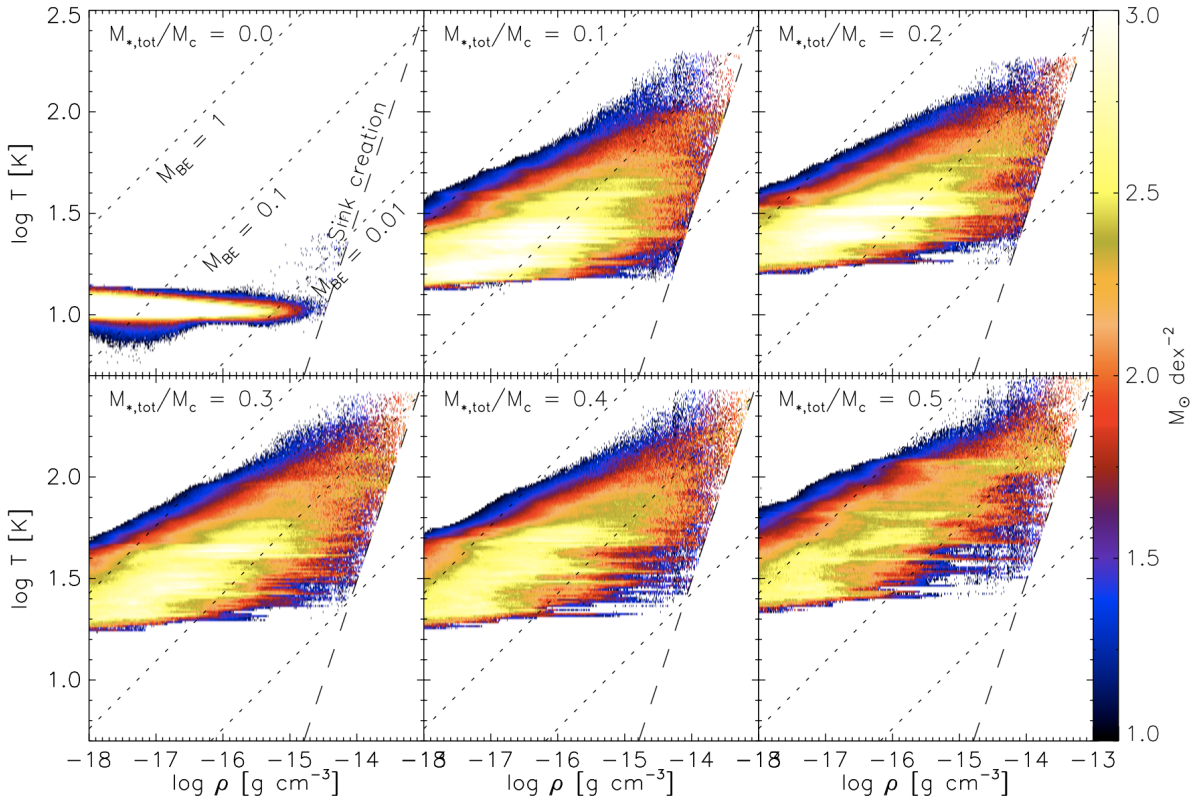

Figure 6: Density-temperature distributions measured from a simulation of the formation of an ONC-like star cluster, including radiative transfer and stellar feedback (Krumholz et al., 2011a). The panels show the distribution at different times in the simulation, characterized by the fraction of mass that has been turned into stars. Doted lines show lines of constant Bonnor-Ebert mass (in $M_\odot$), while dashed lines show the threshold for sink particle formation in the simulation. Histograms show the mass distributions measured for the stars. Figure taken from Draine (2011).

gas is cooler and M_{BE} falls.

Now let us compare this mass to the mass enclosed within the radius R, which is $M = (4/3)\pi R^3 \rho$. At small radii, M_{BE} greatly exceeds the enclosed mass, while at large radii M_{BE} is much less than the enclosed mass. A reasonable hypothesis is that fragmentation will be suppressed out to the point where $M \approx M_{BE}$. If we solve for the radius R and mass M at which this condition is met, we obtain

$$M \approx \left(\frac{1}{36\pi}\right)^{1/10} \left(\frac{k_B}{G\mu m_H}\right)^{6/5} \left(\frac{L}{\sigma}\right)^{3/10} \rho^{-1/5} \; [\mathrm{M}_\odot].$$

To go further, we need to know the luminosity L. The good news is that the luminosity is dominated by accretion, and the energy produced by accretion is simply the accretion rate multiplied by a roughly fixed energy yield per unit mass. In other words, we can write

$$L \approx \phi \dot{M} \; [\mathrm{erg\,s^{-1}}].$$

where $\phi = 10^{14}\,\mathrm{erg\,g^{-1}}$, and can in fact be written in terms of fundamental constants. Taking this on faith for now, if we further assume that stars form over a time of order a free-fall time, then

$$\dot{M} \approx M\sqrt{G\rho} \; [\mathrm{M_\odot\,yr^{-1}}],$$

and substituting this into the equation for M above and solving gives

$$M \approx \left(\frac{1}{36\pi}\right)^{1/7} \left(\frac{k_B}{G\mu m_H}\right)^{12/7} \left(\frac{\phi}{\sigma}\right)^{3/7} \rho^{-1/14} = 0.3 \left(\frac{n}{100\,\mathrm{cm^{-3}}}\right)^{-1/14} \; [\mathrm{M}_\odot],$$

where $n = \rho/(\mu m_H)$. Thus we get a characteristic mass that is a good match to the IMF peak, and that depends only very, very weakling on the ambient density.

Simulations including radiation seem to support the idea that this effect can pick out a characteristic peak ISM mass. The main downside to this hypothesis is that it has little to say by itself about the power-law tail of the IMF. This is not so much a problem with the model as an omission, and a promising area of research seems to be joining a non-isothermal model such as this onto a turbulent fragmentation

or competitive accretion model to explain the full IMF.
Measuring the IMF: There are two major strategies for measuring the IMF. One is to use direct star counts in regions where we can resolve individual stars. The other is to use integrated light from more distant regions where we cannot. The former itself contains two different methods, one using field stars and the other using young clusters.
Field stars (resolved stars): The first attempts to measure the IMF were by Salpeter (1955) (for those counting, nearly 5000 citations as of this writing), using stars in the Solar neighborhood, and the use of Solar neighborhood stars remains one of the main strategies for measuring the IMF today. Suppose that we want to measure the IMF of the field stars within some volume or angular region around the Sun. What steps must we carry out?
The first step is to construct a luminosity function for the stars in our survey volume in one or more photometric bands. This by itself is a non-trivial task, because we require absolute luminosities, which means we require distances. If we are carrying out a volume-limited instead of a flux-limited survey, we also require distances to determine if the target stars are within our survey volume.
The most accurate distances available are from parallax, but this presents a challenge. To measure the IMF, we require a sample of stars that extends down to the lowest masses we wish to measure. As one proceeds to lower masses, the stars very rapidly become dimmer, and as they become dimmer it becomes harder and harder to obtain parallax distances. For $\sim 0.1\,\mathrm{M}_\odot$ stars, typical absolute V band magnitudes are $M_V \sim 14$, and parallax catalogues at such magnitudes are only complete out to $\sim 5 - 10\,\mathrm{pc}$. A survey of this volume only contains $200 - 300$ stars and brown dwarfs, and this sample size presents a fundamental limit on how well the IMF can be measured. If one reduces the mass range being studied, parallax catalogues can go out somewhat further, but then one is trading off sample size against the mass range that the study can probe. Hopefully Gaia will improve this situation significantly.
For these reasons, more recent studies have tended to rely on less accurate spectroscopic or photometric distances. These introduce significant uncertainties in the luminosity function, but they are more than compensated for by the vastly larger number of stars available, which in the most recent studies can be $> 10^6$. The general procedure for photometric distances is to construct color-magnitude (CMD) diagrams in one or more colors for Solar neighborhood stars using the limited sample of stars with measured parallax distances, perhaps aided by theoretical models. Each observed star with an unknown distance is then assigned an absolute magnitude based on its color and the CMD. The absolute magnitude plus the observed magnitude also gives a distance. The spectroscopic parallax method is analogous, except that one uses spectral type - magnitude diagrams (STMD) in place of color-magnitude ones to assign absolute magnitudes. This can be more accurate, but requires at least low resolution spectroscopy instead of simply photometry.
Once that procedure is done, one has in hand an absolute luminosity function, either over a defined volume or (more commonly) a defined absolute magnitude limit. The next step is to correct it for a series of biases. We will not go into the technical details of how the corrections are made, but it is worth going through the list just to understand the issues, and why this is not a trivial task:

- **Metallicity bias**: The reference CMDs or STMDs used to assign absolute magnitudes are constructed from samples very close to the Sun with parallax distances. However, there is a known negative metallicity gradient with height above the galactic plane, so a survey going out to larger distances will have a lower average metallicity than the reference sample. This matters because stars with lower metallicity have higher effective temperature and earlier spectral type than stars of the same mass with lower metallicity. (They have slightly higher absolute luminosity as well, but this is a smaller effect.) As a result, if the CMD or STMD used to assign absolute magnitudes is constructed for Solar metallicity stars, but an actual star being observed is sub-Solar, then we will tend to assign too high an absolute luminosity based on the color, and, when comparing with the observed luminosity, too large a distance. We can correct for this bias if we know the vertical metallicity gradient of the galaxy.

- **Extinction bias**: The reference CMDs/STMDs are constructed for nearby stars, which are systematically less extincted than more distant stars because their light travels through less of the dusty galactic disk. Dust extinction reddens starlight, which causes the more distant stars to be assigned artificially red colors, and thus artificially low magnitudes. This in turn causes their absolute magnitudes and distances to be underestimated, moving stars from their true luminosities to lower values. These effects can be mitigated with knowledge of the shape of the dust extinction curve and estimates of how much extinction there is likely to be as a function of distance.

- **Malmquist bias**: There is some scatter in the magnitudes of stars at fixed color, both due to the intrinsic physical width of the main sequence (e.g., due to varying metallicity, age, stellar rotation) and due to measurement error. Thus at fixed color magnitudes can scatter up or down. Consider

how this affects stars that are near the distance of magnitude limit for the survey: stars whose true magnitude should place them just outside the survey volume or flux limit will be artificially scatter into the survey if they scatter up but not if they scatter down, and those whose true magnitude should place them within the survey will be removed if they scatter to lower magnitude. This asymmetry means that, for stars near the distance or magnitude cutoff of the survey, the errors are not symmetric; they are much more likely to be in the direction of positive than negative flux. This effect is known as Malmquist bias. It can be corrected to the extent that one has a good idea of the size of the scatter in magnitude and understands the survey selection.

- **Binarity**: Many stars are members of binary systems, and all but the most distant of these will be unresolved in the observations and will be mistaken for a single star. This has a number of subtle effects, which we can think of in two limiting cases. If the binary is far from equal mass, say $M2/M1 \sim 0.3$ or less, then the colors and absolute magnitude will not be that different from those of the primary stuff. Thus the main effect is that we do not see the lower mass member of the system at all. We get a reasonable estimate for the properties of the primary, but we miss the secondary entirely, and therefore under-count the number of low luminosity stars. On the other hand, if the mass ratio $M2/M1 \sim 1$ then the main effect is that the color stays about the same, but using our CMD we assign the luminosity of a single star when the true luminosity is actually twice that. We therefore underestimate the distance, and artificially scatter things into the survey (if it is volume limited) or out of the survey (if it is luminosity limited). At intermediate mass ratios, we get a little of both effects. The means of correcting for this, if we have a reasonable estimate of the binary fraction of mass ratio distribution, to guess a true luminosity function, determine which stars are binaries, add them together as they would be added in observations, filter the resulting catalogue through the survey selection, and compare to the observed luminosity function. This procedure is then repeated, adjusting the guessed luminosity function, until the simulated observed luminosity function matches the actually observed one.

Once all these bias corrections are made, the result is a corrected luminosity function that (should) faithfully reproduce the actual luminosity function in the survey volume. Figure 7 shows an example of raw and corrected luminosity functions.

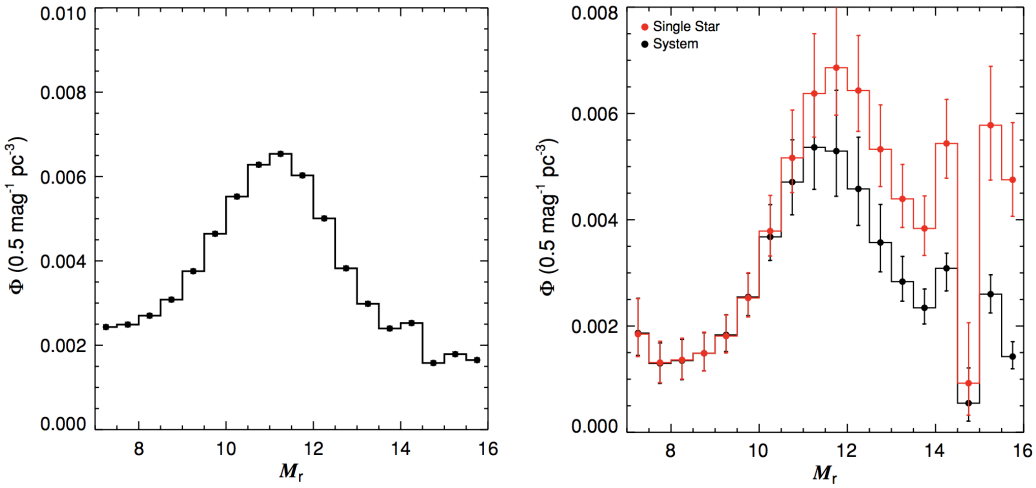

Figure 7: Luminosity function for MW stars before (left) and after (right) bias correction. Figure taken from Draine (2011).

The next step is to convert the luminosity function into a mass function, which requires knowledge of the mass-magnitude relation (MMR) in whatever photometric band we have used for our luminosity function. This must either be determined by theoretical modeling, empirical calibration, or both. Particularly at the low-mass end, the theoretical models tend to have significant uncertainties arising from complex atmospheric chemistry that affects the optical and even NIR colours. For empirical calibrations, the data are only as good as the empirical mass determinations, which must come from orbit modeling. This requires the usual schemes for measuring stellar masses from orbits, e.g., binaries that are both spectroscopic and eclipsing, and thus have known inclinations, or visual binaries with measured radial velocities.
As with the luminosity function, there are a number of possible biases because the stars are not uniform in either age or metallicity, and as a result there is no true single MMR. This would only introduce

a random error if the age and metallicity distribution of the sample used to construct the MMR were the same as that in the IMF survey, but there is no reason to believe that this is actually the case. The selection function used to determine the empirical mass-magnitude sample is complex and poorly characterized, but it is certainly biased towards systems closer to the Sun, for example. Strategies to mitigate this are similar to those used to mitigate the corresponding biases in the luminosity function. Once the mass-magnitude relationship and any bias corrections have been applied, the result is a measure of the field IMF. The results appear to be well-fit by a log-normal distribution or a broken power-law, along the lines of the Chabrier (2005) and Kroupa & Boily (2002) IMFs.

The strategy we have just described works fine for stars up to $\sim 0.7\,\mathrm{M}_\odot$ in mass. However, it fails with higher mass stars, for one obvious reason: stars with masses larger than this can evolve off the main sequence on timescales comparable to the mean stellar age in the Solar neighborhood. Thus the quantity we measure from this procedure is the present-day mass function (PDMF), not the IMF. Even that is somewhat complicated because stars' luminosities start to evolve non-negligibly even before they leave the main sequence, so there are potential errors in assigning masses based on a MMR calibrated from younger stars.

One option in this case is simply to give up and not say anything about the IMF at higher masses. However, there is another option, which is to try to correct for the bias introduced by stellar evolution. Suppose that we think we know both the star formation history of the region we're sampling, $\dot{M}_*(t)$, and the initial mass-dependent main-sequence stellar lifetime, $t_{MS}(M)$. Let dN/dM be the IMF. In this case, the total number of stars formed of the full lifetime of the galaxy in a mass bin from M to $M + dM$ is

$$\frac{dN_{\mathrm{form}}}{dM} = \frac{dN}{dM} \int_{-\infty}^{0} \dot{M}_*(t)\mathrm{d}t \ [\mathrm{M}_\odot^{-1}]$$

where $t = 0$ represents the present. In contrast, the number of stars per unit mass still on the main sequence is

$$\frac{dN_{\mathrm{MS}}}{dM} = \frac{dN}{dM} \int_{-t_{\mathrm{MS}}(M)}^{0} \dot{M}_*(t)\mathrm{d}t \ [\mathrm{M}_\odot^{-1}].$$

Thus if we measure the main sequence mass distribution dN_{MS}/dM, we can correct it to the IMF just by multiplying:

$$\frac{dN}{dM} \propto \left(\frac{dN_{\mathrm{MS}}}{dM}\right) \frac{\int_{-t_{\mathrm{MS}}(M)}^{0} \dot{M}_*(t)\mathrm{d}t}{\int_{-\infty}^{0} \dot{M}_*(t)\mathrm{d}t} \ [\mathrm{M}_\odot^{-1}].$$

This simply reduces to scaling the number of observed stars by the fraction of stars in that mass bin that are still alive today.

Obviously this correction is only as good as our knowledge of the star formation history, and it becomes increasingly uncertain as the correction factor becomes larger. Thus attempts to measure the IMF from the galactic field even with age correction are generally limited to masses of no more than a few $\mathrm{M}_\odot$.

Young clusters (resolved stars): To measure the IMF for more massive stars requires a different technique: surveys of young star clusters. The overall outline of the technique is essentially the same as for the field: construct a luminosity function, correct for biases, then use a mass-magnitude relation to convert to a mass function. However, compared to the field, studying a single cluster offers numerous advantages:

- If the population is young enough, then even the most massive stars will remain on the main sequence, so there is no need to worry about correcting from the PDMF to the IMF. Even for somewhat older clusters, one can probe to higher masses than would be possible with the $\sim 5\,\mathrm{Gyr}$ old field population.
- The stellar population is generally uniform in metallicity or very close to it, so there are no metallicity biases.
- The entire stellar population is at roughly the same distance, so there are no Malmquist or extinction biases. Moreover, in some cases the distance to the cluster is known to better than 10% from radio parallax – some young stars flare in the radio, and with radio interferometry it is possible to obtain parallax measurements at much larger distances than would be possible for the same stars in the optical.

- Low-mass stars and brown dwarfs are significantly more luminous at young ages, and so the same magnitude limit will correspond to a much lower mass limit, making it much easier to probe into the brown dwarf regime.

These advantages also come with some significant costs:

- The statistics are generally much worse than for the field. The most populous young cluster that is close enough for us to resolve individual stars down to the hydrogen burning limit is the ONC, and it contains only $\sim 10^3 - 10^4$ stars, as compared to $\sim 10^6$ for the largest field surveys.
- The MMR that is required to convert an observed magnitude into a mass is much more complex in a young cluster, because a significant fraction of the stars may be pre-main sequence. For such stars, the magnitude is a function not just of the mass but also the age, and one must fit both simultaneously, and with significant theoretical uncertainty. How much of a problem this is depends on the cluster age – for a 100 Myr old cluster like the Pleiades, all the stars have reached the main sequence, while for a $\sim 1-2$ Myr old cluster like Orion, almost none have. However, there is an obvious trade-off here: in a Pleiades-aged cluster, the correction for stars leaving the main sequence is significant, while for an Orion-aged cluster it is negligible.
- For the youngest clusters, there is usually significant dust in the vicinity of the stars, which introduces extinction and reddening that is not the same from star to star. This introduces scatter, and also potentially bias because the extinction may vary with position, and there is a systematic variation between position and mass (see next point).
- Mass segregation can be a problem. In young clusters, the most massive stars are generally found closer to the center – whether this is a result of primordial mass segregation (the stars formed there), dynamical mass segregation (they formed elsewhere but sank to the center), the result is the same. Conversely, low mass stars are preferentially on the cluster outskirts. This means that studies must be extremely careful to measure the IMF over the full cluster, not just its outskirts or core; this can be hard in the cluster center due to problems with crowding. Moreover, if the extinction is not spatially uniform, more massive stars toward the cluster center are likely to suffer systematically more extinction that low-mass ones.
- Dynamical effects can also be a problem. A non-trivial fraction of O and B stars are observed to be moving with very high spatial velocities, above $50\,\mathrm{km\,s^{-1}}$. There are known as runaways. They are likely created by close encounters between massive stars in the core of a newly-formed cluster that lead to some stars being ejected at speeds comparable to the orbital velocities in the encounter. Regardless of the cause, the fact that this happens means that, depending on its age and how many ejections occurred, the cluster may be missing some of its massive stars. Conversely, because low-mass stars are further from the center, if there is any tidal stripping, that will preferentially remove low-mass stars.
- Binary correction is harder for young stars because the binary fraction as a function of mass is much less well known for young clusters than it is for field stars.

Probably the best case for studying a very young cluster is the Orion Nebula Cluster, which is 415 pc from the Sun. Its distance is known to a few percent from radio interferometry. It contains several thousand stars, providing relatively good statistics, and it is young enough that all the stars are still on the main sequence. It is close enough that we can resolve all the stars down to the brown dwarf limit, and even beyond. However, the ONC's most massive star is only $38\,\mathrm{M}_\odot$, so to study the IMF at even higher masses requires the use of more distant clusters within which we can't resolve down to to low masses.
For somewhat older clusters, the best case is almost certainly the Pleiades, which has an age of about 120 Myr. It obviously has no very massive stars left, but there are still $\sim 10\,\mathrm{M}_\odot$ stars present, and it is also close and very well-studied. The IMF inferred for the Pleiades appears to be consistent with that measured in the ONC.

1.1.3 Follow-up Questions

- How did Salpeter determine the IMF?
- How do you normalize the IMF?
- Is the upper or lower limit on mass more important for normalization?

1.2 Question 2

Describe the orbits of stars in a galactic disk and in galactic spheroid.

1.2.1 Short answer

As a first approximation, the stars in the disk move around the Galactic center on circular orbits. However, these orbits are not perfectly circular: besides the orbital velocity (which is about 220 $km\,s^{-1}$ in the Solar vicinity), they have additional random velocity components.

1.2.2 Additional context

Although galaxies are composed of stars, we shall neglect the forces from individual stars and consider only the large-scale forces from the overall mass distribution, which is made up of thousands of millions of stars. In other words, we assume that the gravitational fields of galaxies are smooth, neglecting small-scale irregularities due to individual stars or larger objects like globular clusters or molecular clouds. The gravitational fields of galaxies are sufficiently smooth that these irregularities can affect the orbits of stars only after many crossing times.

Disk galaxies

1. Orbits in symmetric potentials

We first consider orbits in a static, spherically symmetric gravitational field. Such fields are appropriate for globular clusters, which are usually nearly spherical, but, more important, the results we obtain provide an indispensable guide to the behavior of orbits in more general fields.

The motion of a star in a centrally directed gravitational field is greatly simplified by the familiar law of conservation of angular momentum. Thus if

$$\vec{r} = r\hat{e}_r \text{ [pc]}$$

denotes the position vector of the star with respect to the center, and the radial acceleration is

$$\vec{g} = g(r)\hat{e}_r \text{ [m s}^{-2}\text{]},$$

the equation of motion of the star is

$$\frac{\mathrm{d}^2\vec{r}}{\mathrm{d}t^2} = g(r)\hat{e}_r \text{ [m}^2\text{ s}^{-2}\text{]}.$$

If we remember that the cross product of any vector with itself is zero, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\vec{r}\times\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}\right) = \frac{\mathrm{d}\vec{r}}{\mathrm{d}t}\times\frac{\mathrm{d}\vec{r}}{\mathrm{d}t} + \vec{r}\times\frac{\mathrm{d}^2\vec{r}}{\mathrm{d}t^2} = g(r)\vec{r}\times\hat{e}_r = 0.$$

This equation says that $\vec{r}\times\dot{\vec{r}}$ is some constant vector which we will denote $\vec{L}$:

$$\vec{r}\times\frac{\mathrm{d}\vec{r}}{\mathrm{d}t} \equiv \vec{L} \text{ [m}^2\text{ s}^{-1}\text{ kg}^{-1}\text{]}.$$

Of course, $\vec{L}$ is simply the **angular momentum per unit mass**, a vector perpendicular to the plane defined by the star's instantaneous position and velocity vectors. Since this vector is constant, we conclude that the star moves in a plane, the orbital plane. This finding greatly simplifies the determination of the star's orbit, for now that we have established that the star moves in a plane, we may simply use plane polar coordinates (r,ψ) in which the center of attraction is at $r = 0$ and ψ is the **azimuthal angle** in the orbital plane. In terms of these coordinates, the **Lagrangian per unit mass** is

$$\mathcal{L} = \frac{1}{2}[\dot{\vec{r}}^2 + (r\dot{\psi})^2] - \Phi(r) \text{ [J kg}^{-1}\text{]},$$

where $\Phi(r)$ is the **gravitational potential** and $g(r) = \mathrm{d}\Phi/\mathrm{d}r$. The equations of motion are

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\mathcal{L}}{\partial\dot{\vec{r}}} - \frac{\partial\mathcal{L}}{\partial\vec{r}} = \ddot{\vec{r}} - r\dot{\psi}^2 + \frac{\mathrm{d}\Phi}{\mathrm{d}r}$$
$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\mathcal{L}}{\partial\dot{\psi}} - \frac{\partial\mathcal{L}}{\partial\psi} = \frac{\mathrm{d}}{\mathrm{d}t}(r^2\dot{\psi}).$$

The second of these equations implies that

$$r^2\dot{\psi} = \text{constant} \equiv L.$$

It is not hard to show that L is actually the length of the vector $\vec{r} \times \dot{\vec{r}}$, and hence that $r^2\dot{\psi} \equiv L$ is just a restatement of the conservation of angular momentum. Geometrically, L is equal to twice the rate at which the radius vector sweeps out area.
To proceed further we use $r^2\dot{\psi} = \text{constant} \equiv L$ to replace time t by angle ψ as the independent variable in the first EOM. Since the former implies

$$\frac{\mathrm{d}}{\mathrm{d}t} = \frac{L}{r^2}\frac{\mathrm{d}}{\mathrm{d}\psi},$$

the first EOM becomes

$$\frac{L^2}{r^2}\frac{\mathrm{d}}{\mathrm{d}\psi}\left(\frac{1}{r^2}\frac{\mathrm{d}r}{\mathrm{d}\psi}\right) - \frac{L^2}{r^3} = -\frac{\mathrm{d}\Phi}{\mathrm{d}r}.$$

This equation can be simplified by the substitution

$$u \equiv \frac{1}{r},$$

which puts the EOM into the form

$$\frac{\mathrm{d}^2u}{\mathrm{d}\psi^2} + u = \frac{1}{L^2u^2}\frac{\mathrm{d}\Phi}{\mathrm{d}r}\left(\frac{1}{u}\right).$$

The solutions of this equation are of two types: along unbound orbits $r \to \infty$ and hence $u \to 0$, while on bound orbits r and u oscillate between finite limits. Thus each bound orbit is associated with a periodic solution of this equation. We give several analytic examples later in this section, but in general the solutions of this EOM must be obtained numerically.
Some additional insight is gained by deriving a "radial energy" equation from this EOM in much the same way as we can derive the conservation of kinetic plus potential energy; we multiply the EOM by $\mathrm{d}u/\mathrm{d}\psi$ and integrate over ψ to obtain

$$\left(\frac{\mathrm{d}u}{\mathrm{d}\psi}\right)^2 + \frac{2\Phi}{L^2} + u^2 = \text{constant} \equiv \frac{2E}{L^2},$$

where we have used the relation $\mathrm{d}\Phi/\mathrm{d}r = u^2(\mathrm{d}\Phi/\mathrm{d}u)$.
This result can also be derived using Hamiltonians. From the original EOMs we have that the momenta are $p_r = \partial\mathcal{L}/\partial\dot{r} = \dot{r}$ and $p_\psi = \partial\mathcal{L}/\partial\dot{\psi} = r^2\dot{\psi}$, so we find that the **Hamiltonian per unit mass** is

$$\begin{aligned} H(r, p_r, p_\psi) &= p_r\dot{r} + p_\psi\dot{\psi} - \mathcal{L} \ [\mathrm{J\,kg^{-1}}] \\ &= \frac{1}{2}\left(p_r^2 + \frac{p_\psi^2}{r^2}\right) + \Phi(r) \\ &= \frac{1}{2}\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 + \frac{1}{2}\left(r\frac{\mathrm{d}\psi}{\mathrm{d}t}\right)^2 + \Phi(r). \end{aligned}$$

We find that the constant E in this equation is simply the numerical value of the Hamiltonian, which we refer to as the energy of that orbit.
For bound orbits the equation $\mathrm{d}u/\mathrm{d}\psi = 0$ or

$$u^2 + \frac{2[\Phi(1/u) - E]}{L^2} = 0,$$

will normally have two roots u_1 and u_2 between which the star oscillates radially as it revolves in ψ. Thus the orbit is confined between an inner radius $r_1 = u_1^{-1}$, known as the **pericenter** distance, and an outer radius $r_2 = u_2^{-1}$, called the **apocenter** distance. The pericenter and apocenter are equal for a circular orbit. When the apocenter is nearly equal to the pericenter, we say that the orbit has small eccentricity, while if the apocenter is much larger than the pericenter, the eccentricity is said to be near unity. The term "eccentricity" also has a mathematical definition, but only for Kepler orbits.
The radial period T_r is the time required for the star to travel from apocenter to pericenter and back. To determine T_r we use equation $L \equiv r^2\dot{\psi}$ to eliminate $\dot{\psi}$ from the Hamiltonian. We find

$$\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 = 2(E - \Phi) - \frac{L^2}{r^2},$$

which may be rewritten

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \pm\sqrt{2[E - \Phi(r)] - \frac{L^2}{r^2}}.$$

The two possible signs arise because the star moves alternately in and out. Comparing this equation with the equation that has roots u_1 and u_2, we see that $\dot{r} = 0$ at the pericenter and apocenter distances r_1 and r_2, as of course it must. It follows from the last equation that the radial period is

$$T_r = 2\int_{r_1}^{r_2} \frac{\mathrm{d}r}{\sqrt{2[E - \Phi(r)] - \dfrac{L^2}{r^2}}} \text{ [yr]}.$$

In traveling from pericenter to apocenter and back, the azimuthal angle ψ increases by an amount

$$\Delta\psi = 2\int_{r_1}^{r_2} \frac{\mathrm{d}\psi}{\mathrm{d}r}\mathrm{d}r = 2\int_{r_1}^{r_2} \frac{L}{r^2}\frac{\mathrm{d}t}{\mathrm{d}r}\mathrm{d}r \text{ [rad]}.$$

Substituting for $\mathrm{d}t/\mathrm{d}r$ from above, this becomes

$$\Delta\psi = 2L\int_{r_1}^{r_2} \frac{\mathrm{d}r}{r^2\sqrt{2[E - \Phi(r)] - \dfrac{L^2}{r^2}}} \text{ [rad]}.$$

The **azimuthal period** is

$$T_\psi = \frac{2\pi}{|\Delta\psi|}T_r \text{ [yr]};$$

in other words, the mean angular speed of the particle is $2\pi/T_\psi$. In general $\Delta\psi/2\pi$ will not be a rational number. Hence the orbit will not be closed: a typical orbit resembles a rosette and eventually passes close to every point in the annulus between the circles of radii r_1 and r_2 (see Figure 8). There are, however, two and only two potentials in which all bound orbits are closed.

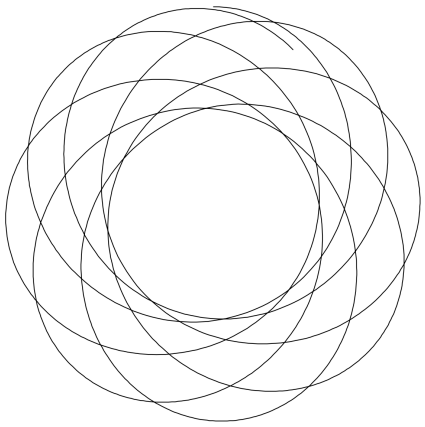

Figure 8: A typical orbit in a spherical potential forms a rosette. Figure taken from Binney & Tremaine (2011).

Spherical harmonic oscillator
We call a potential of the form

$$\Phi(r) = \frac{1}{2}\Omega^2 r^2 + \text{constant [J]}$$

a spherical harmonic oscillator potential. This potential is generated by a homogeneous sphere of matter. Our modified EOM could be solved analytically in this case, but it is simpler to use Cartesian coordinates A(x, y) defined by $x = r\cos\psi$, $y = r\sin\psi$. In these coordinates, the equations of motion are simply

$$\ddot{x} = -\Omega^2 x; \quad \ddot{y} = -\Omega^2 y,$$

with solutions

$$x = X\cos(\Omega t + \epsilon_x); \quad y = Y\cos(\Omega t + \epsilon_y),$$

where X, Y , ϵ_x, and ϵ_y are arbitrary constants. Every orbit is closed since the periods of the oscillations in x and y are identical. The orbits form ellipses centered on the center of attraction. The azimuthal period is $T_\psi = 2\pi/\Omega$ because this is the time required for the star to return to its original azimuth. During this time, the particle completes two in-and-out cycles, so the radial period is

$$T_r = \frac{1}{2}T_\psi = \frac{\pi}{\Omega}\ [\text{yr}].$$

Kepler potential: When the star is acted on by an inverse-square field $g(r) = GM/r2$ due to a point mass M, the corresponding potential is $\Phi = GM/r = GMu$. Motion in this potential is often called **Kepler motion**. Our modified EOM becomes

$$\frac{\mathrm{d}^2u}{\mathrm{d}\psi^2} + u = \frac{GM}{L^2},$$

the general solution of which is

$$u(\psi) = C\cos(\psi - \psi_0) + \frac{GM}{L^2},$$

where $C > 0$ and ψ_0 are arbitrary constants. Defining the orbit's **eccentricity** by

$$e \equiv \frac{CL^2}{GM}\ [\text{dimensionless}],$$

and its **semi-major axis** by

$$q \equiv \frac{L^2}{GM(1-e^2)}\ [\text{AU}],$$

the general solution may now be written as

$$r(\psi) = \frac{a(1-e^2)}{1 + e\cos(\psi - \psi_0)}\ [\text{AU}].$$

An orbit for which $e \geq 1$ is unbound, since $r \to \infty$ as $(\psi\psi_0) \to \pm\cos^{-1}(1/e)$. Bound orbits have $e < 1$ and along them r is a periodic function of ψ with period 2π, so the star returns to its original radial coordinate after exactly one revolution in π. Thus bound Kepler orbits are closed, and one may show that they form ellipses with the attracting center at one focus. The pericenter and apocenter distances are

$$r_1 = a(1-e)\ [\text{AU}]; \quad r_2 = a(1+e)\ [\text{AU}].$$

In many applications, $r(\psi)$ for r along a bound Kepler orbit is less convenient than the parameterization

$$r = a(1 - e\cos\eta)\ [\text{AU}],$$

where the parameter η is called the eccentric anomaly to distinguish it from the true anomaly, $\psi - \psi_0$. By equating $r(\psi)$ with this parameterization and using the identity $\cos\theta = (1\tan^2(\theta/2))/(1+\tan^2(\theta/2))$, it is straightforward to show that the true and eccentric anomalies are related by

$$\sqrt{1-e}\tan\frac{1}{2}(\psi - \psi_0) = \sqrt{1+e}\tan\frac{1}{2}\eta.$$

Taking $t = 0$ to occur at pericenter passage, from $L = r^2\dot{\psi}$ we have

$$t = \int_{\psi_0}^{\psi}\frac{\mathrm{d}\psi}{\dot{\psi}} = \int\frac{r^2}{L}\mathrm{d}\psi = \frac{a^2}{L}\int_0^{\eta}\frac{\mathrm{d}\psi}{\mathrm{d}\eta}(1 - e\cos\eta)^2\mathrm{d}\eta\ [\text{yr}].$$

Evaluating $\mathrm{d}\psi/\mathrm{d}\eta$, integrating, and using trigonometrical identities to simplify the result, we finally obtain

$$t = \frac{a^2}{L}\sqrt{1-e^2}(\eta - e\sin\eta) = \frac{T_r}{2\pi}(\eta - e\sin\eta)\ [\text{yr}],$$

where the second equality follows because the bracket on the right increases by 2π over an orbital period. This is called **Kepler's equation**, and the quantity $2\pi t/T_r$ is sometimes called the **mean anomaly**. Hence

$$T_r = T_\psi = \frac{a^2}{L}\sqrt{1-e^2} = 2\pi\sqrt{\frac{a^3}{GM}}\ [\text{yr}].$$

The energy per unit mass of a particle on a Kepler orbit is

$$E = -\frac{GM}{2a}\ [\text{J kg}^{-1}].$$

To unbind the particle, we thus must add the binding energy E.

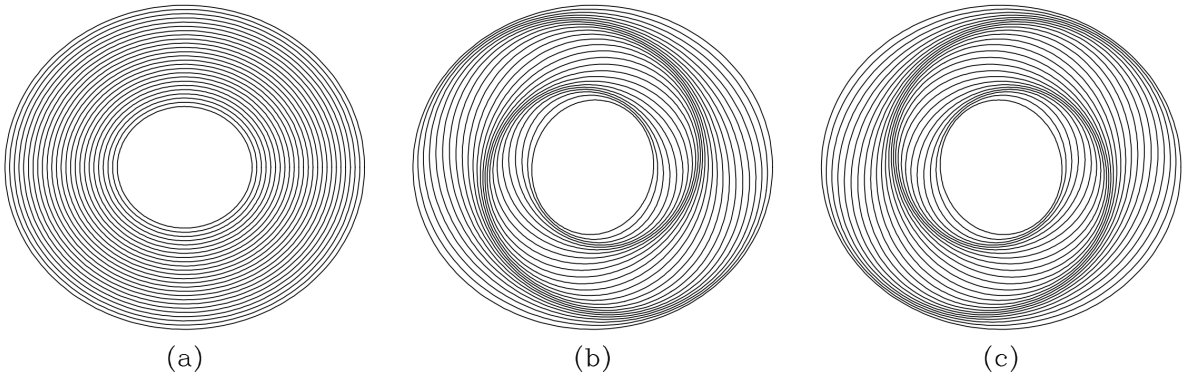

Figure 9: Arrangement of closed orbits in a galaxy to create bars and spiral patterns. Figure taken from Binney & Tremaine (2011).

The study of motion in nearly Kepler potentials is central to the dynamics of planetary systems.
We have shown that a star on a Kepler orbit completes a radial oscillation in the time required for ψ to increase by $\Delta\psi = 2\pi$, whereas a star that orbits in a harmonic-oscillator potential has already completed a radial oscillation by the time ψ has increased by $\Delta\psi = \pi$. Since galaxies are more extended than point masses, and less extended than homogeneous spheres, a typical star in a spherical galaxy completes a radial oscillation after its angular coordinate has increased by an amount that lies somewhere in between these two extremes; $\pi < \Delta\psi < 2\pi$. Thus, we expect a star to oscillate from its apocenter through its pericenter and back in a shorter time than is required for one complete azimuthal cycle about the galactic center.
It is sometimes useful to consider that an orbit in a non-Kepler force field forms an approximate ellipse, though one that precesses by $\psi_p = \Delta\psi 2\pi$ in the time needed for one radial oscillation. For the orbit shown in Figure 8, and most galactic orbits, this precession is in the sense opposite to the rotation of the star itself. The angular velocity Ω_p of the rotating frame in which the ellipse appears closed is

$$\Omega_p = \frac{\psi_p}{T_r} = \frac{\Delta\psi - 2\pi}{T_r} \ [\mathrm{rad\,yr^{-1}}].$$

Hence we say that Ω_p is the **precession rate** of the ellipse. The concept of closed orbits in a rotating frame of reference is crucial to the theory of spiral structure – see Figure 9.

Isochrone potential

The harmonic oscillator and Kepler potentials are both generated by mass distributions that are qualitatively different from the mass distributions of galaxies. The only known potential that could be generated by a realistic stellar system for which all orbits are analytic is the isochrone potential:

$$\Phi(r) = -\frac{GM}{b + \sqrt{b^2 + r^2}} \ [\mathrm{J}].$$

2. Orbits in axisymmetric potentials

Few galaxies are even approximately spherical, but many approximate figures of revolution. Thus we begin to explore the types of orbits that are possible in many real galaxies. We shall usually employ a cylindrical coordinate system (R, ϕ, z) with origin at the galactic center, and shall align the z axis with the galaxy's symmetry axis.
Stars whose motions are confined to the equatorial plane of an axisymmetric galaxy have no way of perceiving that the potential in which they move is not spherically symmetric. Therefore their orbits will be identical with in symmetric potentials; the radial coordinate R of a star on such an orbit oscillates between fixed extrema as the star revolves around the center, and the orbit again forms a rosette figure.

Motion in the meridional plane

The situation is much more complex and interesting for stars whose motions carry them out of the equatorial plane of the system. The study of such general orbits in axisymmetric galaxies can be reduced to a two-dimensional problem by exploiting the conservation of the z-component of angular momentum of any star. Let the potential, which we assume to be symmetric about the plane $z = 0$, be $\Phi(R, z)$. Then the motion is governed by the Lagrangian

$$\mathcal{L} = \frac{1}{2}[\dot{R}^2 + (R\dot{\phi})^2 + \dot{z}^2] - \Phi(R, z) \ [\mathrm{J}].$$

The momenta are

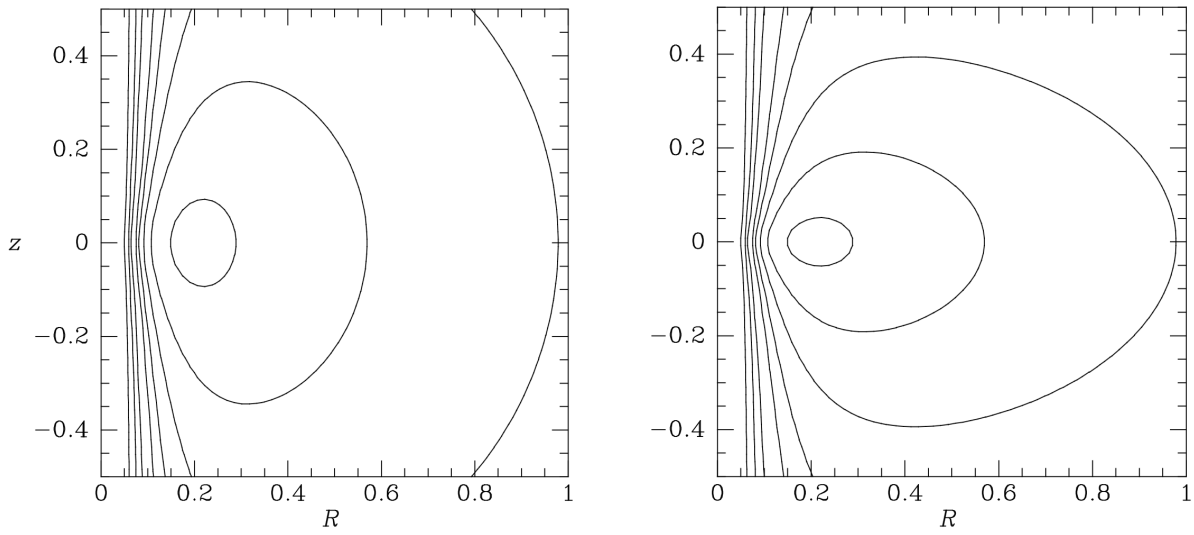

Figure 10: Two orbits in an effective potential $\Phi_{\rm eff}$ with $q = 0.9$. Both orbits are at energy $E = 0.8$ and angular momentum $L_z = 0.2$, and we assume $v_0 = 1$. Figure taken from Binney & Tremaine (2011).

$$p_R = \dot{R}; \quad p_\phi = R^2\dot{\phi}; \quad p_z = \dot{z},$$

so the Hamiltonian is

$$H = \frac{1}{2}\left(P_R^2 + \frac{p_\phi^2}{R^2} + p_z^2\right) + \Phi(R,z) \text{ [J]}.$$

From Hamilton's equations we find that the equations of motion are

$$\dot{p_r} = \ddot{R} = \frac{p_\phi^2}{R^3} - \frac{\partial\Phi}{\partial R}$$
$$\dot{p_\phi} = \frac{\mathrm{d}}{\mathrm{d}t}(R^2\dot{\phi}) = 0$$
$$\dot{p_z} = \ddot{z} = -\frac{\partial\Phi}{\partial z}.$$

The second of these EOMs expresses conservation of the component of angular momentum about the z axis, $p_\phi = L_z$ (a constant), while the first and second describe the coupled oscillations of the star in the R and z-directions.
After replacing p_ϕ in the first of these EOMs by its numerical value L_z, the first and last of these EOMs can be written

$$\ddot{R} = -\frac{\partial\Phi_{\rm eff}}{\partial R}; \quad \ddot{z} = -\frac{\partial\Phi_{\rm eff}}{\partial z},$$

where

$$\Phi_{\rm eff} \equiv \Phi(R,z) + \frac{L_z^2}{2R^2}$$

is called the **effective potential**. Thus the three-dimensional motion of a star in an axisymmetric potential $\Phi(R,z)$ can be reduced to the two-dimensional motion of the star in the (R,z) plane (the meridional plane) under the Hamiltonian

$$H_{\rm eff} = \frac{1}{2}(p_R^2 + p_z^2) + \Phi_{\rm eff}(R,z) \text{ [J]}.$$

Notice that $H_{\rm eff}$ differs from the full Hamiltonian only in the substitution of the constant L_z for the azimuthal momentum p_ϕ. Consequently, the numerical value of $H_{\rm eff}$ is simply the orbit's total energy E. The difference $E\Phi_{\rm eff}$ is the kinetic energy of motion in the (R,z) plane, equal to $(p_R^2 + p_z^2)/2$. Since kinetic energy is non-negative, the orbit is restricted to the area in the meridional plane satisfying the inequality $E \geq \Phi_{\rm eff}$. The curve bounding this area is called the zero-velocity curve, since the orbit can only reach this curve if its velocity in the (R,z) plane is instantaneously zero.
Figure 10 shows contour plots of the effective potential

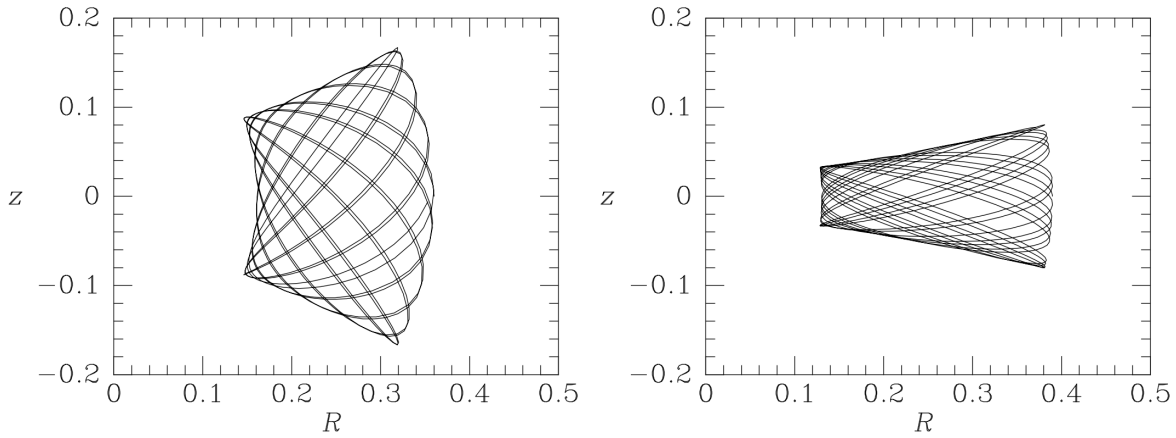

Figure 11: Two orbits in the effective potential Φ_{eff} with $q = 0.9$. Both orbits are at energy $E = 0.8$ and angular momentum $L_z = 0.2$, and we assume $v_0 = 1$. Figure taken from Binney & Tremaine (2011).

$$\Phi_{\text{eff}} = \frac{1}{2}v_0^2 \ln\left(R^2 + \frac{z^2}{q^2}\right) + \frac{L_z^2}{2R^2} \text{ [J]},$$

for $v_0 = 1$, $L_z = 0.2$ and axial ratios $q = 0.9$ and 0.5. This resembles the effective potential experienced by a star in an oblate spheroidal galaxy that has a constant circular speed v_0. Notice that Φ_{eff} rises very steeply near the z axis, as if the axis of symmetry were protected by a **centrifugal barrier**.
The minimum in Φ_{eff} has a simple physical significance. The minimum occurs where

$$\frac{\partial \Phi_{\text{eff}}}{\partial R} = \frac{\partial \Phi}{\partial R} - \frac{L_z^2}{R^3} = 0; \quad \frac{\partial \Phi_{\text{eff}}}{\partial z} = 0.$$

The second of these conditions is satisfied anywhere in the equatorial plane $z = 0$ on account of the assumed symmetry of Φ about this place, and the first is satisfied at the guiding-center radius R_g where

$$\left(\frac{\partial \Phi}{\partial R}\right)_{(R_g,0)} = \frac{L_z^2}{R_g^3} = R_g\dot{\phi}^2.$$

This is simply the condition for a circular orbit with angular speed $\dot{\phi}$. Thus the minimum of Φ_{eff} occurs at the radius at which a circular orbit has angular momentum L_z, and the value of Φ_{eff} at the minimum is the energy of this circular orbit.
Unless the gravitational potential Φ is of some special form, the EOMs for $\ddot{R}$ and $\ddot{z}$ cannot be solved analytically. However, we may follow the evolution of $R(t)$ and $z(t)$ by integrating them numerically, starting from a variety of initial conditions. Figure 11 shows the result of two such integrations for the effective potential Φ_{eff} with $q = 0.9$. The orbits shown are of stars of the same energy and angular momentum, yet they look quite different in real space, and hence the stars on these orbits must move through different regions of phase space. Is this because the equations of motion admit a third isolating integral $I(R, z, p_R, p_z)$ in addition to E and L_z?
Surfaces of section
The phase space associated with the motion we are considering has four dimensions, R, z, p_R, and p_z, and the four-dimensional motion of the phase-space point of an individual star is too complicated to visualize. Nonetheless, we can determine whether orbits in the (R, z) plane admit an additional isolating integral by use of a simple graphical device. Since the Hamiltonian $H_{\text{eff}}(R, z, p_R, p_z)$ is constant, we could plot the motion of the representative point in a three-dimensional reduced phase space, say (R, z, p_R), and then p_z would be determined (to within a sign) by the known value E of H_{eff}. However, even three-dimensional spaces are difficult to draw, so we simply show the points where the star crosses some plane in the reduced phase space, say the plane $z = 0$; these points are called **consequents**. To remove the sign ambiguity in p_z, we plot the (R, p_R) coordinates only when $p_z > 0$. In other words, we plot the values of R and p_R every time the star crosses the equator going upward. Such plots were first used by Poincaré and are called **surfaces of section**. The key feature of the surface of section is that, even though it is only two-dimensional, no two distinct orbits at the same energy can occupy the same point. Also, any orbit is restricted to an area in the surface of section defined by the constraint $Heff \geq (\dot{R}^2 + \Phi_{\text{eff}})/2$; the curve bounding this area is often called the zero-velocity curve of the surface of section, since it can only be reached by an orbit with $p_z = 0$.

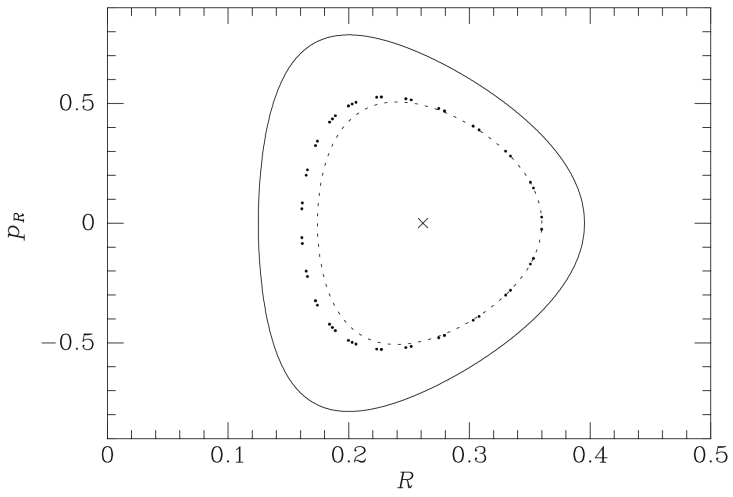

Figure 12: Points generated by the orbit of the left panel of Figure 11 in the (R, p_R) surface of section. If the total angular momentum L of the orbit were conserved, the points would fall on the dashed curve. The full curve is the zero-velocity curve at the energy of this orbit. The × marks the consequent of the shell orbit. Figure taken from Binney & Tremaine (2011).

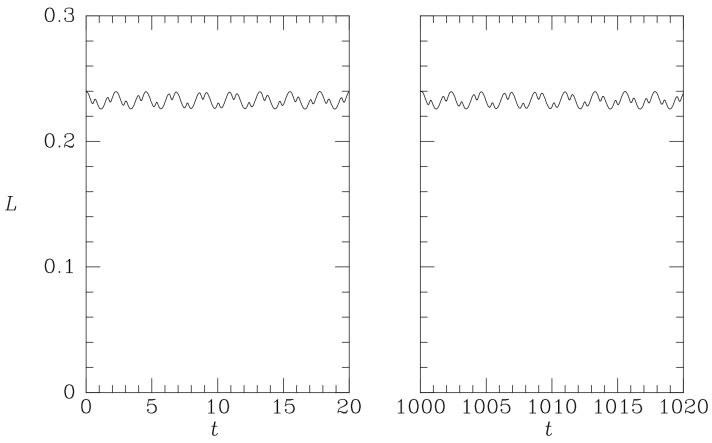

Figure 13: The total angular momentum is almost constant along the orbit shown in the left panel of Figure 11. For clarity $L(t)$ is plotted only at the beginning and end of a long integration. Figure taken from Binney & Tremaine (2011).

Figure 12 shows the (R, p_R) surface of section at the energy of the orbits of Figure 11: the full curve is the zero-velocity curve, while the dots show the consequents generated by the orbit in the left panel of Figure 11. The cross near the center of the surface of section, at $(R = 0.26, p_R = 0)$, is the single consequent of the shell orbit, in which the trajectory of the star is restricted to a two-dimensional surface. The shell orbit is the limit of orbits such as those shown in Figure 11 in which the distance between the inner and outer boundaries of the orbit shrinks to zero.

In Figure 12 the consequents of the orbit of the left panel of Figure 11 appear to lie on a smooth curve, called the invariant curve of the orbit. The existence of the invariant curve implies that some isolating integral I is respected by this orbit. The curve arises because the equation $I =$ constant restricts motion in the two-dimensional surface of section to a one-dimensional curve (or perhaps to a finite number of discrete points in exceptional cases). It is often found that for realistic galactic potentials, orbits do admit an integral of this type. Since I is in addition to the two classical integrals H and p_ϕ, it is called the **third integral**. In general there is no analytic expression for I as a function of the phase-space variables, so it is called a **non-classical integral**.

We may form an intuitive picture of the nature of the third integral by considering two special cases. If the potential Φ is spherical, we know that the total angular momentum $|\vec{L}|$ is an integral. This suggests that for a nearly spherical potential (this one has axis ratio $q = 0.9$) the third integral may be approximated by $|\vec{L}|$. The dashed curve in Figure 12 shows the curve on which the points generated by the orbit of the left panel of Figure 11 would lie if the third integral were $|\vec{L}|$, and Figure 13 shows the actual time evolution of $|\vec{L}|$ along that orbit – notice that although $|\vec{L}|$ oscillates rapidly, its mean value does not change even over hundreds of orbital times. From these two figures we see that $|\vec{L}|$ is an approximately conserved quantity, even for orbits in potentials that are significantly flattened. We may think of these orbits as approximately planar and with more or less fixed peri- and apocenter radii. The approximate orbital planes have a fixed inclination to the z axis but precess about this axis, at a rate that gradually tends to zero as the potential becomes more and more nearly spherical.

The second special case is when the potential is separable in R and z:

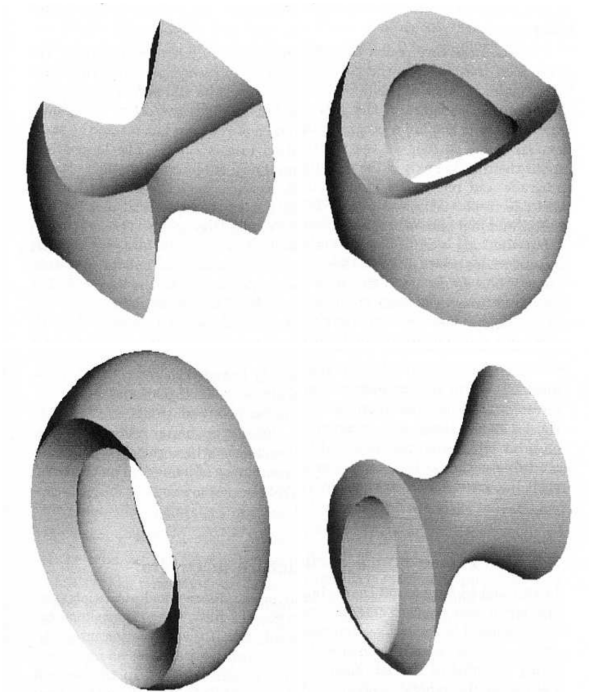

Figure 14: Orbits in a non-rotating triaxial potential. Clockwise from top left: (a) box orbit; (b) short-axis tube orbit; (c) inner long-axis tube orbit; (d) outer long-axis tube orbit. From Statler (1987), by permission of the AAS. Figure taken from Binney & Tremaine (2011).

$$\Phi(R,z) = \Phi_R(R) + \Phi_z(z) \text{ [J]}.$$

Then the third integral can be taken to be the energy of vertical motion

$$H_z = \frac{1}{2}p_z^2 + \Phi_z(z) \text{ [J]}.$$

Along nearly circular orbits in a thin disk, the potential is approximately separable, so this equation provides a useful expression for the third integral.

Elliptical galaxies

Elliptical galaxies nearly always have cusps in their central density profiles in which $\rho \sim r^{-\alpha}$ with $0.3 \lesssim \alpha \lesssim 2$. Black holes with masses $\sim 0.2\%$ of the mass of the visible galaxy are believed to reside at the centers of these cusps. Further out the mass distributions of many elliptical galaxies are thought to be triaxial. These features make the orbital dynamics of elliptical dynamics especially rich.

A useful basic model of the orbital dynamics of a triaxial elliptical galaxy is provided by extensions to three dimensions of the two-dimensional Stäckel potentials. The simplest three-dimensional system that generates a Stäckel potential through Poisson's equation is the **perfect ellipsoid**, in which the density is given by

$$\rho(x) = \frac{\rho_0}{(1+m^2)^2}$$

where

$$m^2 \equiv \frac{x^2 + (y/q_1)^2 + (z/q_2)^2}{a_0^2}.$$

In this formula q_1 and q_2 are the axis ratios of the ellipsoidal surfaces of constant density, and a_0 is a scale length. At radii significantly smaller than a_0, the density is approximately constant, while at $r \gg a_0$ the density falls off $\propto r^{4-}$. Since these asymptotic forms differ from those characteristic of elliptical galaxies, we have to expect the orbital structures of real galaxies to differ in detail from that of the perfect ellipsoid, but nevertheless the model exhibits much of the orbital structure seen in real elliptical galaxies.

By an analysis similar to that used to explore the potential of a planar bar, one can show that the perfect ellipsoid supports four types of orbits. Figure 3.46 depicts an orbit of each type. At top left we have a box orbit. The key feature of a box orbit is that it touches the isopotential surface for its energy at its eight corners. Consequently, the star comes to rest for an instant at these points; a box orbit is conveniently generated numerically by releasing a star from rest on the equipotential surface. The potential's longest

axis emerges from the orbits convex face. The other three orbits are all tube orbits: stars on these orbits circulate in a fixed sense around the hole through the orbit's center, and are never at rest. The most important tube orbits are the short-axis loops shown at top right, which circulate around the potential's shortest axis. These orbits are mildly distorted versions of the orbits that dominate the phase space of a flattened axisymmetric potential. The tube orbits at the bottom of Figure 3.46 are called outer (left) and inner long-axis tube orbits, and circulate around the longest axis of the potential. Tube orbits around the intermediate axis are unstable. All these orbits can be quantified by a single system of angle-action coordinates $(J_\lambda, J_\mu, J_\nu)$ that are generalizations of the angle-action coordinates for spherical potentials (J_r, J_θ, J_ϕ).

- If we perturb a star in the disk in the z-direction, what happens?
- If we perturb a star in the disk in the radial-direction, what happens?
- What are the observed quantities in each scenario?
- How many integrals of motion are there in the disk?
- What symmetry leads to energy conservation?

1.3 Question 3

Every now and then a supernova explosion occurs within 3 pc of the Earth. Estimate how long one typically has to wait for this to happen. Why are newborn stars likely to experience this even when they are much younger than the waiting time you have just estimated?

1.3.1 Short answer

Answer.

1.3.2 Additional context

Additional context.

1.4 Question 4

Galactic stars are described as a collision-less system. Why? (Dont forget the influence of gravity.)

1.4.1 Short answer

Answer.

1.4.2 Additional context

Additional context.

1.4.3 Follow-up Questions

- What happens when stars collide?
- Why choose a cross-section that's larger than the star's radius?
- What impact parameter do we need for the stars to end up physically touching (calculate it)?

1.5 Question 5

Given that only a tiny fraction of the mass of the interstellar medium consists of dust, why is dust important to the chemistry of the medium and to the formation of stars?

1.5.1 Short answer

Dust is understood to play many critical roles in galactic evolution. By sequestering selected elements in the solid grains, and by catalyzing formation of the H_2 molecule, dust grains are central to the chemistry of interstellar gas. Photoelectrons from dust grains can dominate the heating of gas in regions where UV starlight is present, and in dense regions the infrared emission from dust can be an important cooling mechanism. Last, dust grains can be important in interstellar gas dynamics, communicating radiation pressure from starlight to the gas, and coupling the magnetic field to the gas in regions of low fractional ionization.

1.5.2 Additional context

Our strongest constraints on interstellar dust come from observations of its interaction with electromagnetic radiation:

- Wavelength-dependent attenuation ("extinction") of starlight by absorption and scattering, now observable at wavelengths as long as $20\,\mu$m ("mid-infrared"), and as short as $0.1\,\mu$m ("vacuum ultraviolet"). The extinction includes a number of spectral features that provide clues to grain composition.
- Polarization-dependent attenuation of starlight, resulting in wavelength-dependent polarization of light reaching us from reddened stars.
- Scattered light in reflection nebulae.
- Thermal emission from dust, at wavelengths ranging from the sub-mm to $2\,\mu$m.
- Small-angle scattering of x-rays, resulting in "scattered halos" around x-ray point sources.
- Microwave emission from dust, probably from rapidly spinning ultra-small grains.
- Luminescence when dust is illuminated by starlight – the so-called extended red emission.

In addition to these electromagnetic studies, our knowledge of dust is also informed by other, less direct, evidence:

- Pre-solar grains preserved in meteorites – a selective but not well-understood sampling of the interstellar grains that were present in the Solar nebula 4.5 Gyr ago.
- "Depletion" of certain elements from the interstellar gas, with the missing atoms presumed to be contained in dust grains.
- The observed abundance of H_2 in the ISM, which can only be understood if catalysis on dust grains is the dominant formation avenue.
- The temperature of interstellar diffuse HI and H_2, in part a result of heating by photo-electrons ejected from interstellar grains.

Interstellar extinction:
Starlight polarization: The polarization of starlight was discovered serendipitously in 1949. When it was realized that the degree of polarization tended to be larger for stars with greater reddening, and that stars in a given region of the sky tended to have similar polarization directions, it became obvious that the polarization is produced by the ISM: initially upolarized light propagating through the ISM becomes linearly polarized as a result of preferential extinction of one linear polarization mode relative to the other. Figure 15 shows the direction of polarization and the strength of polarization for 5453 stars with galactic latitudes b between $-80°$ and $+80°$. The large-scale organization of the polarization vectors can be understood if dust grains are somehow aligned by the interstellar magnetic field.
The polarization percentage typically peaks near the V band ($5,500$Å), and can be empirically described by the "Serkowski law":

$$p(\lambda) \approx p_{\max} \exp\left[-K \ln^2\left(\frac{\lambda}{\lambda_{\max}}\right)\right] \text{ [dimensionless]}$$

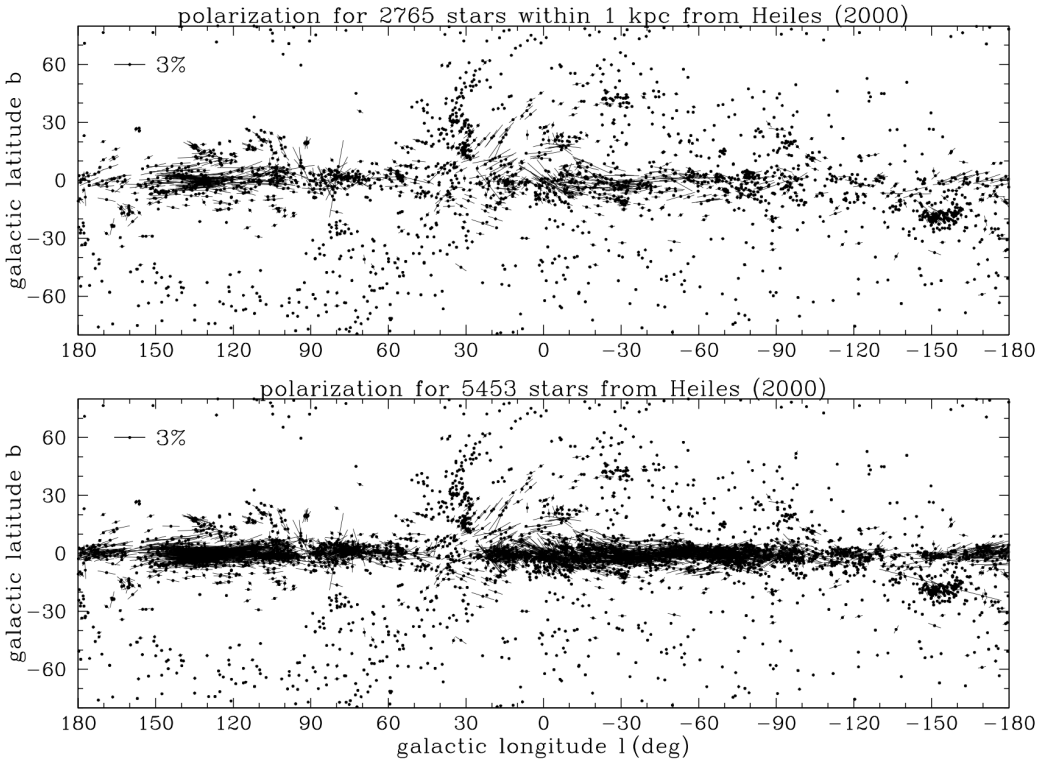

Figure 15: Linear polarization of starlight plotted in galactic coordinates, for stars within 1 kpc, and for all stars in the catalogue of Heiles (2000). The length of each line segment is proportional to the degree of polarization. Figure taken from Draine (2011).

with $\lambda_{\max} \approx 5500$Å and $K \approx 1.15$. The peak polarization $p_{\max}$ is found to fall within an envelope

$$0 < p_{\max} \leq 0.09 \left[\frac{E(B-V)}{\text{mag}}\right] \approx 0.03 \left[\frac{A_V}{\text{mag}}\right] \text{ [dimensionless]}$$
$$0 < p_V \lesssim 0.03\tau_V \text{ [dimensionless]}.$$

The polarization is produced by dust grains that are somehow partially aligned by the interstellar magnetic field. It appears that the grains are aligned with their shortest axes tending to be parallel to the magnetic field direction. The largest values of $p_{\max}/E(B-V)$ are presumed to arise on sightlines where the magnetic field is uniform and perpendicular to the line of sight. While the Serkowski law was originally put forward as an empirical fit to the observed polarization at $0.3\,\mu\text{m} \lesssim \lambda \lesssim 1\,\mu\text{m}$, it turns out to give a surprisingly good approximation to the measured linear polarization in the vacuum UV, although there are some sightlines where the Serkowski law under-predicts the UV polarization, and one sightline where the 2175Å feature appears to be weakly polarized.
The mechanism responsible for the grain alignment remains a fascinating puzzle . Independent of the grain alignment mechanism, however, we can infer the sizes of the interstellar grains responsible for this polarization by noting that the extinction rises rapidly into the UV whereas the polarization declines . This can be understood if the grains responsible for the polarization have diameters $2a$ such that $a \approx (\lambda_{\max}/2\pi) \approx 0.1\,\mu$m: as one proceeds into the UV, one moves toward the "geometric optics" limit where both polarization modes suffer the same extinction, so the polarization goes to zero. Thus we conclude that:

- The extinction at $\lambda \approx 0.55\,\mu$m has an appreciable contribution from grains with sizes $a \approx 0.1\,\mu$m. These grains are non-spherical and substantially aligned.
- The grains with $a \lesssim 0.05\,\mu$m, which dominate the extinction at $\lambda \lesssim 0.3\,\mu$m, are either spherical (which seems unlikely) or minimally aligned.

Scattering of Starlight: When an interstellar cloud happens to be unusually near one or more bright stars, we have a reflection nebula, where we see starlight photons that have been scattered by the dust in the cloud. The spectrum of the light coming from the cloud surface shows the stellar absorption lines, thus demonstrating that scattering rather than some emission process is responsible. By comparing the observed scattered intensity with the estimated intensity of the starlight incident on the cloud, it is

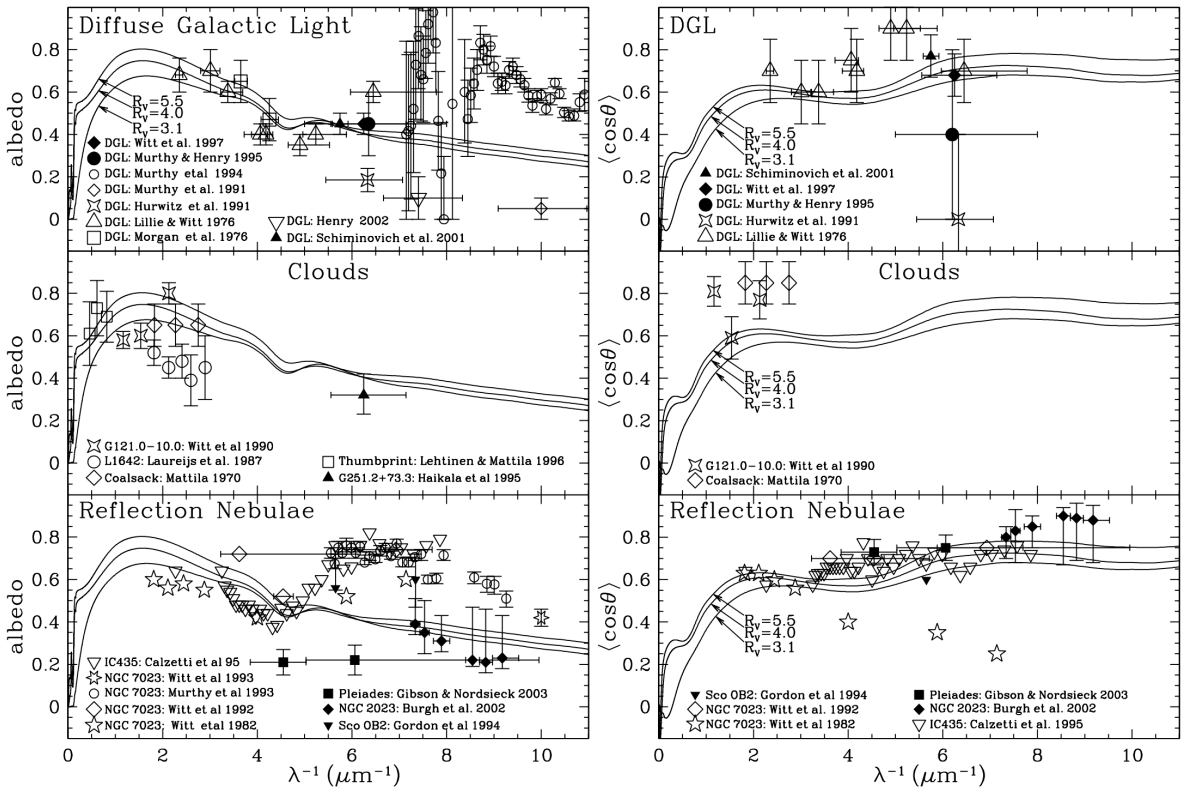

Figure 16: Albedo and scattering asymmetry factor $\langle\cos\theta\rangle$ inferred from observations of the diffuse galactic light, reflection nebulae, and dark clouds. Figure taken from Draine (2011).

possible to infer the albedo ω of the dust – the ratio of scattering cross section to extinction cross section. It is also possible to infer $\langle\cos\theta\rangle$ for the dust, where θ is the scattering angle.

Figure 16 shows ω and $\langle\cos\theta\rangle$ for (1) the dust in the general diffuse ISM producing the "diffuse galactic light," (2) dust in individual clouds illuminated by the general starlight, and (3) dust in clouds that are illuminated by a nearby bright star. In the optical, the interstellar dust mixture has an albedo $\omega \approx 0.5$ (scattering is about as important as absorption) and the grains are somewhat forward-scattering, with $\langle\cos\theta\rangle \approx 0.5$. Rayleigh scattering by particles small compared to the wavelength has $\langle\cos\theta\rangle \approx 0$, so this tells us that the particles dominating the scattering at $\lambda \approx 0.6\,\mu$m have $a \gtrsim \lambda/2\pi \approx 0.1\,\mu$m.

IR emission: Dust grains are heated by starlight, and cool by radiating in the IR. The emission from dust at high galactic latitudes has been studied by a number of satellites. Figure 17 shows the emission spectrum from $800\,\mu$m to $3\,\mu$m. The 3 to $12\,\mu$m spectrum is estimated from observations of the Galactic plane near $l \approx 45°$, if we assume that the ratio of 3 to $12\,\mu$m emission to the $100\,\mu$m emission is unchanged in going from observations of the Galactic plane to high galactic latitudes. The correlation of the IR emission with HI 21 cm emission at high latitudes is used to estimate the power radiated per H nucleon: $5.0 \times 10^{-24}\,\mathrm{erg\,s^{-1}\,H^{-1}}$.

Interstellar dust is heated primarily by starlight, and the total power radiated requires, therefore, that the absorption cross section of interstellar dust be such that the power absorbed per H (for the estimated spectrum of the starlight heating the dust) match the observed emission, $5.0 \times 10^{-24}\,\mathrm{erg\,s^{-1}\,H^{-1}}$. The IR spectrum provides very strong constraints on grain models, as the dust must include a component that can account for the fact that $\sim 35\%$ of the radiated power is short-ward of $50\,\mu$m, including the strong emission features at $\sim 12\,\mu$m and $6 - 8\,\mu$m.

Luminescence: The energy absorbed by dust grains is primarily reradiated in the mid- and far-IR, but there is evidence that dust grains also emit light at optical and near-IR wavelengths. Studies of reflection nebulae indicate that there is more light emerging at wavelengths $6000 - 8000$Å than can be accounted for by scattering alone, and this excess is ascribed to luminescence from dust grains following absorption of shorter-wavelength photons. Luminescence at $6000 - 8000$Å is also termed "extended red emission," or ERE. Luminescence in the blue has also been reported. Candidate materials to explain this luminescence must of course reproduce the observed luminescence spectrum. The luminescing materials have not yet been conclusively identified. The blue luminescence may be produced by neutral PAHs, and PAH di-cations ($\mathrm{PAH^{++}}$) may be responsible for the ERE.

H_2 formation: Molecular hydrogen is a lower energy state than atomic hydrogen, so an isolated box

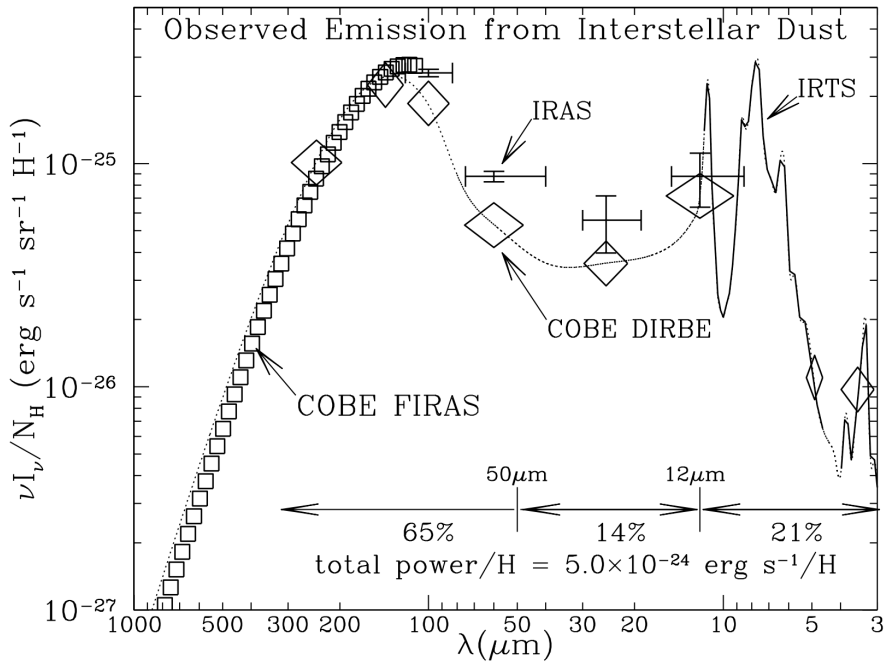

Figure 17: Observed IR emission per H nucleon from dust heated by the average starlight background in the local MW. Crosses: IRAS; squares: COBE-FIRAS; diamonds: COBE-DIRBE; heavy curve: IRTS;. The interpolated dotted line is used to estimate the total power. Figure taken from Draine (2011).

of hydrogen left for an infinite amount of time will eventually become predominantly molecular. In interstellar space, though, the atomic versus molecular fraction in a gas is determined by a balance between formation and destruction processes.

Atomic hydrogen can turn into molecular hydrogen in the gas phase, but this process is extremely slow. This is ultimately due to the symmetry of the hydrogen molecule. To form an H_2 molecule, two H atoms must collide and then undergo a radiative transition that removes enough energy to leave the resulting pair of atoms in a bound state. However, two H atoms that are both in the ground state constitute a symmetric system, as does an H_2 molecule in its ground state. Because both the initial and final states are symmetric, one can immediately show from symmetry considerations that the system cannot emit dipole radiation.

Due to this limitation, the dominant formation process is instead formation on the surfaces of dust grains. In this case the excess energy released by forming the molecule is transferred into vibrations in the dust grain lattice, and there is no need for forbidden photon emission.

Composition of interstellar dust: There is ample evidence for the presence of substantial amounts of submicron-sized dust particles in interstellar space. What is this dust made of? This question has been difficult to answer.

The preferred approach would be spectroscopy: ideally, we would observe spectroscopic features that would uniquely identify the materials, and, furthermore, allow us to measure the amounts of each material present. This is the approach that is followed for atoms, ions, and small molecules, but unfortunately it is difficult to apply to solid materials because: (1) the optical and UV absorption is largely a continuum; and (2) the spectral features that do exist are broad, making them difficult to identify conclusively.

An alternative approach is to ask: what materials could plausibly be present in the ISM in quantities sufficient to account for the observed extinction? A Kramers-Kronig integral over the observed extinction indicates that the total grain mass relative to total hydrogen mass $M_{\rm dust}/M_{\rm H} \gtrsim 0.0083$.

Additionally, certain elements appear to be underabundant, or "depleted," in the gas phase; observed depletions can tell us about the major elemental composition of interstellar dust. The available evidence indicates that the overall abundances in the ISM are close to the values in the solar photosphere. Because there is no way to have hydrogen contribute appreciably to the grain mass (even polyethylene $(CH_2)_n$ is 86% carbon by mass), and He and Ne are chemically inert, the only way to have a dust/H mass ratio of 0.0056 or higher is to build the grains out of the most abundant condensible elements: C, O, Mg, Si, S, and Fe.

With the elements providing the bulk of the grain volume identified, we can limit consideration to the following possible materials:

- Silicates (e.g., pyroxene composition or olivine composition)
- Oxides of silicon, magnesium, and iron (e.g., SiO_2, MgO, Fe_3O_4)
- Carbon solids (graphite, amorphous carbon, and diamond)

- Hydrocarbons (e.g., PAHs)
- Carbides, particularly silicon carbide (SiC)
- Metallic Fe

Other elements (e.g., Ti, Cr) are also present in interstellar grains, but, because of their low abundances, they contribute only a minor fraction of the grain mass.

1.5.3 Follow-up Questions

- Why is molecular hydrogen (H_2) so difficult to detect?
- What are other ways in which molecular cloud cores cool?

1.6 Question 6

The ISM mainly consists of hydrogen and helium, which are very poor coolants. How, then, do molecular cloud cores ever manage to lose enough heat to collapse and form stars? Why are H and He such poor coolants?

1.6.1 Short answer

Answer.

1.6.2 Additional context

Cooling of atomic gas:
Most of the interstellar gas in the Milky Way is neutral, and $\sim 78\%$ of the neutral hydrogen is atomic, or HI. The most abundant elements in the Universe after H and He are O, C, and N. Just as in the bulk of the ISM hydrogen is mostly H, in the bulk of the ISM the oxygen is mostly O and the carbon is mostly C^+. It's C^+ rather than C because the ionization potential of carbon is less than that of hydrogen, and as a result it tends to be ionized by starlight.
An example of the "**cooling function**" Λ for predominantly neutral gas, as a function of temperature, is shown in Figure 18 for abundances appropriate to diffuse HI in the Milky Way, and for two different fractional ionizations: $x_e = 0.017$ (WNM conditions) and $x_e = 4 \times 10^{-4}$ (CNM conditions). For $10 \lesssim T \lesssim 10^4\,\mathrm{K}$, the [CII]158 μm fine structure line is a major coolant. The [OI]63 μm fine structure line is important for $T \gtrsim 100\,\mathrm{K}$. Lyman$\alpha$ cooling dominates only at $T \gtrsim 1 \times 10^4\mathrm{K}$.
The critical densities for [CII]158 μm and [OI]63 μm are $\sim 4 \times 10^3\,\mathrm{cm}^{-3}$ and $\sim 10^5\,\mathrm{cm}^{-3}$, respectively, implying that collisional de-excitation of these levels is unimportant in the diffuse ISM of the MW. Thus, for fixed composition (and ionization fraction x_e), the cooling power per volume $\Lambda \propto n_\mathrm{H}^2 \times \lambda(T)$, where the cooling rate coefficient λ depends only on T.

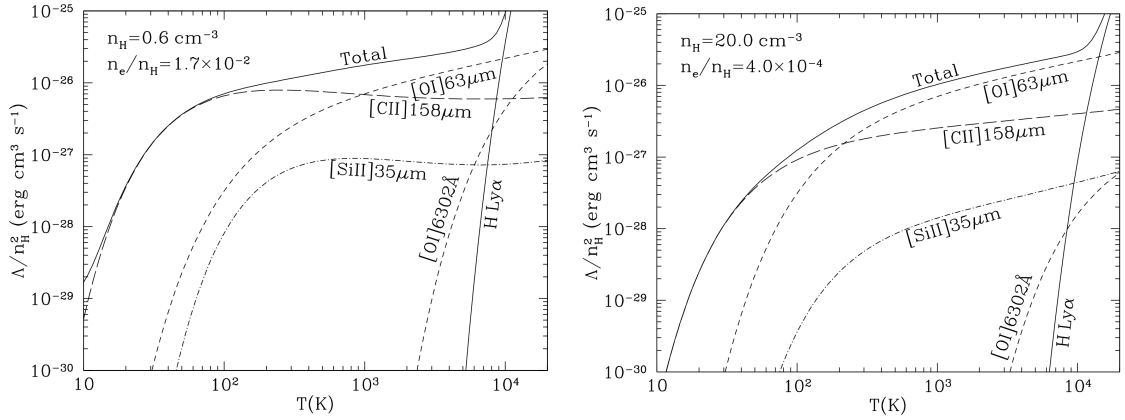

Figure 18: Cooling rate for neutral HI gas at temperatures $10 \lesssim T \lesssim 2 \times 10^4\,\mathrm{K}$ for two fractional ionizations. For $T < 10^4\,\mathrm{K}$, the cooling is dominated by two fine structure lines: [CII]158 μm and [OI]63 μm. Figure taken from Draine (2011).

Cooling of molecular gas:
At the high densities where stars form, hydrogen tends to be molecular rather than atomic, and H_2 very rarely emits radiation via de-excitation. To understand why, we can look at an energy level diagram for rotational levels of H_2 (Figure 19). A diatomic molecule like H_2 has three types of excitation: **electronic** (corresponding to excitations of one or more of the electrons), **vibrational** (corresponding to vibrational motion of the two nuclei), and **rotational** (corresponding to rotation of the two nuclei about the center of mass). Generally electronic excitations are highest in energy scale, vibrational are next, and rotational are the lowest in energy.
For H_2, the first thing to notice is that the first excited state, the $J = 1$ rotational state, is 175 K above the ground state. Since the dense ISM where molecules form is often also cold, $T \sim 10\,\mathrm{K}$, almost no molecules will be in this excited state. However, it gets even worse: H_2 is a **homonuclear molecule**, and for reasons of symmetry $\Delta J = 1$ radiative transitions are forbidden in homonuclear molecules. Indeed, there is no electronic process by which a hydrogen molecule with odd J to turn into one with even J, and vice versa, because the allowed parity of J is determined by the spins of the hydrogen nuclei. We refer to the even J state as **para-H_2**, and the odd J state as **ortho-H_2**.

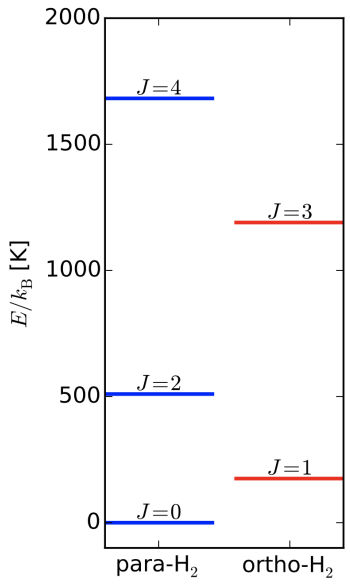

Figure 19: Level diagram for the rotational levels of para- and ortho-H_2, showing the energy of each level. Level data are taken from http://www.gemini.edu/sciops/instruments/nir/wavecal/h2lines.dat. Figure taken from Draine (2011).

The significance of this is that *there is no* $J = 1 \rightarrow 0$ *emission*. Instead, the lowest-lying transition is the $J = 2 \rightarrow 0$ quadrupole. This is very weak, because it's a quadrupole. More importantly, however, the $J = 2$ state is 510 K above the ground state. This means that, for a population in equilibrium at a temperature of 10 K, the fraction of molecules in the $J = 2$ state is $\sim e^{-510/10} \approx 10^{-22}$! In effect, in a molecular cloud there are simply no H_2 molecules in states capable of emitting and therefore cannot cool the gas. The very high temperature required to excite the H_2 molecular is its low mass: for a quantum oscillator or rotor the level spacing varies with reduced mass as $m^{-1/2}$. It is the low mass of the hydrogen atom that creates our problems.

In molecular clouds there are two main cooling processes: molecular lines and dust radiation. Dust can cool the gas efficiently because dust grains are solids, so they are thermal emitters. However, dust is only able to cool the gas if collisions between dust grains and hydrogen molecules occur often enough to keep them thermally well-coupled. Otherwise the grains cool off, but the gas stays hot. The density at which grains and gas become well-coupled is around $10^4 - 10^5\,\mathrm{cm}^{-3}$ which is higher than the typical density in a GMC.

The remaining cooling process is line emission, and by far the most important molecule for this purpose is CO. H_2 is the dominant species in molecular regions, but it is very hard to observe directly – the temperatures are too low for it to be excited. Moreover, H_2 is also not the dominant coolant for the same reason. Instead, that role falls to the CO molecule. The physics is fairly simple. CO molecules are excited by inelastic collisions with hydrogen molecules, and such collisions convert kinetic energy to potential energy within the molecule. If the molecule de-excites radiatively, and the resulting photon escapes the cloud, the cloud loses energy and cools.

1.7 Question 7

The stars in the solar neighbourhood, roughly the 300 pc around us, have a range of ages, metallicities and orbital properties. How are those properties related?

1.7.1 Short answer

Answer.

1.7.2 Additional context

Age-metallicity relation: Assuming that at the beginning of its evolution the MW had a chemical composition with only low metal content, the metallicity should be strongly related to the age of a stellar population. With each new generation of stars, more metals are produced and ejected into the ISM, partially by stellar winds, but mainly by SN explosions. Stars that are formed later should therefore have a higher metal content than those that were formed in the early phase of the Galaxy. One would thus expect that a relation exists between the age of a star and its metallicity.
For instance, under this assumption the iron abundance [Fe/H] can be used as an age indicator for a stellar population, with the iron predominantly being produced and ejected in SNe of Type Ia. Therefore, a newly formed generation of stars has a higher fraction of iron than their predecessors, and the youngest stars should have the highest iron abundance. Indeed one finds [FeH]$= 4.5$ (i.e., 3×10^{-5} of the Solar iron abundance) for extremely old stars, whereas very young stars have [FeH]$= 1$, so their metallicity can significantly exceed that of the Sun.
However, this age-metallicity relation is not very tight. On the one hand, SNe Ia occur only $\gtrsim 10^9$ yr after the formation of a stellar population. The exact time-span is not known because even if one accepts the accretion scenario for SN Ia described above, it is unclear in what form and in what systems the accretion of material onto the white dwarf takes place and how long it typically takes until the limiting mass is reached. On the other hand, the mixing of the SN ejecta in the ISM occurs only locally, so that large inhomogeneities of the [Fe/H] ratio may be present in the ISM, and thus even for stars of the same age. An alternative measure for metallicity is [O/H], because oxygen is produced and ejected mainly in supernova explosions of massive stars. These happen just $\sim 10^7$ yr after the formation of a stellar population, which is virtually instantaneous.
Velocity dispersion of stars: The dispersion of stellar velocities relative to the LSR can be determined, i.e., the mean square deviation of their velocities from the velocity of the LSR. For young stars (A stars, for example), this dispersion happens to be small. For older K giants it is larger, and is larger still for old, metal-poor red dwarf stars. We observe a very well-defined velocity-metallicity relation which, when combined with the age-metallicity relation, suggests that the oldest stars have the highest peculiar velocities. This effect is observed in all three coordinates and is in agreement with the relation between the age of a stellar population and its scale-height, the latter being linked to the velocity dispersion via σ_z.
Stellar populations: The chemical composition of stars in the thin and the thick disks differs: we observe the clear tendency that stars in the thin disk have a higher metallicity than those in the thick disk. In contrast, the metallicity of stars in the Galactic halo and in the bulge is smaller. To paraphrase these trends, one distinguishes between stars of population I (pop I) which have a Solar-like metallicity ($Z \sim 0.02$) and are mainly located in the thin disk, and stars of population II (pop II) that are metal-poor ($Z \sim 0.001$) and predominantly found in the thick disk, in the halo, and in the bulge. In reality, stars cover a wide range in Z, and the figures above are only characteristic values. For stellar populations a somewhat finer separation was also introduced, such as "extreme population I", "intermediate population II", and so on. The populations also differ in age (stars of pop I are younger than those of pop II), in scale height (as mentioned above), and in the velocity dispersion perpendicular to the disk (σ_z is larger for pop II stars than for pop I stars).
Stellar age distribution in the bulge: The stars in the bulge cover a large range in metallicity, $-1 \lesssim$ [Fe = H] $\lesssim 0.6$, with a mean of about 0.3, i.e., the mean metallicity is about twice that of the Sun. The metallicity also changes as a function of distance from the center, with more distant stars having a smaller value of [Fe/H].
The high metallicity means that either the stars of the bulge formed rather late, according to the age-metallicity relation, or that it is an old population with very intense star formation activities at an early cosmic epoch. We can distinguish between these two possibilities from the chemical composition of stars in the bulge, obtained from spectroscopy. Bulge stars have a significantly higher abundance of Mg, relative to iron, than the stars from the thin disk, but much more similar to thick disk stars. This implies that the enrichment must have occurred predominantly by core-collapse supernovae, since they produce a high ratio of elements like magnesium compared to iron, whereas Type Ia SNe produce

mainly iron-group elements. Therefore, most of the bulge stars must have formed before the Type Ia SNe exploded. Whereas the time lag between the birth of a stellar population and the explosion of the bulk of Type Ia SN is not well known (it depends on the evolution of binary systems), it is estimated to be between 1 and 3 Gyr. Hence, most of the bulge stars must have formed on a rather short time-scale: the bulge consists mainly of an old stellar population, formed within ~ 1 Gyr.

1.8 Question 8

What are the main sources of heat in the interstellar medium?

1.8.1 Short answer

Possible mechanisms for heating HI regions include:

- Ionization by cosmic rays
- Photoionization of H and He by x rays
- Photoionization of dust grains by starlight UV
- Photoionization of C, Mg, Si, Fe, etc., by starlight UV
- Heating by shock waves and other MHD phenomena

1.8.2 Additional context

Ionization by cosmic rays: The great advantage of cosmic rays over FUV photons is that, because they are relativistic particles, they have much lower interaction cross sections, and thus are able to penetrate into regions where light cannot. The process of cosmic ray heating works as follows. The first step is the interaction of a cosmic ray with an electron, which knocks the electron off a molecule:

$$\mathrm{CR} + \mathrm{H_2} \rightarrow \mathrm{H_2^+} + \mathrm{e^-} + \mathrm{CR}.$$

The free electron's energy depends only weakly on the CR's energy, and is typically $\sim 30\,\mathrm{eV}$.
The electron cannot easily transfer its energy to other particles in the gas directly, because its tiny mass guarantees that most collisions are elastic and transfer no energy to the impacted particle. However, the electron also has enough energy to ionize or dissociate other hydrogen molecules, which provides an inelastic reaction that can convert some of its $\sim 30\,\mathrm{eV}$ to heat. Secondary ionizations do indeed occur, but in this case almost all the energy goes into ionizing the molecule ($15.4\,\mathrm{eV}$), and the resulting electron has the same problem as the first one: it cannot effectively transfer energy to the much more massive protons. Instead, there are a number of other channels that allow electrons to dump their energy into motion of protons.
Photoionization of dust grains by starlight UV: The dominant heating process in the atomic ISM is the **grain photoelectric effect**: photons from stars with energies of $\sim 8 - 13.6\,\mathrm{eV}$ hit dust grains and eject fast electrons via the photoelectric effect. The fast electrons then thermalize and deposit their energy as heat in the gas. The rate per H nucleus at which this process deposits energy can be written approximately as

$$\Gamma_{\mathrm{PE}} \approx 4.0 \times 10^{-26} \chi_{\mathrm{FUV}} Z'_d e^{-\tau_d}\ [\mathrm{erg\,s^{-1}}],$$

where χ_{FUV} is the intensity of the FUV radiation field scaled to its value in the Solar neighborhood, Z'_d is the dust abundance scaled to the Solar neighborhood value, and τ_d is the dust optical depth to FUV photons. The result is, not surprisingly, proportional to the radiation field strength (and thus the number of photons available for heating), the dust abundance (and thus the number of targets for those photons),and the $e^{-\tau_d}$ factor by which the radiation field is attenuated.
At FUV wavelengths, typical dust opacities are $\kappa_d \approx 500\,\mathrm{cm^2\,g^{-1}}$, so at a typical molecular cloud surface density $\Sigma \approx 50 - 100\,\mathrm{M_\odot\,pc^2}$, $\tau_d \sim 5 - 10$, and thus $e^{-\tau_d} \approx 10^{-3}$. Thus in the interiors of molecular clouds, photoelectric heating is strongly suppressed simply because the FUV photons cannot get in. Typical photoelectric heating rates are therefore of order a few $10^{29}\,\mathrm{erg\,s^{-1}}$ per H atom deep in cloud interiors, though they can obviously be much larger at cloud surfaces or in regions with stronger radiation fields.
Shocks: Before discussing individual feedback mechanisms in detail, it is also helpful to lay out two general categories that can be used to understand them. Let us consider a population of stars surrounded by initially-uniform interstellar gas. Those stars eject both photons and baryons (in the form of stellar winds) into the surrounding gas, and these photons and baryons carry both momentum and energy. We want to characterize how the ISM will respond. One important consideration is that it is very hard to raise the temperature of molecular gas (or even dense atomic gas) because it is able to radiate so efficiently. A factor of 10 increase in the radiative heating rate might yield only a tens of percent increase in temperature. This is true as long as the gas is cold and dense, but at sufficiently high temperatures or if the gas is continuously illuminated then the cooling rate begins to drop off, and it is possible for gas to remain hot.

A critical distinction is therefore between mechanisms that are able to keep the gas hot for a time that is long enough to be significant (generally of order the crossing time of the cloud or longer), and those where the cooling time is much shorter. For the latter case, the energy delivered by the photons and baryons will not matter, only the momentum delivered will. The momentum cannot be radiated away. We refer to feedback mechanism where the energy is lost rapidly as momentum-driven feedback, and to the opposite case where the energy is retained for at least some time as energy-driven, or explosive, feedback.
To understand why the distinction between the two is important, let us consider two extreme limiting cases. We place a cluster of stars at the origin and surround it by a uniform region of gas with density ρ. At time $t = 0$, the stars "turn on" and begin emitting energy and momentum, which is then absorbed by the surrounding gas. Let the momentum and energy injection rates be $\dot{p}_w$ and $\dot{E}_w$; it does not matter if the energy and momentum are carried by photons or baryons, so long as the mass swept up is significantly greater than the mass carried by the wind.
The wind runs into the surrounding gas and causes it to begin moving radially outward, which in turn piles up material that is further away, leading to an expanding shell of gas. Now let us compute the properties of that shell in the two extreme limits of all the energy being radiated away, and all the energy being kept. If all the energy is radiated away, then at any time the radial momentum of the shell must match the radial momentum injected up to that time, i.e.,

$$p_{\rm sh} = M_{\rm sh} v_{\rm sh} = \dot{p}_{\rm sh} t \ [{\rm kg\,m\,s^{-1}}].$$

The kinetic energy of the shell is

$$E = \dot{p}_{\rm sh}^2 2 M_{\rm sh} = \frac{1}{2} v_{\rm sh} \dot{p}_w t \ [{\rm J}].$$

For comparison, if none of the energy is radiated away, the energy is simply

$$E = \dot{E}_w t \ [{\rm J}].$$

Thus the energy in the energy-conserving case is larger by a factor of

$$\frac{1}{v_{\rm sh}} \cdot \frac{2\dot{E}_w}{\dot{p}_w}.$$

If the energy injected by the stars is carried by a wind of baryons, then $2\dot{E}_w/\dot{p}_w$ is simply the speed of that wind, while if it is carried by photons, then $2\dot{E}_w/\dot{p}_w = 2c$. Thus the energy in the energy-conserving case is larger by a factor of $2c/v_{sh}$ for a photon wind, and $v_w/v_{\rm sh}$ for a baryon wind. These are not small factors: observed expanding shells typically have velocities of at most a few tens of $km\,s^{-1}$, while wind speeds from massive stars, for example, can be thousands of $km\,s^{-1}$. Thus it matters a great deal where a particular feedback mechanism lies between the energy- and momentum-conserving limits.
Momentum-driven feedback mechanisms include radiation pressure (probably, since the majority of the radiant energy deposited in the ISM will be re-radiated immediately), protostellar jets (due to their characteristic speeds) while energy-driven feedback mechanisms include ionizing radiation, stellar winds, and supernovae.

1.8.3 Follow-up Questions

- Are there any non-ionization sources of heat in the ISM? (shocks)
- How do shock waves heat the gas?
- Are shock waves adiabatic?
- Where do the x-rays for x-ray photoionization come from?
- What phases and temperatures of the ISM apply to each example?

1.9 Question 9

Draw an interstellar extinction curve (i.e., opacity), from the X-ray to the infrared. What are the physical processes responsible?

1.9.1 Short answer

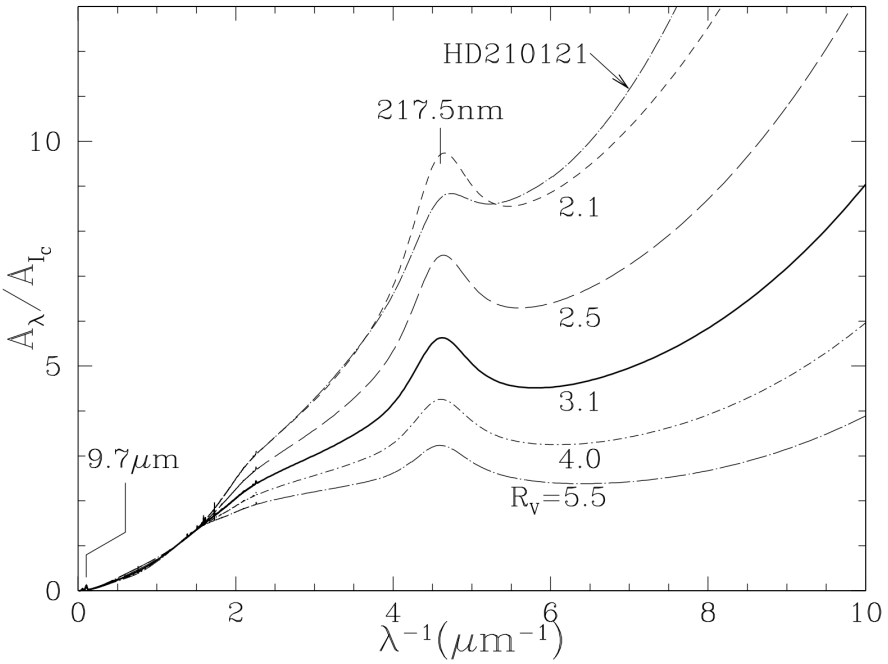

Figure 20: Extinction at wavelength λ, relative to the extinction in the Cousins I band ($I_C = 8020$Å), as a function of inverse wavelength λ^{-1}, for Milky Way regions characterized by different values of $R_V \equiv AV/(AB - AV) \equiv AV/E(BV)$, where A_B is the extinction at $B = 0.44\,\mu$m, A_V is the extinction at $V = 0.55\,\mu$m, and the "reddening" $E(BV) \equiv A_B A_V$. The curves shown are from the one-parameter family of curves $f_1^{\rm CCM}(\lambda)$ parameterized by R_V. Also shown is the extinction curve toward the star HD210121 (with $R_V = 2.1$), showing that it differs from the CCM extinction curve $f_1^{\rm CCM}$ for $R_V = 2.1$. Note the rapid rise in extinction in the vacuum ultraviolet ($\lambda \lesssim 0.15\,\mu$m) for regions with $R_V \lesssim 4$. The normalization per H nucleon is approximately $A_{I_C}/N_H \approx 2.9 \times 10^{-22}\,{\rm mag\,cm^2\,H^{-1}}$. The silicate absorption feature at $9.7\,\mu$m and the diffuse interstellar bands are barely visible. Figure taken from Draine (2011).

1.9.2 Additional context

Dust plays an important role in astrophysics, and the need to characterize and understand dust is increasingly appreciated. Historically, interstellar dust was first recognized for its obscuring effects, and the need to correct observed intensities for attenuation by dust continues today. But with the increasing sensitivity of IR, FIR, and sub-mm telescopes, dust is increasingly important as a diagnostic, with its emission spectrum providing an indicator of physical conditions, and its radiated power bearing witness to populations of obscured stars of which we might otherwise be unaware.

More fundamentally, dust is now understood to play many critical roles in galactic evolution. By sequestering selected elements in the solid grains, and by catalyzing formation of the H2 molecule, dust grains are central to the chemistry of interstellar gas. Photoelectrons from dust grains can dominate the heating of gas in regions where ultraviolet starlight is present, and in dense regions the infrared emission from dust can be an important cooling mechanism. Last, dust grains can be important in interstellar gas dynamics, communicating radiation pressure from starlight to the gas, and coupling the magnetic field to the gas in regions of low fractional ionization.

Barnard was apparently the first to realize that some stars were dimmed by an absorbing medium. This was confirmed by Trumpler who showed that the stars in distant open clusters were dimmed by something in addition to the inverse square law, and concluded that interstellar space in the galactic plane contained "fine cosmic dust particles of various sizes... producing the observed selective absorption." Over the succeeding eight decades, we have built on these pioneering studies, but many aspects of interstellar dust (including its chemical composition!) remain uncertain. Let us, therefore, begin by reviewing the different ways in which nature permits us to study interstellar dust.

Trumpler analyzed the interaction of light with interstellar dust, and this remains our most direct way to study interstellar dust. Using stars as "standard candles," we study the "selective extinctio" (or

Band	$\lambda(\mu\text{m})$	A_λ/A_{I_C}	Band	$\lambda(\mu\text{m})$	A_λ/A_{I_C}
M	4.75	0.0573	i	0.7480	1.125
L'	3.80	0.0842	R_C	0.6492	1.419
L	3.45	0.101	R_J	0.6415	1.442
K	2.19	0.212	r	0.6165	1.531
H	1.65	0.315	V	0.5470	1.805
J	1.22	0.489	g	0.4685	2.238
z	0.893	0.830	B	0.4405	2.396
I_J	0.8655	0.879	U	0.3635	2.813
I_C	0.8020	1.000	u	0.3550	2.867

Table 2: Extinction for Standard Photometric Bands for $R_V = 3.1$. Figure taken from Draine (2011).

reddening") of starlight by the dust. It is assumed that we know what the spectrum of the star is before reddening by dust takes place; this is usually accomplished by observation of another star with similar spectral features in its atmosphere but with negligible obscuration between us and the star. (This is known as the "**pair method**".)

With the assumption that the extinction ($\equiv$ absorption + scattering) goes to zero at wavelengths $\lambda \to \infty$, and including observations of the star at sufficiently long wavelength where extinction is negligible, one can determine the attenuation of the starlight by dust as a function of wavelength. Because atomic hydrogen absorbs strongly for $h\nu > 13.6\,\text{eV}$, it is possible to measure the contribution of dust to the attenuation of light only at $h\nu < 13.6\,\text{eV}$, or $\lambda > 912\,\text{Å}$. Astronomers customarily characterize the attenuating effects of dust by the "extinction" A_λ at wavelength λ. The extinction A_λ (measured in "magnitudes") is defined by

$$A_\lambda = 2.5 \log_{10}\left(\frac{F_{\lambda,0}}{F_\lambda}\right) \text{ [mag]},$$

where F_λ is the observed flux from the star, and $F_{\lambda,0}$ is the flux that would have been observed had the only attenuation been from the inverse square law. The extinction measured in magnitudes is proportional to the optical depth:

$$A_\lambda = 2.5 \log(e)\tau_\nu = 1.086\tau_\nu \text{ [mag]}.$$

A typical "extinction curve" (the extinction A_λ as a function of wavelength or frequency) is shown in Figure 20, showing the rapid rise in extinction in the vacuum ultraviolet. Because the extinction increases from red to blue, the light reaching us from stars will be "reddened" owing to greater attenuation of the blue light. The detailed wavelength dependence of the extinction (the "reddening law") is sensitive to the composition and size distribution of the dust particles.

Observed extinction curves vary in shape from one line of sight to another. The slope of the extinction at visible wavelengths is characterized by the dimensionless ratio

$$R_V \equiv \frac{A_V}{A_B - A_V} \equiv \frac{A_V}{E(B-V)} \text{ [dimensionless]},$$

where A_B and A_V are the extinctions measured in the B (4405 Å) and V (5470 Å) photometric bands, and $E(B-V) \equiv A_B - A_V$ is the "reddening."

Sightlines through diffuse gas in the MW have $R_V \approx 3.1$ as an average value. The extinction A_λ, relative to A_V, is given in Table 2 for a number of standard photometric bands for sightlines characterized by $R_V \approx 3.1$. The smallest well-determined value is $R_V = 2.1$ toward the star HD 210121; the extinction toward HD 201021 is shown in 20. Sightlines through dense regions tend to have larger values of R_V; the sightline toward HD 36982 has $R_V \approx 5.7$.

A very useful parametrization of the extinction curve within the MW was provided by Cardelli et al. (1989), who showed that the extinction relative to some reference wavelength λ_{ref} can be well-described as a function of λ by a fitting function

$$\frac{A_\lambda}{A_{\text{ref}}} \approx f_7^{\text{CCM}}(\lambda) \text{ [mag]},$$

where f_7^{CCM} has seven adjustable parameters. At wavelengths $3.5\,\mu\text{m} > \lambda > 3030\,\text{Å}$, the function $f_7^{\text{CCM}}(\lambda)$ depends only on λ and the single parameter R_V.

Six parameters are required to describe the UV extinction. Three parameters specify the strength, central wavelength, and width of the 2175 Å "bump" (relative to A_V), and three specify the slope and curvature of the continuous extinction underlying the bump and extending to shorter wavelengths. So-called **CCM extinction curves** are obtained using the function $f_7^{\rm CCM}$ with suitable choices for the 7 seven fit parameters.
Cardelli et al. (1989) showed that if the single quantity R_V is known, it is possible to estimate the values of the other six parameters so that the optical-UV extinction can be approximated by a one-parameter family of curves:

$$\frac{A_\lambda}{A_{\rm ref}} \approx f_1^{\rm CCM}(\lambda; R_V) \text{ [mag]},$$

It is clear that if the dust grains were large compared to the wavelength, we would be in the "geometric optics" limit, and the extinction cross section would be independent of wavelength, with $R_V = \infty$. The tendency for the extinction to rise with decreasing λ, even at the shortest UV wavelengths where we can measure it, tells us that grains smaller than the wavelength must be making an appreciable contribution to the extinction at all observed wavelengths, down to $\lambda = 0.1\,\mu$m. "Small" means (approximately) that $2\pi a/\lambda \lesssim 1$. Thus interstellar dust must include a large population of grains with $a \lesssim 0.015\,\mu$m.
A number of different quantities are used to characterize the absorption, scattering, and emission of electromagnetic radiation by a (non-rotating) dust grain:

- The absorption cross section at wavelength λ: $C_{\rm abs}(\lambda)$
- The scattering cross section: $C_{\rm sca}(\lambda)$
- The extinction cross section: $C_{\rm ext}(\lambda) \equiv C_{\rm abs}(\lambda) + C_{\rm sca}(\lambda)$
- The albedo:

$$\omega \equiv \frac{C_{\rm sca}(\lambda)}{C_{\rm abs}(\lambda) + C_{\rm sca}(\lambda)} = \frac{C_{\rm sca}(\lambda)}{C_{\rm ext}(\lambda)} \text{ [dimensionless]}.$$

- The differential scattering cross section:

$$\frac{{\rm d}C_{\rm sca}}{{\rm d}\Omega}$$

for incident unpolarized light to be scattered by an angle θ. This is related to the dimensionless Muller matrix element S_{11} by

$$\frac{{\rm d}C_{\rm sca}}{{\rm d}\Omega} \equiv \frac{S_{11}(\theta)}{k^2}$$

where $k \equiv 2\pi/\lambda$.

- The mean value of $\cos\theta$ for scattered light:

$$\langle \cos\theta \rangle = \frac{1}{C_{\rm sca}} \int_0^\pi \cos\theta \frac{{\rm d}C_{\rm sca}}{{\rm d}\Omega} 2\pi \sin\theta {\rm d}\theta \text{ [dimensionless]}.$$

- The radiation pressure cross section:

$$C_{\rm pr}(\lambda) \equiv C_{\rm abs}(\lambda) + (1 - \langle \cos\theta \rangle) C_{\rm sca}(\lambda) \text{ [cm}^2\text{]}.$$

- The degree of polarization $P(\theta)$ for light scattered through an angle θ (for incident unpolarized light).

For a given direction of incidence relative to a fixed grain, we would obviously need two angles (θ, ϕ) to fully specify the scattering direction. However, for spherical grains, or for an ensemble of randomly oriented grains, the scattering properties can be described as a function of a single scattering angle θ.
In some cases, one wants to consider scattering of polarized light. For this case, it is usual to use the four-element **Stokes vector** to specify the intensity and state of polarization of radiation propagating in a particular direction. The ability of a grain to scatter radiation with incident Stokes vector $V_{\rm in}$ to outgoing Stokes vector $V_{\rm sca}$ is conveniently specified by a 4×4 dimensionless scattering matrix S_{ij}, known as the Muller matrix.

It is convenient to normalize the absorption and scattering cross sections $C_{\rm abs}$ and $C_{\rm sca}$ to some area characterizing the grain. In the case of a spherical grain, it is natural to use the grain geometric cross section πa^2.
For non-spherical grains, some authors choose to normalize using the geometric cross section as seen from the direction of the incident radiation; other authors choose to normalize using the average geometric cross section for random orientations.
Here, we will instead normalize to the geometric cross section of an equal-solid-volume sphere. For a target with solid volume V (V does not include the volume of any voids, if present), we define efficiency factors $Q_{\rm sca}$, $Q_{\rm abs}$ and $Q_{\rm ext} \equiv Q_{\rm abs} + Q_{\rm sca}$ by

$$Q_{\rm sca} \equiv \frac{C_{\rm sca}}{\pi a_{\rm eff}^2} \ [{\rm cm}^{-2}]$$
$$Q_{\rm abs} \equiv \frac{C_{\rm abs}}{\pi a_{\rm eff}^2} \ [{\rm cm}^{-2}]$$
$$a_{\rm eff} \equiv \left(\frac{3V}{4\pi}\right)^{1/3} \ [\mu{\rm m}].$$

Here, $a_{\rm eff}$ is the radius of an equal-volume sphere. This is a natural choice, because it relates the scattering and absorption cross sections directly to the actual volume of grain material.
In order to calculate scattering and absorption of electromagnetic waves by targets, we need to characterize the response of the target material to the local oscillating electric and magnetic fields. At submillimeter frequencies and above, real materials have only a negligible response to an applied magnetic field – this is because the magnetization of materials is the result of aligned electron spins and electron orbital currents, and an electron spin (or orbit) can change direction only on time scales longer than the period for the electron spin (or orbit) to precess in the local (microscopic) magnetic fields within atoms and solids. These fields are at most $B_i \lesssim 10\,{\rm kG}$, and the precession frequencies are $\omega_p \approx \mu_B B_i/\hbar \lesssim 10^{10}\,{\rm s}^{-1}$, where μ_B is the **Bohr magneton** given by the equation

$$\mu_B \equiv \frac{e\hbar}{2m_e c} \ [{\rm erg\,G}^{-1}].$$

When a weak applied field oscillates at frequencies$\omega \ll 10^{10}\,{\rm s}^{-1}$, the magnetization of the material cannot respond. As a result, for frequencies $\nu \geq 10\,{\rm GHz}$ we normally set the magnetic permeability $\mu = 1$, and consider only the material's response to the oscillating electric field.
The response of material to an applied oscillating electric field $E = E_0 e^{-i\omega t}$ is characterized by a complex **dielectric function** of the permittivity ϵ:

$$\epsilon(\omega) = \epsilon_1 + i\epsilon_2 \ [{\rm F\,m}^{-1}].$$

The electrical conductivity σ, if any, can be absorbed within the imaginary part of the dielectric function, with the replacement

$$\epsilon \rightarrow \frac{4\pi i\sigma}{\omega} \ [{\rm S\,m}^{-1}].$$

The complex **refractive index** $m(\omega)$ is related to the complex dielectric function by $m = \sqrt{\sigma}$.
There are two sign conventions for the imaginary part of the dielectric function or refractive index. If we choose to write oscillating quantities $\propto e^{i\vec{k}\cdot\vec{r}} - i\omega t$, then ${\rm Im}(\epsilon) > 0$ and ${\rm Im}(m) > 0$ for absorbing, dissipative materials, where a propagating wave is attenuated. This is the convention that we will use.
In terms of the refractive index, the wave vector

$$k = m(\omega)\frac{\omega}{c} \ [{\rm m}^{-1}]$$

and, therefore, the electric field

$$E \propto e^{i(kx-\omega t)} \propto e^{-{\rm Im}(m)\omega x/c} \ [{\rm N\,C}^{-1}]$$

and the power in the wave ($\propto |E|^2$) decays as $\exp[2{\rm Im}(m)\omega x/c]$. Therefore, the attenuation coefficient κ and attenuation length $L_{\rm abs} \equiv 1/\kappa$ for the wave are simply

$$\kappa(\omega) = 2{\rm Im}(\omega)\frac{\omega}{c}$$
$$L_{\rm abs}(\omega) = \frac{c}{2\omega{\rm Im}(m)} = \frac{\lambda}{4\pi{\rm Im}(m)}$$

where $\lambda = 2\pi c/\omega$ is the wavelength in vacuo.

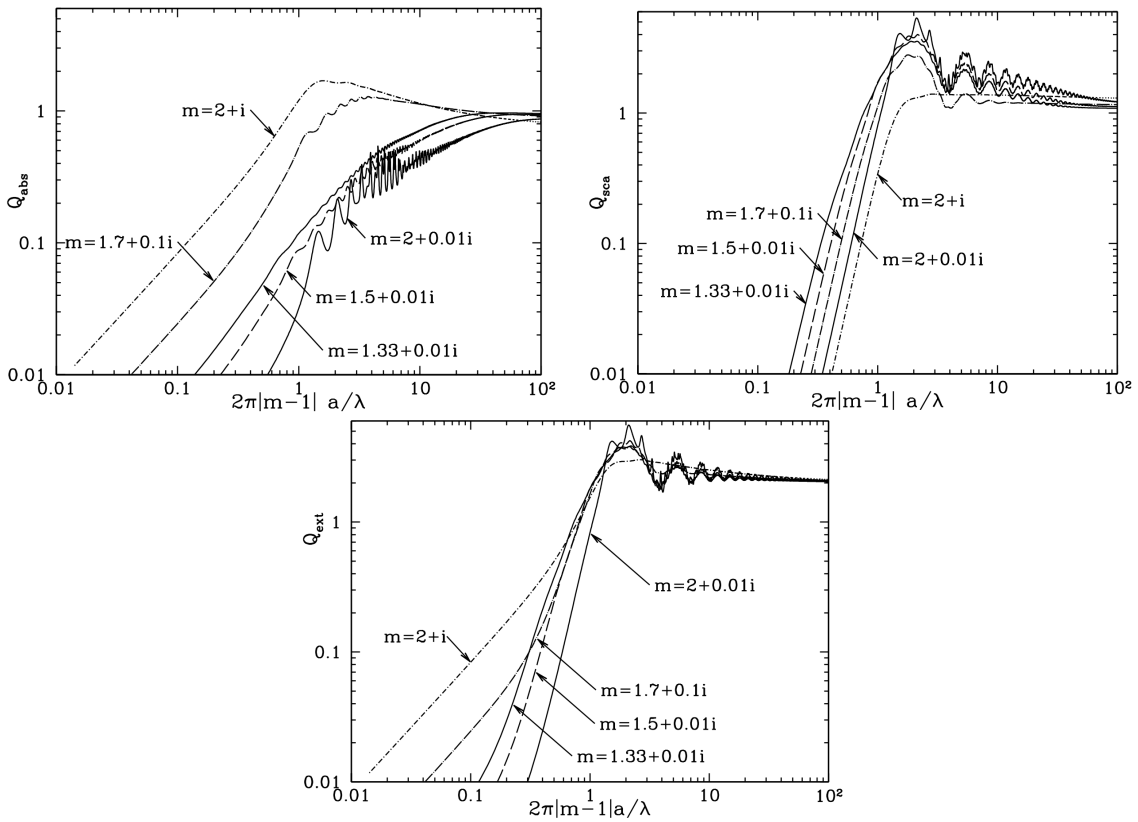

Figure 21: **Top left**: Absorption efficiency factors $Q_{\rm abs}$ for spheres with various refractive indices m. **Top right**: Scattering efficiency factors $Q_{\rm sca}$ for spheres with various refractive indices m. **Bottom**: Extinction efficiency factors $Q_{\rm ext}$ for spheres with various refractive indices m. Note that $Q_{\rm ext} \to 2$ for $|m-1|a/\lambda \to \infty$. Figure taken from Draine (2011).

We are often interested in situations where the grain is much smaller than the wavelength of the incident EM wave. In this situation, the small grain is subject to an incident applied electric field that is nearly uniform in space. The electric field inside the grain will be proportional to the applied external electric field $\mathrm{Re}(E)0e^{-i\omega t})$. Averaged over one cycle, the rate per volume at which energy is absorbed within the grain is proportional to $\omega\epsilon 2E_0^2$.
The absorption and scattering cross sections can be written

$$C_{\rm abs} = \frac{4\pi\omega}{c}{\rm Im}(\alpha)\ [{\rm cm}^2]$$
$$C_{\rm sca} = \frac{8\pi}{3}\left(\frac{\omega}{c}\right)^4 |\alpha|^2\ [{\rm cm}^2],$$

where α is the electric polarizability of the grain: the electric dipole moment of the grain $\vec{\mu} = \alpha\vec{E}$, where $\vec{E}$ is the instantaneous applied electric field. Calculating the polarizability in the limit $\omega a/c \to 0$ becomes a problem in electrostatics.
At optical and UV wavelengths, the dust particles are not necessarily small compared to the wavelength, and the electric dipole approximation is no longer applicable. We must find the solution to Maxwell's equations with an incident plane wave, for an object of specified size and shape, composed of material with a specified dielectric function ϵ or refractive index m.
For the special case of a sphere, an elegant analytic solution was found by Mie (1908) and Debye (1909), and is known as **Mie theory**. In brief, the EM field inside and outside the sphere can be decomposed into spherical harmonics with appropriate radial functions, with coefficients determined by the need to give an incident plane wave at infinity and to satisfy the continuity conditions at the surface of the sphere. Computer programs to evaluate the Mie theory solution are widely available.
The character of the EM scattering will depend on the dimensionless ratio a/λ and on the dimensionless refractive index $m(\omega)$. One relevant parameter will be the phase shift of a wave traveling a distance equal to the grain radius within the grain, expressed in radians. For non-absorptive material, this would be just $2\pi a|m-1|/\lambda$. Figure 21 shows five examples, where we plot the absorption (top), scattering (middle), and extinction (bottom) efficiency factors against this phase shift.
The details depend on the refractive index m, but the general trend is for $Q_{\rm ext}$ to rise to a value $Q_{\rm ext} \approx 3{-}5$ near $|m-1|2\pi a/\lambda \approx 2$. For dielectric functions with small imaginary components (i.e., weakly absorbing

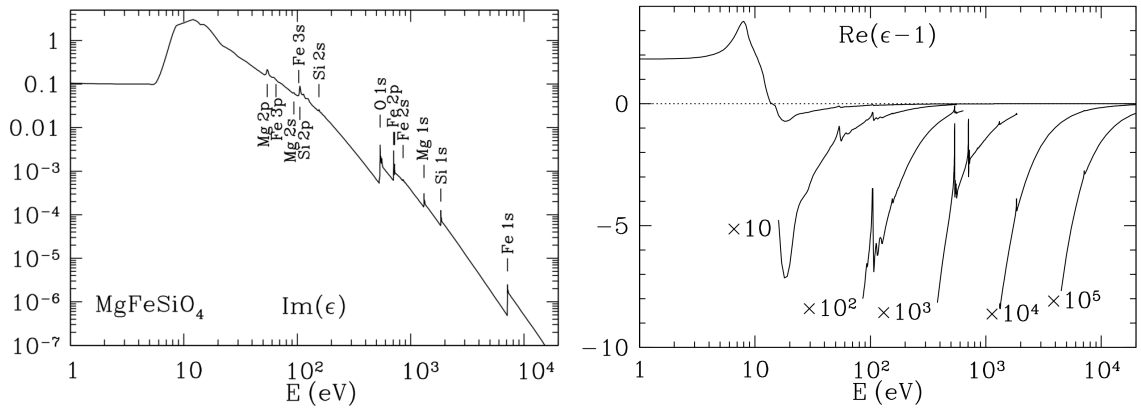

Figure 22: Dielectric function ϵ for Mg Fe Si O_4 material. Various absorption edges are labelled in the plot of Im(ϵ). This dielectric function, and its continuation at lower energies, will be referred to as "**astrosilicate**". From Draine (2003b), reproduced by permission of the AAS. Figure taken from Draine (2011).

material, Im(m) $\ll 1$) $Q_{\rm ext}$ as a function of a/λ shows oscillatory behavior due to interference effects, but the oscillations are minimal for strongly absorbing materials (Im(m) $\gtrsim 1$]).
For $(a/\lambda) \to \infty$, all of these examples have $Q_{\rm ext} \to 2$. This is a general result, sometimes referred to as "**the extinction paradox**": for $x \equiv 2\pi a/\lambda \to \infty$ and $|m-1|x \to \infty$, the extinction cross section is equal to exactly twice the geometric cross section.
Ray-tracing arguments would lead us to expect the extinction cross section to be equal to the geometric cross section, but diffraction around the target leads to additional small-angle scattering, with the total extinction cross section equal to twice the geometric cross section.
Mie theory is a powerful and robust computational tool with which one can efficiently calculate scattering and absorption by spheres with a wide range of dielectric constants, for $x \equiv 2\pi a/\lambda \lesssim 10^4$. For $x > 10^4$, cancellation in the alternating series leads to round-off errors on machines with 64-bit arithmetic, but for the size distributions that are present in the ISM, scattering by the dust mixture is usually dominated by particles with $x \approx 1$, and particles with $x \gg 1$ can generally be ignored except at x-ray energies.
However, one thing we know for certain about interstellar grains: the observed polarization of starlight implies that they are not spherical. If the grains are not spherical, how are we to calculate scattering and absorption cross sections? Elegant analytic treatments do exist for spheroids or infinite cylinders, but for more general shapes it is necessary to resort to brute force treatments. One approach that has proven useful is to approximate the actual target (with its particular geometry and dielectric function) by an array of "point dipoles." For a target illuminated by an incident monochromatic EM wave, each of these dipoles is assigned a complex polarizability $\alpha(\omega)$. Each dipole has an instantaneous dipole moment $\vec{\mu}_j = \alpha_j \vec{E}_j$, where α_j is the polarizability tensor for dipole j, and $\vec{E}_j$ is the electric field at location j due to the incident wave plus all of the other dipoles. This method is known as the **discrete dipole approximation** (DDA) or coupled dipole approximation.
DDA calculations are CPU-intensive, but many problems of practical interest can be handled by a desktop computer. For example, the DDA has been used to study absorption and scattering by graphite particles and by random agglomerates.
Figure 22.4 shows the real and imaginary components of the dielectric function for Mg Fe Si O_4. In the optical and UV, normal solids have refractive indices $|m-1| \gtrsim 0.3$. At x-ray energies, however, $|m-1| \ll 1$, and the character of the scattering changes considerably. The wavelength $\lambda = 0.00124\,({\rm keV}/h\nu)\,\mu$m is small compared to the sizes $a \approx 0.2\,\mu$m of the particles containing most of the grain mass. The result is that the x-ray scattering is very strongly peaked in the forward direction, with a characteristic scattering angle

$$\theta \approx \frac{\lambda}{\pi a} \approx 800'' \left(\frac{\rm keV}{h\nu}\right)\left(\frac{0.1\,\mu{\rm m}}{a}\right)\ [{\rm rad}].$$

Above we have discussed calculational methods for various regimes. We can now calculate scattering and absorption cross sections for micron- or submicron-sized grains from x-ray to sub-mm wavelengths. Figure 23 shows the extinction efficiency $Q_{\rm ext}$ calculated for grains of amorphous silicate ("astrosilicate") from the x-ray to the submm, for four different sizes. There are several noteworthy features:

- $Q_{\rm ext}$ shows sharp discontinuities at x-ray absorption edges. The amorphous silicate material is assumed to have composition Mg Fe Si O_4. Two conspicuous edges are the Fe K edge at 1.75Å (7.1 keV) and the O K edge at 23Å (528 eV). Note that the appearance of these edges depends on grain size.

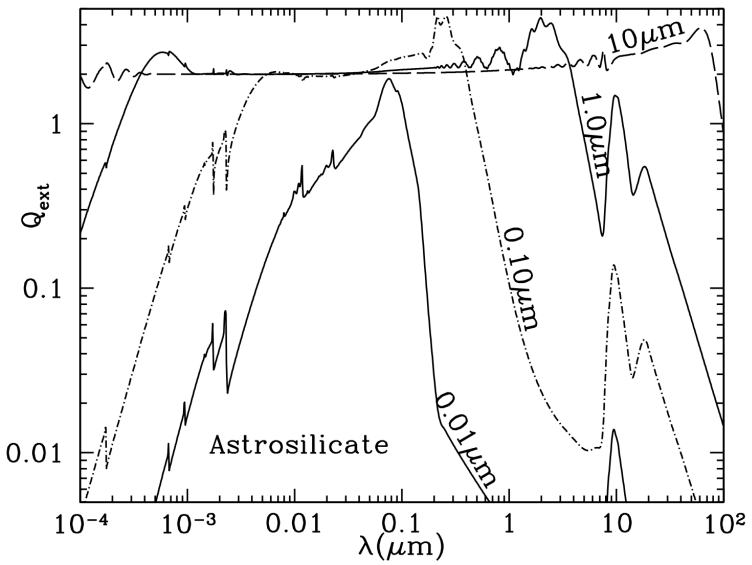

Figure 23: $Q_{\rm ext} \equiv C_{\rm ext}/\pi a 2$ for $a = 0.01, 0.1, 1,$ and $10\,\mu$m amorphous silicate spheres, for wavelengths ranging from $\lambda = 10^{-4}\,\mu{\rm m} = 1$Å ($h\nu = 12.4\,{\rm keV}$) to $\lambda = 10^3\,\mu{\rm m} = 1\,{\rm mm}$. At short wavelengths, the $a = 0.01$ and $0.10\,\mu$m grains show discontinuities in $Q_{\rm ext}$ at x-ray absorption edges. In the IR, the $a = 0.01, 0.1, 1\,\mu$m grains show prominent silicate absorption features at 9.7 and $18\,\mu$m, but these features are suppressed when $a = 10\,\mu$m. Figure taken from Draine (2011).

As the grains become larger, scattering makes an appreciable contribution to $Q_{\rm ext}$, and the long-wavelength side of the O K edge is "filled in" by scattering.

- $a = 1\,\mu$m grains are, in effect, optically thick (with $Q_{\rm ext} \approx 2$) for $0.001 \lesssim \lambda \lesssim 2\,\mu$m; for $\lambda < 10^{-3}\,\mu$m ($h\nu > 1.24\,{\rm keV}$), the absorption length exceeds the grain diameter, and for $\lambda > 2\,\mu$m, the grain is smaller than the wavelength. Similarly, the $a = 10\,\mu$m grain is optically thick for $10^{-4}\,\mu{\rm m} \lesssim \lambda \lesssim 10\,\mu$m.
- The silicate absorption features at 9.7 and $18\,\mu$m are prominent absorption features for the $a = 0.01$, 0.1, $1\,\mu$m cases shown, but are suppressed in the $a = 10\,\mu$m example, because the grain is, in effect, optically thick at wavelengths on either side of the silicate features.

Figure 24 shows the behavior of $Q_{\rm ext}$ for wavelengths running from the vacuum UV into the infrared. The upper panel shows the extinction efficiency factors $Q_{\rm ext}$ for spheres with the "astrosilicate" dielectric function. For the wavelength range shown here, there are no spectral features, although small particles do show a rise in extinction for $\lambda \lesssim 0.2, \mu$m due to the onset of ultraviolet absorption in silicates. Scattering becomes important for $\lambda \lesssim 2\pi a$ (i.e., $x = 2\pi a/\lambda \gtrsim 1$), and $Q_{\rm ext} \gtrsim 2$ for $\lambda \lesssim 4a$.
For the adopted optical constants, the small ($a \lesssim 0.02\,\mu$m) carbonaceous particles show a strong absorption feature near 2175Å, closely matching the observed interstellar feature near this wavelength. However,

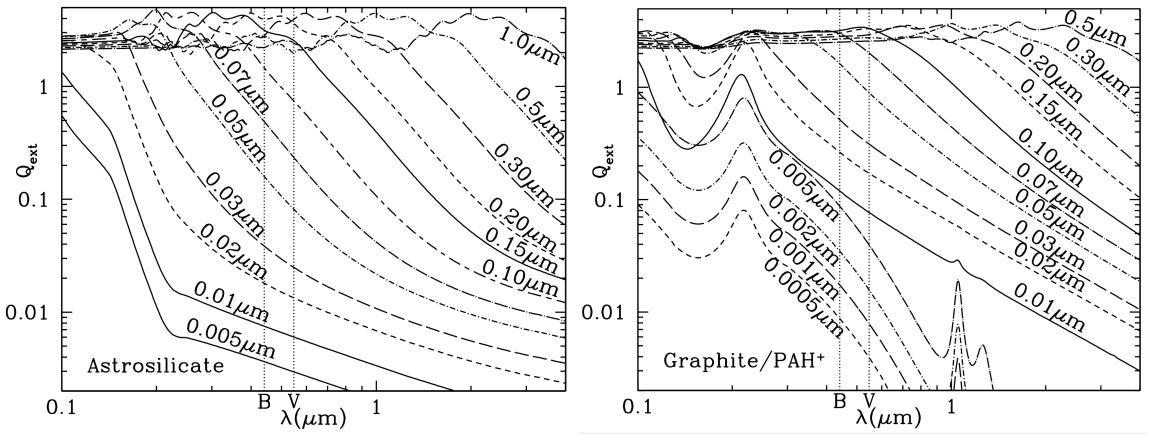

Figure 24: $Q_{\rm ext}$ for astrosilicate spheres (left) and carbonaceous spheres (right) for wavelengths ranging from $\lambda = 0.1, \mu$m to $\lambda = 4, \mu$m. The locations of the B (4405Å) and V (5470Å) bands are shown. Curves are labeled by radius a. Figure taken from Draine (2011).

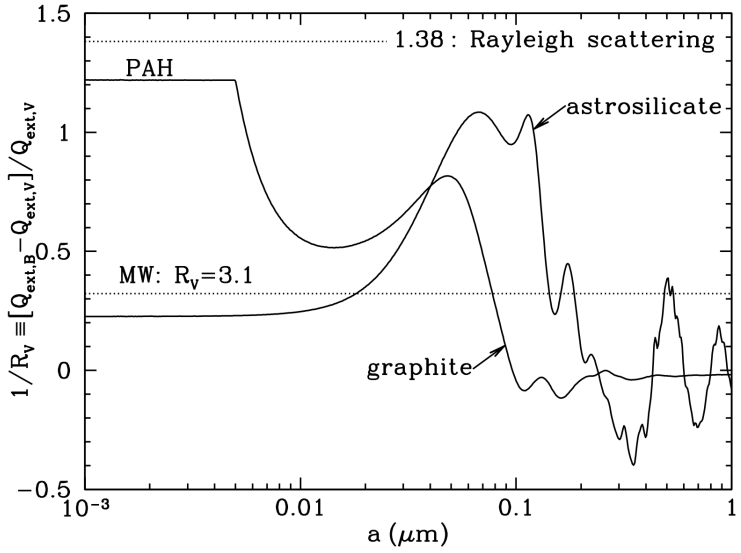

Figure 25: $1/R_V \equiv (C_{\rm ext}(B) - C_{\rm ext}(V))/C_{\rm ext}(V)$ as a function of radius a for astrosilicate and carbonaceous spheres ($B = 0.44\,\mu$m, $V = 0.55\,\mu$m). The carbonaceous spheres are assumed to be graphitic for $a > 0.01\,\mu$m, and PAHs for $a < 0.005\,\mu$m, with a continuous transition between 0.005 and $0.01\,\mu$m. For $a \lesssim 0.02\,\mu$m, scattering is unimportant, and R_V is determined by the absorptive properties of the grain material. For $a \gtrsim 0.12\,\mu$m, scattering resonances move through the wavelength range between B and V, and $1/R_V$ has oscillatory behavior. The dust in the diffuse ISM is observed to have $R_V \approx 3.1$, shown by the dotted line. $R_V \approx 3.1$ for $a \approx 0.08\,\mu$m or $0.14\,\mu$m for graphitic and astrosilicate grains, respectively. Figure taken from Draine (2011).

the theoretically-calculated feature broadens as the grain size increases to $0.03\,\mu$m, and disappears for larger grains because the grain becomes optically thick not only at the wavelength of the resonance, but also at wavelengths above and below the resonance.

As discussed earlier, interstellar extinction curves are often characterized by $R_V \equiv A_V/(A_B - A_V)$, and it is of interest to see what value of RV would apply to the extinction produced by grains of a single size and composition. Because R_V is singular when $A_B = A_V$, it is preferable to instead consider $1/R_V \equiv (A_B - A_V)/A_V$, which is proportional to the slope of the extinction curve between V and B. Figure 25 shows $1/R_V$ versus grain radius for carbonaceous grains and astrosilicate grains. For very small grains, scattering is negligible compared to absorption, and the value of R_V in the limit $a \to 0$ depends on the wavelength dependence of the optical constants– hence the very different limiting values for PAHs and astrosilicates. As the grain radius is increased, scattering begins to contribute significantly to the extinction, but we see that neither the silicate nor carbonaceous particles ever reach the value of $1/R_V = 1/0.726$ appropriate to Rayleigh scattering by particles with a polarizability that is wavelength independent. This is because, for our assumed dielectric functions, when the particles are small enough to be in the Rayleigh limit, absorption makes an important contribution to the extinction.

$R_V \approx 3.1$ is attained by graphitic grains for $a \approx 0.08\,\mu$m, and by astrosilicate grains for $\approx 0.15\,\mu$m. Although a broad size distribution is required to match the full extinction curve, grain models that reproduce the observed extinction should have the extinction in the visible dominated by grains with $a \approx 0.1\,\mu$m.

1.9.3 Follow-up Questions

- What happens at shorter wavelengths, like gamma rays?

1.10 Question 10

What is dynamical friction? Explain how this operates in the merger of a small galaxy into a large one.

1.10.1 Short answer

Answer.

1.10.2 Additional context

Additional context.

1.11 Question 11

Sketch the SED, from the radio to Gamma, of a spiral galaxy like the Milky Way. Describe the source and radiative mechanism of each feature.

1.11.1 Short answer

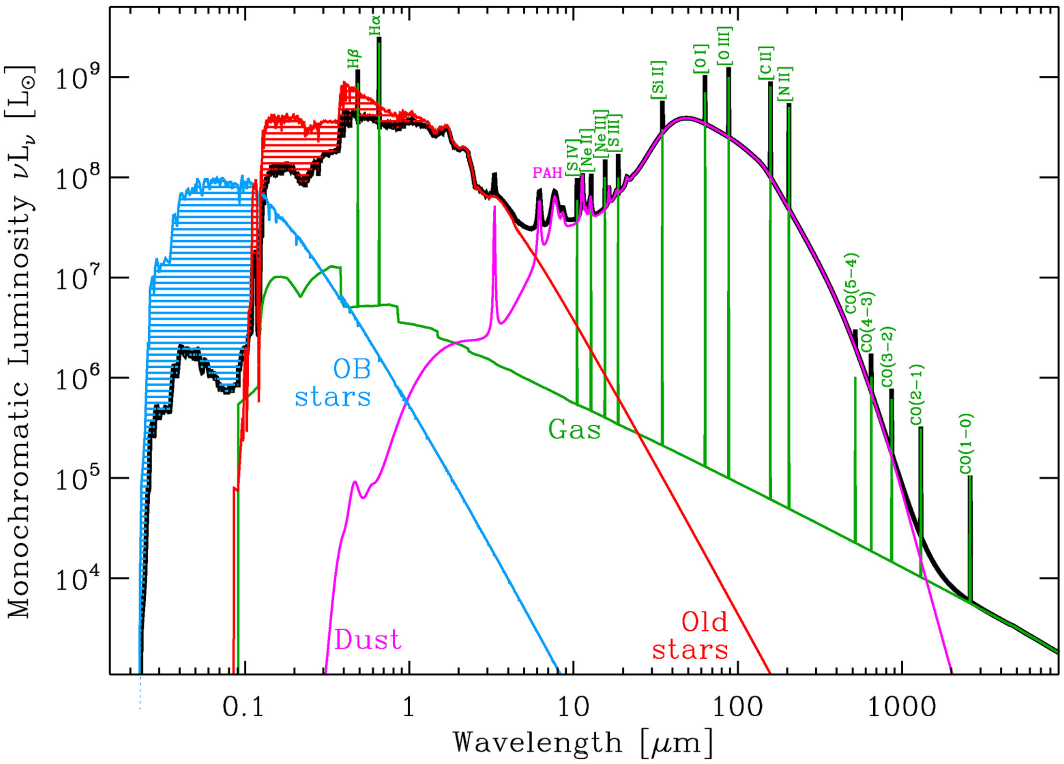

Figure 26: Typical SED of a star forming galaxy. It is based on the model of NGC 1569 by Galliano et al. (2008), with the addition of a typical gas contribution, modelled with Cloudy (Ferland et al., 2013). Only the most widely observed gas lines have been shown. The hatched areas in the stellar spectra represent the fraction of light that has been absorbed by the ISM (mainly the dust). This absorbed light is re-emitted by the dust component. Figure taken from Galliano (2017).

The spectral energy distribution (SED; see Figure 26) of a typical galaxy is dominated by several processes:

- **Stars.** Young OB stars are strong UV emitters, while older stellar populations dominate the visible/near-IR range. The weights of these two main components depend on the SF history of the galaxy. The fraction of light absorbed by dust is represented by the hatched area on Figure 26.
- **Gas.** Atomic and molecular lines are numerous, but the gas also emits continuum radiation (Figure 26): (i) thermal continuum in the form of free-bound and free-free emission; and (ii) non-thermal continuum, mainly synchrotron.
- **Dust.** Dust radiates thermally over the whole IR domain. Several molecular and solid state features can be seen, mainly in the mid-IR. Figure 26 represents the aromatic feature emission believed to be carried by polycyclic aromatic hydrocarbons (PAH). The whole power emitted by the dust equals the absorbed stellar power (hatched areas of Figure 26).

1.11.2 Additional context

Although only accounting for 1% of the ISM mass, dust plays a crucial role in galactic physics. It re-radiates in the infrared (IR) about 30% of the stellar power in normal disk galaxies, and up to 99% in ultra-luminous IR galaxies. It is a catalyst for numerous chemical reactions, including H_2 formation. The photoelectrons it releases in photodissociation regions (PDR) are one of the main heating sources of the gas. However, the detailed microscopic properties of the dust (its chemical composition, size distribution, abundance, etc.) are poorly known and are evidenced to vary strongly from one environment to the other. As a consequence, we are left with large uncertainties on the physics of the ISM, and on galaxy evolution.

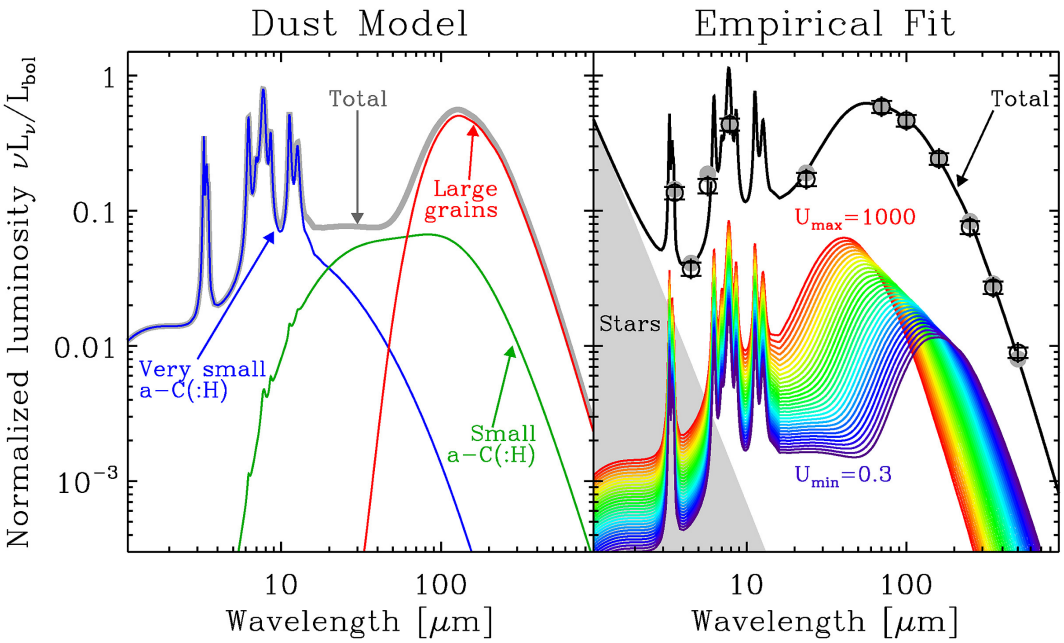

Figure 27: Dust SED modelling. **Left:** Dust model of Jones et al. (2017) uniformly illuminated by the solar neighborhood radiation field ($U = 1$). **Right:** An example of SED fit. The observations are the open circles with error bars (simulated data, for demonstration). The filled circles are the synthetic photometry of the model. The total model is the sum of uniformly illuminated SEDs with radiation field intensity ranging between $U_{\rm min}$ and $U_{\rm max}$ (rainbow curves). Figure taken from Galliano (2017).

In theory, we could infer dust properties by modelling its evolution from its formation in stellar ejecta and dense ISM, to its processing by shock and UV radiation and its recycling in star formation. However, the efficiency of each individual process is not accurately enough known to provide reliable grain properties, purely based on theory. We therefore have to rely on observations to constrain the grain properties in different environments.

Nearby galaxies are particularly suitable environments to conduct such studies, as they harbor a wide diversity of physical conditions (star formation activity, metallicity, etc.). In addition, a wealth of data is available for them, with good spatial resolution and sensitivity. The Milky Way itself is an important laboratory, but it spans a rather narrow range of metallicity and does not contain very massive star forming regions.

Dust and stars are not uniformly mixed. Knowing which stellar population is responsible for the dust heating in a given regime can be observationnally inferred. For instance, by comparing select far-IR Herschel band ratios to tracers of the stellar populations (young: 24 μm and $H\alpha$; old: 3.6 μm), it has been shown that the transition between dust heated by young stars and dust heated by old stars happens between 160 and 350 μm, in most cases. It appears that dust grains are on average hotter when heated by young stars, and colder when heated by older populations. The picture is different in more extreme objects, like low-metallicity dwarf galaxies, where the young stellar population can dominate the whole emission.

A more comprehensive approach to this problem is provided by panchromatic radiative transfer models. Such codes can solve the radiative transfer equation into a complex 3D spatial distribution (thin and thick disks, bulges, clumps, etc.). They estimate the dust heating in every region of the galaxy and compute the resulting escaping SED (far-IR optical depths are small). Studies have been able to reproduce the morphology of the galaxy at different wavelengths, however they were left with a deficit in emission in the far-IR. These discrepancies could be due to: (i) a lack of constraint on the 3D structure of these edge-on galaxies; (ii the presence of compact cold clumps; (iii) a higher far-IR grain emissivity.

The left panel of Figure 27 shows the dust mixture, heated by the solar neighborhood radiation field ($U = 1$). The far-IR bump is emitted by the large grains (silicates and a $-$ C(: H)), at thermal equilibrium with the radiation field. The mid-IR continuum is carried by out-of-equilibrium small a $-$ C(: H) (radius $a \lesssim 20$ nm). The aromatic features are carried by the smallest a $-$ C(: H) (radius $a \lesssim 1.5$ nm).

Such a dust model can not be used, as is, to fit the SED of galaxies, since there can be significant mixing of physical conditions in the beam or along the line of sight. Ideally, we should model the radiative transfer inside the object, but we usually lack the knowledge of its actual 3D structure. An alternative is to empirically account for the mixing, focusing only on quantities that are weakly dependent on radiative transfer processes.

The far-IR peak is mainly emitted by large grains at thermal equilibrium. Thus, its spectrum does not depend on the spectral shape of the incident radiation field, as it depends only on the total absorbed power. On the contrary, the spectral shape is important for small stochastically heated grains, as their temperature distribution depends on the mean photon energy. Fortunately, these grains do not account for a large fraction of the mass. However, it is not the case for the carriers of the aromatic features, since they are effectively heated by a narrower wavelength range of photons.
SED models being highly non-linear, several degeneracies and biases are encountered with a classical χ^2 minimization fit. This is well-known for modified black body models, where the monochromatic luminosity is parameterized by the dust mass ($M_{\rm dust}$), the equilibrium temperature ($T_{\rm dust}$), and the emissivity index (β):

$$L_\nu = M_{\rm dust} 4\pi\kappa_0 \left(\frac{\nu}{\nu_0}\right)^\beta B_\nu(T_{\rm dust}, \nu) \text{ [W]},$$

where κ_0 is the opacity at frequency ν, and B_ν is the Planck function. There are also biases induced by our ignorance of the origin of certain physical processes. In particular, the "**submm excess**" is an emission excess particularly strong beyond 500 μm in low-metallicity environments, but it has also been detected in the MW. It can not be accounted for by regular dust models, free-free, synchrotron and molecular line emission. Its origin is still debated: (i) very cold dust, although unlikely; (ii) magnetic grains; (iii) temperature dependent grain emissivity; or (iv) intrinsic grain optical properties. The only way to avoid this excess is to not use constraints beyond 500 μm.
Figure 28 demonstrates a panchromatic SED model applied to a galaxy and Figure 29 illustrates its geometry.

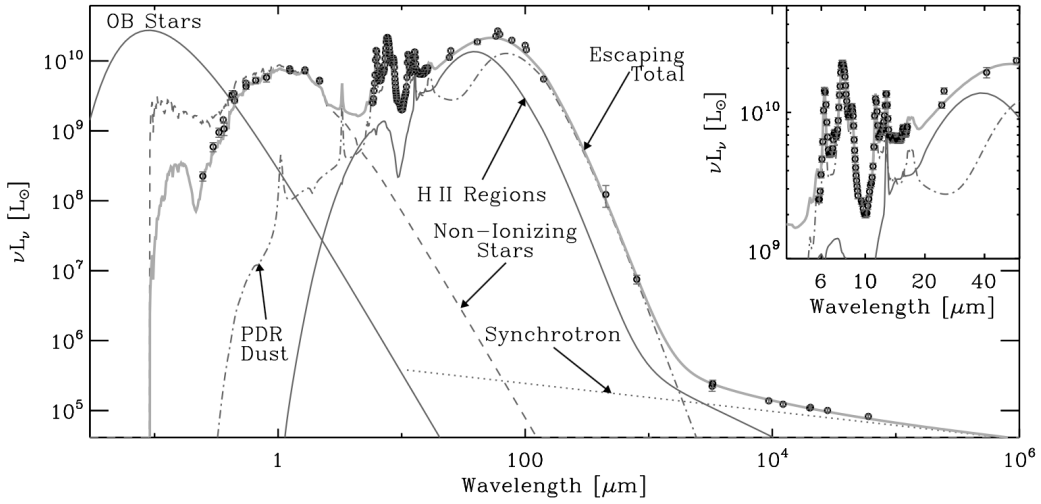

Figure 28: Demonstration of a panchromatic SED model applied to a galaxy. Figure taken from Galliano (2008).

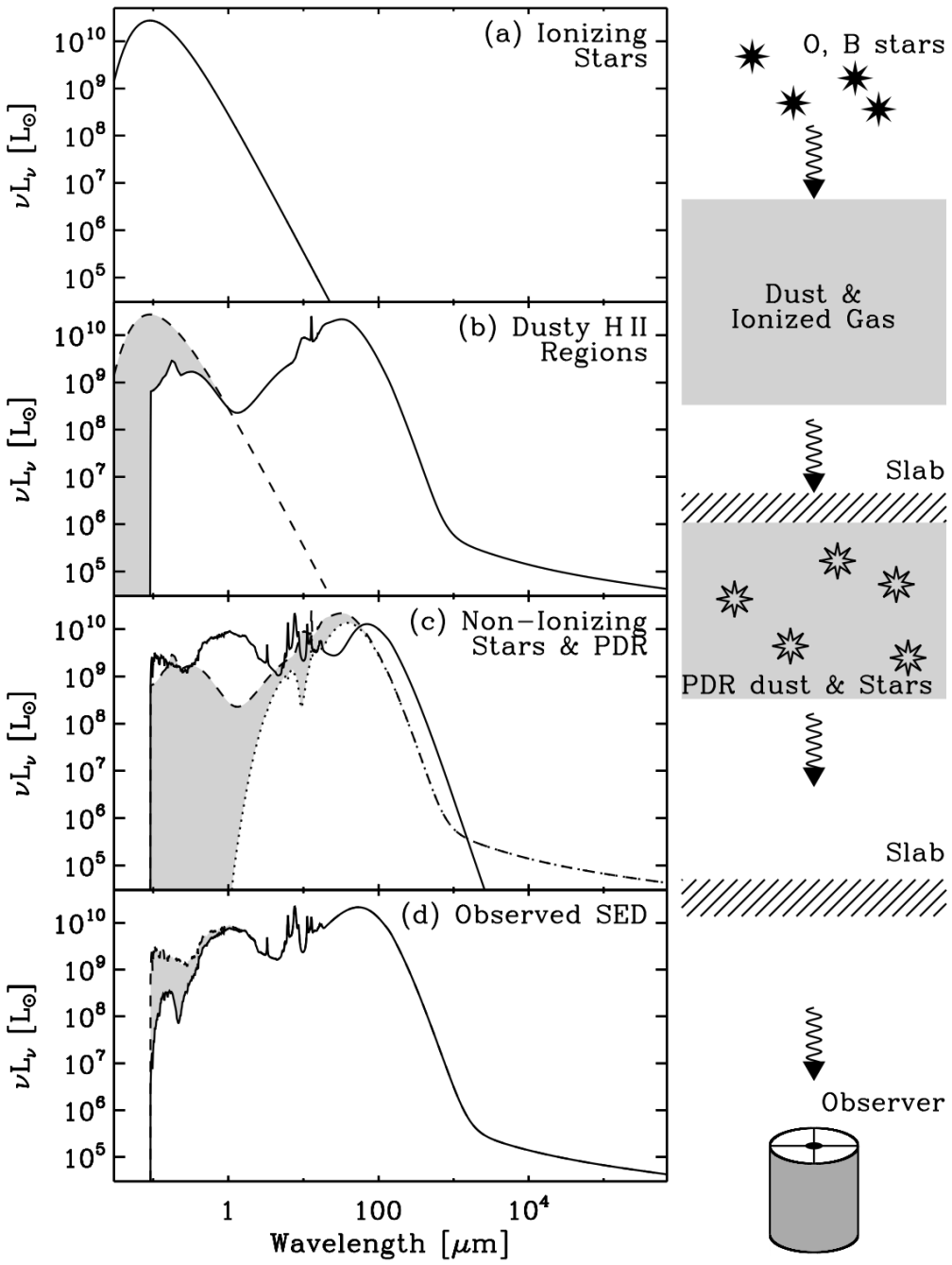

Figure 29: Illustration of the geometry of a panchromatic SED model. The left panels, from top to bottom, show the combination of the various SED building blocks moving from the massive star clusters to the observer. The solid lines are the total SED at each step; the dashed lines are the SED of the previous step; and the part of the SED that has been absorbed is shown in gray. The right panel illustrates the path of the photons from the star clusters to the observer. Figure taken from Galliano (2008).

1.11.3 Follow-up Questions

- How do the relative heights of the optical/FIR peaks change?

1.12 Question 12

How many stars does one expect to find within 100 pc of the Sun? If all stars are distributed evenly across the galaxy, how many of these will be B spectral type or earlier? How many are younger than 100 Myrs?

1.12.1 Short answer

Answer.

1.12.2 Additional context

Additional context.

1.12.3 Follow-up Questions

- Justify the assumptions made and explain why they do not match observations (e.g., number density of stars in the MW is not a flat distribution, the SFR isn't constant etc.).
- Where are most B-type and other early-type stars actually found?
- How do we know how many stars are in the MW?
- How do we measure the IMF?
- Are high- or low-mass stars more important to constrain the total number of stars?
- How many B stars are visible from your backyard? Are there any star forming regions visible from your backyard?

1.13 Question 13

Describe what happens as a cloud starts to collapse and form a star. What is the difference between the collapse and contraction stages? What happens to the internal temperature in both? When does the contraction phase end, and why does the end point depend on the mass of the object?

1.13.1 Short answer

Answer.

1.13.2 Additional context

Gravity is responsible for gathering gas into self-gravitating structures ranging in size from stars to giant molecular cloud complexes. Star formation involves extreme compression: part of a gas cloud collapses from a size $\sim 10^{18}\,\mathrm{cm}$ down to a stellar size, $\sim 10^{11}\,\mathrm{cm}$, with an accompanying increase in density by a factor $\sim 10^{21}$.
Here, we consider the conditions necessary for gravitational collapse to occur. There are several barriers to gravitational collapse. Gravity must of course overcome the resistance of pressure, both gas pressure and magnetic pressure. If the collapse is to produce a huge increase in density (as is necessary to form a star), then nearly all of the angular momentum in the collapsing gas must be transferred to nearby material. Last, the observed magnetic fields of young stars require that most of the magnetic field lines initially present in the gas not be swept into the forming protostar.
The **Jeans instability** occurs for a **Jeans length**

$$\lambda > \lambda_J \equiv \frac{2\pi}{\kappa_J} = \sqrt{\frac{\pi c_s^2}{G\rho_0}}\ [\mathrm{m}],$$

where

$$k_J \equiv \sqrt{\frac{4\pi G\rho_0}{c_s^2}}\ [\mathrm{m}^{-1}]$$

.
The **Jeans mass** is defined as

$$\begin{aligned} M_J &\equiv \frac{4\pi}{3}\rho_0\left(\frac{\lambda_J}{2}\right)^3\ [\mathrm{M}_\odot] \\ &= \frac{1}{8}\left(\frac{\pi kT}{\mu G}\right)^{3/2}\rho_0^{-1/2}\ [\mathrm{M}_\odot] \\ &= 0.32\left(\frac{T}{10\,\mathrm{K}}\right)^{3/2}\left(\frac{m_\mathrm{H}}{\mu}\right)^{3/2}\left(\frac{10^6\,\mathrm{cm}^{-3}}{n_\mathrm{H}}\right)^{1/2}\ [\mathrm{M}_\odot]. \end{aligned}$$

It is gratifying that when we substitute densities and temperatures observed for quiescent dark clouds, we find a mass typical of stars! If we plug in some typical numbers for a GMC, $c_s = 0.2\,\mathrm{km\,s^{-1}}$ and $\rho_0 = 100\,m_p$, we get $\lambda_J = 3.4\,\mathrm{pc}$. Since every GMC we have seen is larger than this size, and there are clearly always perturbations present, this means that molecular clouds cannot be stabilized by gas pressure against collapse.
In the limit of $k \ll k_J$ (long wavelength), the exponentiation time or "**growth time**" is

$$\tau_J = \frac{1}{k_J c_s} = \frac{1}{\sqrt{4\pi G\rho_0}} = 2.3\times10^4\left(\frac{n_\mathrm{H}}{10^6\,\mathrm{cm}^{-3}}\right)^{-1/2}\ [\mathrm{yr}].$$

To understand the value of the growth time τ_J, note that a uniform pressureless sphere of initially stationary gas with density ρ_0 will collapse with all shells reaching the center simultaneously in a finite time known as the **free-fall time**

$$\tau_\mathrm{ff} = \sqrt{\frac{3\pi}{32G\rho_0}} = 4.4\times10^4\left(\frac{n_\mathrm{H}}{10^6\,\mathrm{cm}^{-3}}\right)^{-1/2}\ [\mathrm{yr}].$$

The free-fall time τ_ff is only slightly longer than the Jeans growth time τ_J.
Self-gravitating density peaks within an isolated "dark cloud" are usually referred to as **cores**. The cores have masses of order $0.3\,\mathrm{M}_\odot$ to $10\,\mathrm{M}_\odot$. Each core is likely to form a single star or a binary star.
In the case of GMCs, the term **clump** is used to refer to self-gravitating regions with masses as large as $\sim 10^3\,\mathrm{M}_\odot$. Clumps may or may not be forming stars; those that are, are termed **star-forming clumps**. Such clumps will generally contain a number of cores.

When a core becomes gravitationally unstable, it will begin to collapse. Exactly how this collapse proceeds is uncertain in detail, but we think we understand the overall outlines.
During the initial stages, radiative cooling in molecular lines is able to keep the gas cool. As a result, the gas pressure remains unimportant during this phase, and the matter moves inward nearly in free-fall. The velocities at this stage are not large,

$$v \lesssim \sqrt{\frac{GM_c}{R_c}} = \left(\frac{4\pi}{3}\right)^{1/3} G^{1/2} M^{1/3} \rho_c^{1/6}$$
$$\approx 0.4 \left(\frac{M_c}{\mathrm{M}_\odot}\right)^{1/3} n_6^{1/6} \ [\mathrm{km\,s^{-1}}],$$

where M_c and $\rho_c = 1.4 n_{\mathrm{H}} m_{\mathrm{H}}$ are the mass and density of the core, and $n_{\mathrm{H}} = 10^6 n^6 \, cm^{-3}$.
Because the density is higher in the interior, the free-fall time is shortest there, and the collapse proceeds in an "inside-out" manner, with the center collapsing first, and the outer material later falling onto the central matter.
If cores had no angular momentum, and if magnetic fields were negligible, the collapse process would be relatively simple to understand and model. However, molecular clouds appear to have magnetic energies comparable to the kinetic energy (contributed mainly by the "turbulent" motions), and sufficient angular momentum to become dynamically important long before stellar densities are reached.
The infalling gas will generally have nonzero angular momentum, and (if it remains cold) the material will collapse to form a rotationally supported disk, with the material with the lowest specific angular momentum collected in a "**protostar**" at the center of the disk. Energy is dissipated as the infalling gas hits the disk. Angular momentum transport (due to the **magneto-rotational instability** (MRI) if the gas is sufficiently ionized, or due to gravitational torques or turbulent viscosity if the ionization is too low to support the MRI) will cause some material in the disk to move inward, with additional release of gravitational energy. The energy so released will heat the disk, and will be radiated away.
The dominant sources of energy are (1) the gravitational energy released as material is added to the protostar and as the protostar contracts, and (2) the energy released when the protostar is able to ignite fusion reactions to first "burn" deuterium, and then hydrogen. The protostar will have a significant luminosity, allowing it and the surrounding core to be observed as a luminous infrared source.
Thermodynamics of a Collapsing Core: We will now focus on the structure and evolution of protostars. Our goal will be to understand when and why collapse stops to form a pressure-supported object, and how those objects subsequently evolve into main sequence stars. This chapter focuses on the dynamics and thermal behavior of the material at the center of a collapsing core as it settles into something we can describe as a star, and on the structure of the envelope around this protostar. Later we will focus on the evolution of this object, both internally and in its appearance on the HR diagram.
Thermodynamics of a Collapsing Core: We will begin by considering what happens at the center of a collapsing core where the density in the center is rising rapidly as it collapses. We would like to understand the structure forming at the center of this collapsing object.
The Isothermal-Adiabatic Transition: The first important point to make is the assumption of isothermality for cores must break down at some point. Even at low density there are minor deviations from isothermality that result from the changeover in heating and cooling processes, but these are fairly minor, in the sense that they are unable to significantly affect collapse.
In contrast, if the gas is not able to radiate at all, it will behave adiabatically. This means it will approach a polytrope with $\gamma = 7/5$ or $\gamma = 5/3$, depending on whether the gas temperature is high enough to excite the internal quantum mechanical states of H_2 or not. (In actuality the γ for H_2 is more complicated than that, but that doesn't really matter for our purposes.) Either one is $> 4/3$, and thus sufficient to halt collapse.
Let us make some estimates of when deviations from isothermality that are significant enough to put us into this regime will occur. Since we are dealing with the collapse of the first region to fall in, we can probably safely assume that this material has very low angular momentum and treat the collapse as spherical – higher angular momentum material will only fall in later, since removal of angular momentum by the disk takes a while.
Let e_{th} be the **thermal energy per unit mass** of a particular gas parcel, and let Γ and Λ be the **rates of change in $\mathbf{e_{th}}$ due to heating and cooling processes**, i.e.:

$$\frac{\mathrm{d}e_{\mathrm{th}}}{\mathrm{d}} = \Gamma - \Lambda \ [\mathrm{J\,kg^{-1}\,s^{-1}}].$$

At high densities inside a core immediately before a central star forms and begins to radiate, the dominant source of energy is adiabatic compression of the gas. The first law of thermodynamics tells us that the **heating rate due to compression** is

$$\Gamma = -p\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{\rho}\right)\ [\mathrm{J\,kg^{-1}\,s^{-1}}],$$

where p and ρ are the gas pressure and density. Since $1/\rho$ is the specific volume, meaning the volume per unit mass occupied by the gas, this term is just $p\,\mathrm{d}V$, the work done on the gas in compressing it. If the gas is collapsing in free-fall, the compression timescale is about the free-fall timescale, $t_{\mathrm{ff}} \sim 1/\sqrt{G\rho}$, so we expect

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{\rho}\right) = C_1\sqrt{\frac{4\pi G}{\rho}}$$

where C_1 is a number of order unity that will depend on the exact collapse solution, and the factor of $\sqrt{4\pi}$ has been inserted for future convenience. Writing $p = \rho c_s^2$ and plugging this into the heating rate, we get

$$\Gamma = C_1 c_s^2\sqrt{4\pi G\rho}\ [\mathrm{J\,kg^{-1}\,s^{-1}}].$$

The main cooling source is thermal emission by dust grains, which at the high densities with which we are concerned are thermally very well coupled to the gas. Let us first consider the gas where the gas is optically thin to this thermal radiation. In this case the cooling rate per unit mass is simply given by the rate of thermal emission,

$$\Lambda_{\mathrm{thin}} = 4\kappa_P\sigma T^4\ [\mathrm{J\,kg^{-1}\,s^{-1}}],$$

where σ is the Stefan-Boltzmann constant and κ_P is the Planck mean opacity of the dust grains. As long as $\Lambda > \Gamma$, the gas will remain isothermal. (Strictly speaking if $\Lambda > \Gamma$ the gas will cool, but that's because we've left out other sources of heating, such as cosmic rays and the fact that the protostar is bathed in a background IR radiation field from other stars.)
If we equate the heating and cooling rates, using for T the temperature in the isothermal gas, we therefore will obtain a characteristic density beyond which the gas can no longer remain isothermal. Doing so gives

$$\begin{aligned}\rho_{\mathrm{thin}} &= \frac{4}{\pi}\frac{\kappa_P^2\sigma^2\mu^2T^4}{C_1^2Gk_B^2}\\ &= 5\times10^{-15}C_1^{-2}\left(\frac{100\,\kappa_P}{\mathrm{cm^2\,g^{-1}}}\right)^2\left(\frac{T}{10\,\mathrm{K}}\right)^6\ [\mathrm{g\,cm^{-3}}],\end{aligned}$$

where μ is the mean mass per particle and we have set $c_s = \sqrt{k_BT/\mu}$. Thus we find that compressional heating and optically thin cooling to balance at about $10^{14}\,\mathrm{g\,cm^{-3}}$.
A second important density is the one at which the gas starts to become optically thick to its own re-emitted IR radiation. Suppose that the optically thick region at the center of our core has some mean density ρ and radius R. The condition that the optical depth across it be unity then reduces to

$$2\kappa\rho R \approx 1.$$

If this central region corresponds to the size of the region that is no longer in free-fall collapse and is instead thermally supported, then its size must be comparable to the Jeans length at its lowest temperature, i.e., $R \sim \lambda_J = \sqrt{\pi c_s^2/G\rho}$. Thus we set

$$R = C_2\frac{2\pi c_s}{\sqrt{4\pi G\rho}}\ [\mathrm{pc}],$$

where C2 is again a constant of order unity.
Plugging this into the condition for optical depth unity, we derive the **characteristic density** at which the gas transitions from optically thin to optically thick:

$$\begin{aligned}\rho_{\tau\sim1} &= \frac{1}{4\pi}C_2^{-2}\frac{\mu G}{\kappa_P^2k_BT}\\ &= 1.5\times10^{-13}\left(\frac{100\,\kappa_P}{\mathrm{cm^2\,g^{-1}}}\right)^{-2}\left(\frac{T}{10\,\mathrm{K}}\right)^{-1}\ [\mathrm{g\,cm^{-3}}].\end{aligned}$$

This is not very different from the value for ρ_{thin}, so in general for reasonable collapse conditions we expect that cores transition from isothermal to close to adiabatic at a density of $\sim 10^{-13} - 10^{-14}\,\mathrm{g\,cm^{-3}}$. It is worth noting that ratio of ρ_{thin} to $\rho_{\tau\sim1}$ depends extremely strongly on both κ_P (to the 4th power) and T (to the 7th), so any small change in either can render them very different. For example, if the metallicity is super-solar then κ_P will be larger, which will increase ρ_{thin} and decrease $\rho_{\tau\sim1}$ Similarly, if

the region is somewhat warmer, for example due to the presence of nearby massive stars, then $\rho_{\rm thin}$ will increase and $\rho_{\tau\sim1}$ will decrease.
If $\rho_{\tau\sim1} < \rho_{\rm thin}$ the collapsing gas will become optically thick before heating becomes faster than optically thin cooling. In this case we must compare the heating rate due to compression with the cooling rate due to optically thick cooling instead of optically thin cooling. Optically thick cooling is determined by the rate at which radiation can diffuse out of the core. If we have a central region of optical depth $\tau \gg 1$, the effective **speed of the radiation** moving through it is c/τ, so the time required for the radiation to diffuse out is

$$t_{\rm diff} = \frac{\ell\tau}{c} = \frac{\kappa_P \rho \ell^2}{c} \text{ [yr]},$$

where ℓ is the **characteristic size** of the core.
Inside the optically thick region matter and radiation are in thermal balance, so the radiation energy density approaches the blackbody value aT^4. The radiation energy per unit mass is therefore aT^4/ρ. Putting all this together, and taking $\ell = 2R$ as we did before in computing $\rho_{\tau\sim1}$, the optically thick cooling rate per unit mass is

$$\Lambda_{\rm thick} = \frac{aT^4\rho}{t_{\rm diff}} = \frac{\sigma T^4}{\kappa_P \rho^2 R^2} \text{ [J kg}^{-1}\text{ s}^{-1}\text{]},$$

where $\sigma = ca/4$. If we equate $\Lambda_{\rm thick}$ and Γ, we get the characteristic density where the gas becomes non-isothermal in the optically thick regime

$$\begin{aligned}\rho_{\rm thick} &= \left(\frac{C_1^2 G \sigma^2 \mu^4 T^4}{4\pi^3 C_2^4 k_B^4 \kappa_P^2}\right)^{1/3} \\ &= 5\times10^{-14}\left(\frac{C_1^{2/3}}{C_2^{4/3}}\right)\left(\frac{100\,\kappa_P}{\rm cm^2\,g^{-1}}\right)^{-2/3}\left(\frac{T}{10\,\rm K}\right)^{4/3} \text{ [g cm}^{-3}\text{]}\end{aligned}$$

This is much more weakly dependent on κ_P and T, so we can now make the somewhat more general statement that, even for super-solar metallicity or warmer regions, we expect a transition from isothermal to adiabatic behavior somewhere in the vicinity of $10^{-14} - 10^{-13}\,\rm g\,cm^{-3}$.
The first core: The transition to an adiabatic equation of state, with $\gamma > 4/3$, means that the collapse must at least temporarily halt. The result will be a hydrostatic object that is supported by its own internal pressure. This object is known as the **first core**, or sometimes a **Larson's first core**, after Richard Larson, who first predicted this phenomenon.
We can model the first core reasonably well as a simple polytrope, with index n defined by $n = 1/(\gamma-1)$. At low mass when the temperature in the first core is low $\gamma \approx 5/3$ and $\gamma \approx 3/2$, and for a more massive, warmer core $\gamma \approx 7/5$ ($n \approx 5/2$). For a polytrope of central density ρ_c, the radius and mass are

$$\begin{aligned}R &= a\xi_1 \text{ [pc]} \\ M &= -4\pi a^3 \rho_c \left(\xi^2 \frac{{\rm d}\theta}{{\rm d}\xi}\right)_1 \text{ [M}_\odot\text{]},\end{aligned}$$

where $\xi = r/a$ is the **dimensionless radius**, $\theta = (\rho/\rho_c)^{1/n}$ is the **dimensionless density**, the subscript 1 refers to the value at the edge of the sphere (where $\theta = 0$), the factors ξ_1 and $(\xi {\rm d}\theta/{\rm d}\xi)_1$ can be determined by integrating the **Lane-Emden equation**, and the scale factor a is defined by

$$a = \sqrt{\frac{(n+1)K}{4\pi G}\rho_c^{(1-n)/n}} \text{ [pc]}.$$

The factor $K = p/\rho^\gamma$ is the **polytropic constant**, which is determined by the specific entropy of the gas.
For our first core, the specific entropy will just be determined by the density at which the gas transitions from isothermal to adiabatic. If we let $\rho_{\rm ad}$ be the density at which the gas becomes adiabatic, then the pressure at this density is $p = \rho_{\rm ad} c_{s0}^2$, where c_{s0} is the sound speed in the isothermal phase, and $K = c_{s0}^2 \rho_{\rm ad}^{1-\gamma}$. For $\gamma = 5/3$ ($n = 1.5$) we have $\xi_1 = 3.65$ and $(\xi {\rm d}\theta/{\rm d}\xi)_1 = 2.71$, and plugging in we get

$$\begin{aligned}R &= 2.2\left(\frac{10^{10}\rho_c}{\rm g\,cm^{-3}}\right)^{1/6}\left(\frac{T}{10\,\rm K}\right)^{1/2}\left(\frac{10^{13}\rho_{\rm ad}}{\rm cm^{-3}}\right)^{-1/3} \text{ [AU]} \\ M &= 0.059\left(\frac{10^{10}\rho_c}{\rm g\,cm^{-3}}\right)^{1/6}\left(\frac{T}{10\,\rm K}\right)^{1/2}\left(\frac{10^{13}\rho_{\rm ad}}{\rm cm^{-3}}\right)^{-1/3} \text{ [M}_\odot\text{]}.\end{aligned}$$

We can obtain very similar numbers by plugging in for $\gamma = 7/5$ ($n = 2.5$). These results show that the first core is an object a few AU in size, with a mass of a few hundredths of a solar mass.

Second collapse: The first core is a very short-lived phase in the evolution of the protostar. To see why, let us estimate its temperature. The temperature inside the sphere rises as $T \propto \rho^{\gamma-1}$, so the central temperature is

$$T_c = T_0 \left(\frac{\rho_c}{\rho_{\rm ad}} \right)^{\gamma-1} \ [\rm K],$$

where T_0 is the temperature in the isothermal phase. Thus the central temperature will be higher than the boundary temperature by a factor that is determined by how high the central density has risen, which in turn will be determined by the amount of mass that has accumulated on the core.

In general we have $M \propto rho_c^{(3+n)/(2n)}$, or $M \propto \rho_c^{(3\gamma-2)/2}$. We also have $T_c \propto \rho_c^{\gamma-1}$. Combining these results, we get

$$T_c \propto M^{(2\gamma-2)/(3\gamma-2)} \ [\rm K].$$

The exponent is 0.44 for $\gamma = 5/3$ and 0.36 for $\gamma = 7/5$.

Plugging in some numbers, $M = 0.06\,{\rm M}_\odot$, $\rho_{\rm ad} = 10^{-13}\,{\rm g\,cm^{-3}}$, and $\gamma = 5/3$ gives $\rho_c = 10^{-10}{\rm g\,cm^{-3}}$ and $T_c = 1000\,{\rm K}$. Thus we see that by the time anything like $M = 0.1\,{\rm M}_\odot$ of material has accumulated on the first core, compression will have caused its central temperature to rise to 1000 K or more.

This causes yet another change in the thermodynamics of the gas, because all the hydrogen is still molecular, and molecular hydrogen has a binding energy of 4.5 eV. In comparison, the kinetic energy per molecule for molecular hydrogen at a temperature T is $3k_BT = 0.26T_3\,{\rm eV}$, where $T_3 = T/(1000\,{\rm K})$. At 1000 K this means that the mean molecule still has only $\sim 5\%$ of the kinetic energy that would be required to dissociate it. However, there is a non-negligible tail of the Maxwellian distribution that is moving fast enough for collisions to produce dissociation. Each of these dissociative collisions removes 4.5 eV from the kinetic energy budget of the gas and puts it into chemical energy instead. Since dissociations are occurring on the tail of the Maxwellian, any slight increase in the temperature dramatically increases the dissociation rate, moving even more kinetic energy into chemical energy.

This effectively acts as a thermostat for the gas, in much the same way that a boiling pot of water stays near the boiling temperature of water even when energy is added, because all the extra energy that is provided goes into changing the chemical state of the water rather than raising its temperature. Detailed numerical calculations of this effect show that at temperatures above $1000 - 2000\,{\rm K}$, the equation of state becomes closer to $T \propto \rho^{0.1}$, or $\gamma = 1.1$. This is again below the critical value of $\gamma = 4/3$ required to have a hydrostatic object, and as a result the center of the first core again goes into something like free-fall collapse.

This is called the **second collapse**. The time required for it is set by the free-fall time at the central density of the first core, which is only a few years. This collapse continues until all the hydrogen dissociates. The hydrogen also ionizes during this collapse, since the ionization potential of 13.6 eV isn't very different from the dissociation potential of 4.5 eV. Only once all the hydrogen is dissociated and ionized can a new hydrostatic object form.

At this point the gas is warmer than $\sim 10^4\,{\rm K}$, is fully ionized, and the new hydrostatic object is a true protostar. It is supported by degeneracy pressure at first when its mass is low, and then as more mass arrives it heats up and becomes supported by thermal pressure.

An important point to make there is that this discussion implies that brown dwarfs, at least those of sufficiently low mass, do not undergo a prompt second collapse. Instead, their first cores never accumulate enough mass to dissociate the molecules at their center. This isn't to say that dissociation never happens in them, and that second collapse never occurs. A brown dwarf-mass first core will still radiate from its surface and, lacking any internal energy source, this energy loss will have to be balanced by compression. As the gas compresses the temperature and entropy will rise, and, if the object does not become supported by degeneracy pressure first, the central temperature will eventually rise enough to produce second collapse. The difference for a brown dwarf is that this will only occur once slow radiative losses cause a temperature rise, which may take a very long time compared to formation. For stars, in contrast, there is enough mass to reach the critical temperature by compression during formation.

Evolutionary phases for protostars: There are generally a few distinct stages though with forming stars pass, which can be read off from how the radius evolves as the star gains mass. We will use as our primary example the case of a star undergoing hot accretion at $10^{-5}\,{\rm M}_\odot\,{\rm yr}^{-1}$, as illustrated in Figure 30. However, note that the ordering of these phases we'll describe below can vary somewhat depending on the accretion rate and the boundary conditions assumed. Moreover, for low mass stars, some of the later phases may not occur at all, or occur only after the end of accretion.

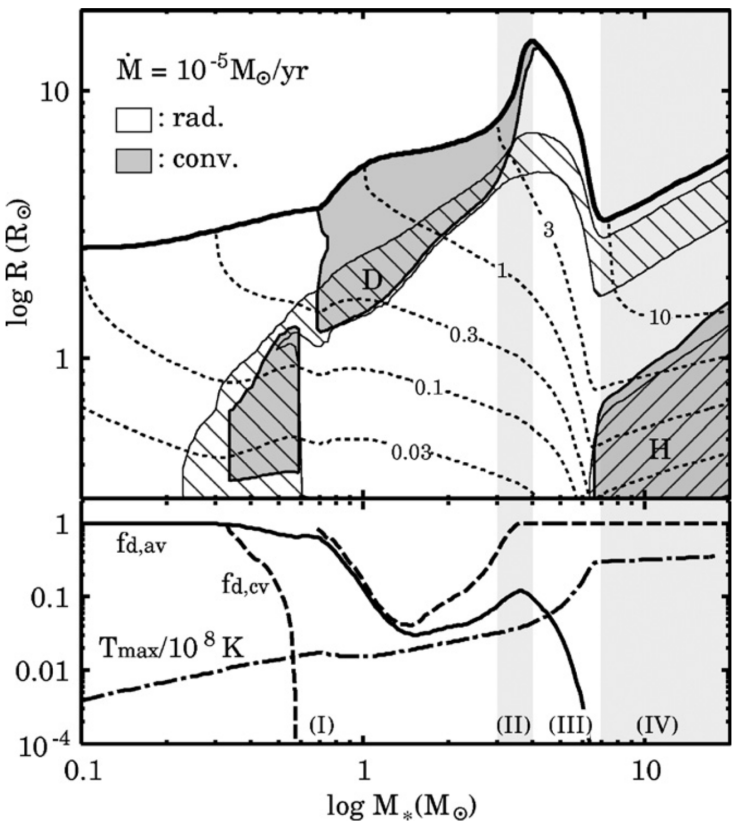

Figure 30: Kippenhahn and composition diagrams for a protostar accreting at $10^{-5}\,M_\odot\,yr^{-1}$. In the top panel, the thick curve shows the protostellar radius as a function of mass, and gray and white bands show convective and radiative regions, respectively. Hatched areas show regions of D and H burning, as indicated. Thin dotted lines show the radii containing 0.1, 0.3, 1, 3, and $10\,M_\odot$, as indicated. Shaded regions show four evolutionary phases: (I) convection, (II) swelling, (III) KH-contraction, and (IV) the main sequence. In the lower panel, the solid line shows the mean deuterium fraction in the star, normalized to the starting value, while the dashed line shows the D fraction only considering the convective parts of the star. The dot-dashed line shows the maximum temperature. Table taken from Krumholz (2015).

The initial phase of evolution is visible in Figure 30 as what takes place up to a mass of $\sim 0.2\,M_\odot$ for the example shown. The first thing that happens during this phase is that the star reaches a radius that is a function solely of M_* and $\dot{M}_*$. This occurs regardless of the initial radius with which we initiate the model, as long as we are using the hot accretion boundary condition. The physical reason for this behavior is easy to understand. The radius of the star is determined by the entropy profile. High entropy leads to high radius. Since the internal energy generated by the star is small compared to the accretion power when the stellar mass is low (i.e., $L_{\rm bb} \ll L_{\rm acc}$), once gas is incorporated into the star it does not lose significant energy by radiation. The only entropy it loses is due to the radiation that occurs at the shock on the star's surface. We could have guessed this result from the large value of $t_{\rm KH}$ compared to the accretion time – in effect, this means that, once a fluid element reaches the stellar surface it will be buried and reach a nearly constant entropy quite quickly. Consequently, we can treat the material falling onto the star during this phase as having an entropy per unit mass that depends only on two factors: (1) the entropy it acquires by striking the stellar surface, and (2) how much it radiates before being buried. The latter factor is just determined by the accretion rate. Higher accretion rates bury accreted material more quickly, leaving it with higher entropy and producing larger radii. The former depends on the velocity of the infalling material just before it strikes the stellar surface, and thus on $v_{\rm ff} \propto \sqrt{M_*/R_*}$. However, this second factor self-regulates. If at fixed M_*, R_* is very large, then $v_{\rm ff}$ is small, and the incoming material gains very little entropy in the shock. Small entropy leads to a smaller radius. Conversely, if R_* is very small, then $v_{\rm ff}$ and the post-shock entropy will be large, and this will produce rapid swelling of the protostar. This effect means that the radius rapidly converges to a value that depends only on M_* and $\dot{M}_*$.

This self-regulation does not happen if the material is assumed undergoing cold accretion. In this case, the radial evolution of the star is determined solely by the amount of entropy that is assumed to remain in the accretion flow when it joins onto the star. One common practice is to assume that the entropy of the accreting material is equal to the entropy of the gas already in the star, and, under this assumption, the choice of initial condition completely determines the subsequent evolution, since the choice of initial condition then determines the entropy content of the star thereafter.

Regardless of the boundary condition assumed, during this phase there is no nuclear burning in the star, as the interior is too cold for any such activity. Since there is no nuclear burning, and this phase generally lasts much less than the Kelvin-Helmholtz timescale on which radiation changes the star's structure, during this phase the entropy content of the star is nearly constant. This phase can therefore be referred to as the **adiabatic stage** in the star's evolution.

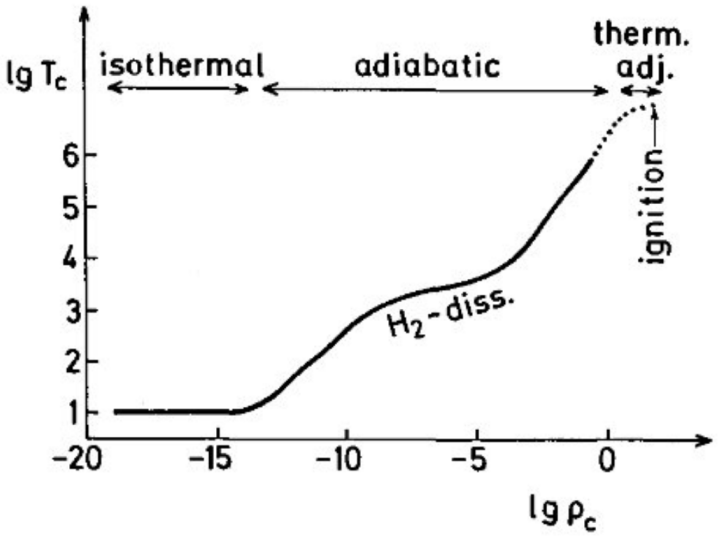

Figure 31: Evolution of a collapsing cloud on a $\rho - T$ diagram. Adiabatic collapse begins due to increased opacity, while H_2 dissociation reduces the temperature increase. Adiabatic collapse ends when the thermal adjustment time is smaller than the mass accretion timescale. During the left/upward travel near 100 K, most of the luminosity is accretion-powered. Once the flat region is reached, accretion has diminished enough that most of the energy comes from deuterium fusion and gravitational contraction Figure taken from Carroll & Ostlie (2006).

Deuterium ignition and convection: In Figure 30, the next evolutionary phase begins at $0.25\,M_\odot$, and continues to $0.7\,M_\odot$. This stage is marked by two distinct but interrelated phenomena: the onset of nuclear burning and the onset of convection. The driving force behind both phenomena is that, as the protostar gains mass, its interior temperature rises. For a polytrope, which is not an unreasonable description of the accreting protostar, the central temperature rises with mass to the $T_c \propto M^{(2\gamma-2)/(3\gamma-2)}$. Thus even at fixed entropy the central temperature must rise as the star gains mass.
Once T_c reaches $\sim 10^6\,K$, deuterium will ignite at the center of the protostar. This generally happens at a mass of hundredths to tenths of $M_\odot$, depending on the choice of accretion rate and boundary condition. This has two significant effects. The first is that deuterium acts as a thermostat for the star's center, much as hydrogen does in a main sequence star. Because the energy generation rate is so incredibly sensitive to T, any slight raise in the temperature causes it to jump enough to raise the pressure and adiabatically expand the star, reducing T. Thus, T_c becomes fixed at $10^6\,K$. The star adjusts its radius accordingly, which generally requires that the radius increase to keep T_c nearly constant as the mass rises. Thus deuterium burning temporarily halts core contraction. Both effects are visible in Figure 30. The halting of core contraction is apparent from the way the dotted lines showing constant mass enclosed bend upward at $\sim 0.3\,M$, and the nearly constant core temperature is visible from the fact that, between $\sim 0.25\,M$ and $3-4\,M$, a factor of more than 10 in mass, the central temperature stays within a factor of 2 of $10^6\,K$.
The second effect of deuterium burning that it causes a rapid rise in the entropy at the center of the star. This has the effect of starting up convection in the star. Before deuterium burning the star is generally stable against convection. That is because the entropy profile is determined by infall, and since shells that fall onto the star later arrive at higher velocities (due to the rising M_*), they have higher entropy. Thus s is an increasing function of M_r, which is the condition for convective stability. Deuterium burning reverses this, and convection follows, eventually turning much of the star convective. This also ensures the star a continuing supply of deuterium fuel, since convection will drag gas from the outer parts of the star down to the core, where they can be burned.
An important caveat here is that, although D burning encourages convection, it is not necessary for it. In the absence of D, or for very high accretion rates, the onset of convection is driven by the increasing luminosity of the stellar core as it undergoes KH contraction. This energy must be transported outwards, and as the star's mass rises and the luminosity goes up, eventually the energy that must be transported exceeds the ability of radiation to carry it. Convection results. For very high accretion rates, this effect drives the onset of convection even before the onset of D burning.
A third effect of the deuterium thermostat is that it forces the star to obey a nearly-linear mass-radius relation, and thus to obey a particular relationship between accretion rate and accretion luminosity. One can show that for a polytrope the central temperature and surface escape speed are related by

$$\psi = \frac{GM}{R} = \frac{1}{2}v_{\rm esc}^2 = T_n \frac{k_B T_c}{\mu m_{\rm H}}\ [{\rm erg\,g^{-1}}]$$

where T_c is the core temperature, T_n is a dimensionless constant of order unity that depends only on the polytropic index, and μ is the mean mass per particle in units of hydrogen masses. For $n = 3/2$, expected for a fully convective star, $T_n = 1.86$. Plugging in this value of T_n, $\mu = 0.61$ (the mean molecular

weight for a fully ionized gas of H and He in the standard abundance ratio), and $T_c = 10^6$ K, one obtains $\psi = 2.5 \times 10^{14}\,\mathrm{erg\,g^{-1}}$ as the energy yield from accretion.

Deuterium exhaustion and formation of a radiative barrier: The next evolutionary phase, which runs from $0.6-3\,\mathrm{M}_\odot$ in Figure 30, is marked by the exhaustion of deuterium in the stellar core. Deuterium can only hold up the star for a finite amount of time. The reason is simply that there isn't that much of it. Each deuterium burned provides 5.5 MeV of energy, comparable the 7 MeV per hydrogen provided by burning hydrogen, but there are only 2×10^{-5} D nuclei per H nuclei. Thus, at fixed luminosity the "main sequence" lifetime for D burning is shorter than that for H burning by a factor of $2 \times 10^{-5} \times 5.5/7 = 1.6 \times 10^{-5}$.

We therefore see that, while a main sequence star can burn hydrogen for 10^{10} yr, a comparable pre-main sequence star of the same mass and luminosity burning deuterium can only do it for only a few times 10^5 yr. To be more precise, the time required for a star to exhaust its deuterium is

$$t_\mathrm{D} = \frac{[\mathrm{D/H}]\Delta E_D M_*}{m_\mathrm{H} L_*} = 1.5 \times 10^6 [\mathrm{yr\,M_*L_{*,0}^{-1}}].$$

Thus deuterium burning will briefly hold up a star's contraction, but cannot delay it for long. However, a brief note is in order here: while this delay is not long compared to the lifetime of a star, it is comparable to the formation time of the star. Recall that typical accretion rates are of order a few times $10^{-6}\,M_\odot\,yr^{-1}$, so a $1\,\mathrm{M}_\odot$ star takes a few times 10^5 yr to form. Thus stars may burn deuterium for most of the time they are accreting.

The exhaustion of deuterium does not mean the end of deuterium burning, since fresh deuterium that is brought to the star as it continues accreting will still burn. Instead, the exhaustion of core deuterium happens for a more subtle reason. As the deuterium supply begins to run out, the rate of energy generation in the core becomes insufficient to prevent it from undergoing further contraction, leading to rising temperatures. The rise in central temperature lowers the opacity, which is govern by Kramers' law: $\kappa \propto \rho T^{-3.5}$. This in turn makes it easier for radiation to transport energy outward. Eventually this shuts off convection somewhere within the star, leading to formation of what is called a **radiative barrier**.

The formation of the barrier ends the transport of D to the stellar center. The tiny bit of D left in the core is quickly consumed, and, without D burning to drive an entropy gradient, convection shuts off through the entire core. This is the physics behind the nearly-simultaneous end of central D burning and central convection that occurs near 0.6 M in Figure 30. After this transition, the core is able to resume contraction, and D continues to burn as fast as it accretes. However, it now does so in a shell around the core rather than in the core.

Swelling: The next evolutionary phase, which occurs from $3-4\,\mathrm{M}$ in Figure 30, is swelling. This phase is marked by a marked increase in the star's radius over a relative short period of time. The physical mechanism driving this is the radiative barrier discussed above. The radiative barrier forms because increasing temperatures drive decreasing opacities, allowing more rapid transport of energy by radiation. The decreased opacity allows the center of the star to lose entropy rapidly, and the entropy to be transported to the outer parts of the star via radiation. The result is a wave of luminosity and entropy that propagates outward through the star.

Once the wave of luminosity and entropy gets near the stellar surface, which is not confined by the weight of overlying material, the surface undergoes a rapid expansion, leading to rapid swelling. The maximum radius, and the mass at which the swelling phase occurs, is a strong function of the accretion rate (Figure 32). However, even at very low accretion rates, swelling does not occur until the mass exceeds $1\,\mathrm{M}_\odot$.

Contraction to the Main Sequence: The final stage of protostellar evolution is contraction to the main sequence. Once the entropy wave hits the surface, the star is able to begin losing energy and entropy fairly quickly, and it resumes contraction. This marks the final phase of protostellar evolution, visible above $\sim 4\,\mathrm{M}_\odot$ in Figure 30. Contraction ends once the core temperature becomes hot enough to ignite hydrogen, landing the star at least on the main sequence.

Observable evolution of protostars: We have just discussed the interior behaviour of an evolving protostar. While this is important, it is also critical to predict the observable properties of the star during this evolutionary sequence. In particular, we wish to understand the star's luminosity and effective temperature, which dictate its location on the HR diagram. The required values can simply be read off from the evolutionary models (Figure 33), giving rise to a track of luminosity versus effective temperature in the HR diagram.

The birthline: Before delving into the tracks themselves, we have to ask what is actually observable. As long as a star is accreting from its parent core, it will probably not be visible in the optical, due to the high opacity of the dusty gas in the core. Thus we are most concerned with stars' appearance in the HR diagram only after they have finished their main accretion phase. We refer to stars that are still accreting and thus not generally optically-observable as **protostars**, and those that are in this post-accretion phase as **pre-main sequence stars**.

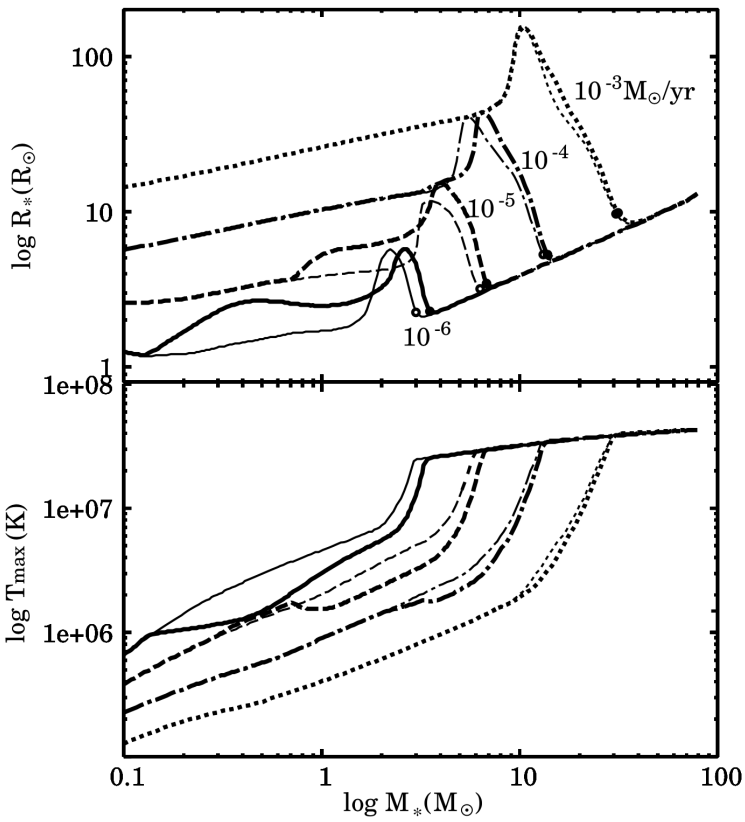

Figure 32: Radius versus mass (top) and maximum temperature versus mass (bottom) for protostars accretting at different rates. The accretion rate is indicated by the line style, as illustrated in the top panel. For each accretion rate there are two lines, one thick and one thin. The thick line is for the observed Milky Way deuterium abundance, while the thin line is the result assuming zero deuterium abundance. Figure taken from Krumholtz (2015).

For stars below 1 M, examining Figure 30, we see that the transition from protostar to pre-main sequence star will occur some time after the onset of deuterium burning, either during the core or shell burning phases depending on the mass and accretion history. More massive stars will become visible only during KH contraction, or even after the onset of hydrogen burning. The lowest mass stars might be observable even before the start of deuterium burning. However, for the majority of the pre-main sequence stars that we can observe, they first become visible during the D burning phase.

Since there is a strict mass-radius relation during core deuterium burning (with some variation due to varying accretion rates), there must be a corresponding relationship between L and T, just like the main sequence. We call this line in the HR diagram, on which protostars first appear, the **birthline**; see Figure 34. Since young stars are larger and more luminous that main sequence stars of the same mass, this line lies at higher L and lower T than the main sequence.

The Hayashi track: Now that we understand what is observable, let us turn to the tracks themselves. The tracks shown in Figures 33 and 34 show several distinct features. One is that, for low mass stars, the initial phases of evolution in the HR diagram are nearly vertically, i.e., at constant $T_{\rm eff}$. The vertical tracks for different masses are very close together. This vertical part of the evolution is called the **Hayashi track**, after its discoverer, who predicted it theoretically (Hayashi, 1961). For low mass stars, the majority of the Hayashi track lies after the birthline, so it is directly observable.

The origin of the Hayashi track is in the physics of opacity in stellar atmospheres at low temperature. At temperatures below about 10^4 K, hydrogen becomes neutral, and the only free electrons available come from metal atoms with lower ionization energies. Some of these electrons become bound with hydrogen atoms, forming H^-, and this ion is the dominant source of opacity. Thus the opacity depends on the number of free electrons provided by metal atoms, which in turn depends extremely sensitively on the temperature.

If the temperature falls too low, the opacity will be so low that, even integrating through the rest of the star's mass, the optical depth to infinity will be $< 2/3$. Since the photosphere must always be defined by a surface of optical depth unity, this effectively establishes a minimum surface temperature for the star required to maintain $\tau \sim 1$. This minimum temperature depends weakly on the star's mass and radius, but to good approximation it is simply $T_{\rm min} = T_H = 3500\,{\rm K}$, where T_H is the **Hayashi temperature**. Low mass protostars, due to their large radii, wind up right against this limit, which is why they all contract along vertical tracks that are packed close together in $T_{\rm eff}$.

The Heyney track: Contraction at nearly constant $T_{\rm eff}$ continues until the star contracts enough to raise its surface temperature above T_H. This increase in temperature also causes the star to transition from convective to radiative, since the opacity drops with temperature at high temperatures, and a lower

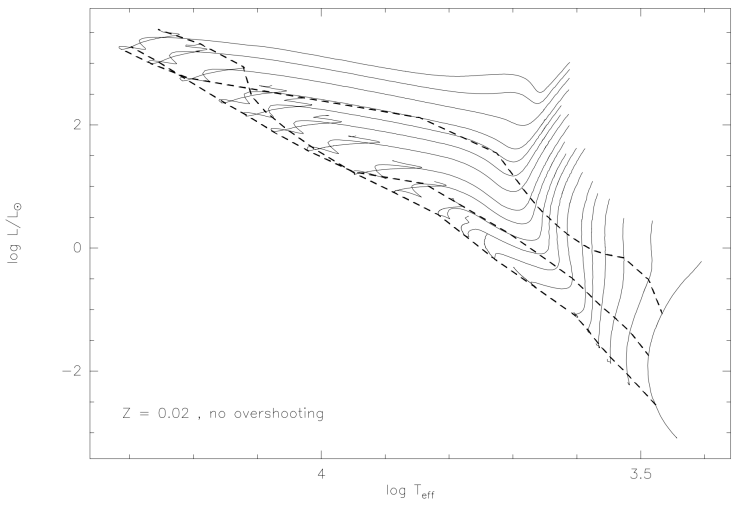

Figure 33: Solid lines show tracks taken by stars of varying masses from $0.1\,M_\odot$ (rightmost line) to $7\,M_\odot$ (leftmost line) in the theoretical HR diagram of luminosity versus effective temperature. Stars begin at the upper right of the tracks and evolve to the lower left; tracks end at the main sequence. Dashed lines represent isochrones corresponding to 10^6, 10^7, and 10^8 years from top right to bottom left. Figure taken from Krumholtz (2015).

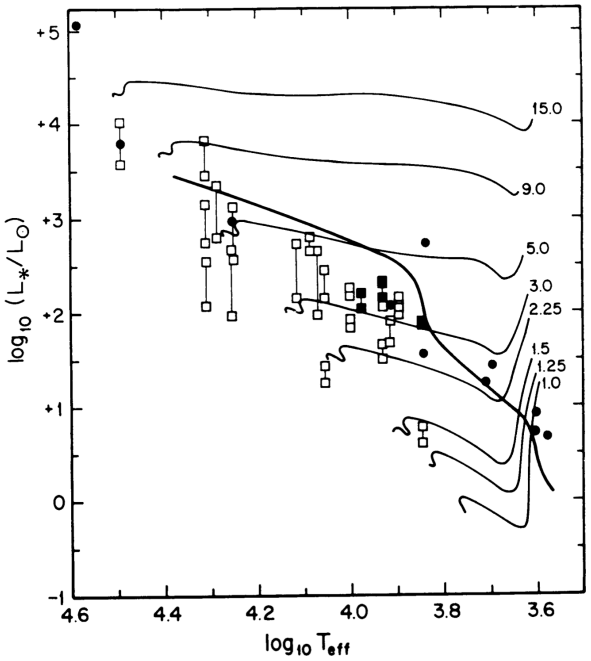

Figure 34: Thin lines show tracks taken by stars of varying masses (indicated by the annotation, in $M_\odot$) in the theoretical HR diagram of luminosity versus effective temperature. Stars begin at the upper right of the tracks and evolve to the lower left; tracks end at the main sequence. The thick line crossing the tracks is the birthline, the point at which the stars stop accreting and become optically visible. Squares and circles represent the properties of observed young stars. Figure adapted from from Palla & Stahler (1990). Figure taken from Krumholtz (2015).

opacity lets radiation rather than convection carry the energy outward.
In the HR diagram, the contraction and increase in $T_{\rm eff}$ produces a vaguely horizontal evolutionary track. This is called the **Heyney track**. The star continues to contract until its center becomes warm enough to allow H burning to begin. At that point it may contract a small additional amount, but the star is essentially on the main sequence. The total time required depends on the stellar mass, but it ranges from several hundred Myr for $0.1\,M_\odot$ stars to essentially zero time for very massive stars, which reach the main sequence while still accreting.

1.13.3 Follow-up Questions

- How do you calculate the Jeans mass?
- What happens to the temperature during adiabatic contraction?
- Draw a plot of density versus temperature to distinguish between the contracting and collapsing phases.

1.14 Question 14

Sketch the rotation curve for a typical spiral galaxy. Show that a flat rotation curve implies the existence of a dark matter halo with a density profile that drops off as $1/r^2$.

1.14.1 Short answer

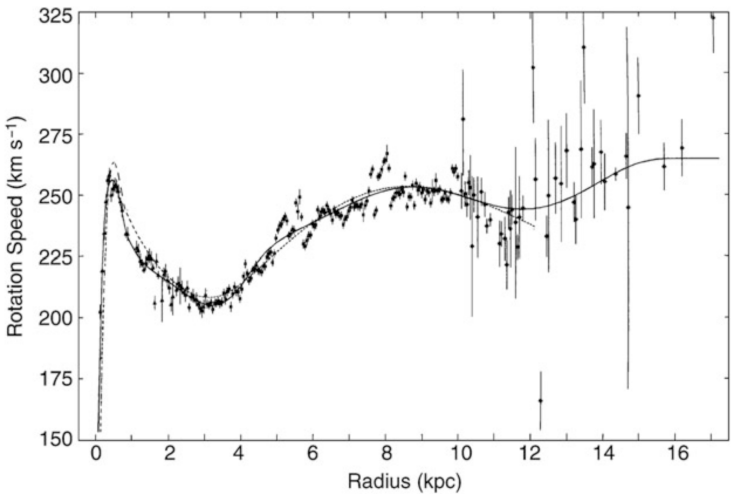

Figure 35: Rotation curve of the Milky Way. Inside the "Solar circle", that is at $R < R_0$, the radial velocity is determined quite accurately using the tangent point method; the measurements outside have larger uncertainties. Source: D. Clemens 1985, Massachusetts-Stony Brook Galactic plane CO survey - The Galactic disk rotation curve ApJ 295, 422, p.429, Fig. 3. Figure taken from Schneider (2015).

1.14.2 Additional context

From observations of the velocity of stars or gas around the Galactic center, the rotational velocity V can be determined as a function of the distance R from the Galactic center.

Decomposition of rotational velocity: We consider an object at distance R from the Galactic center which moves along a circular orbit in the Galactic plane, has a distance D from the Sun, and is located at a Galactic longitude ℓ (see Figure 36). In a Cartesian coordinate system with the Galactic center at the origin, the positional and velocity vectors (we only consider the two components in the Galactic plane because we assume a motion in the plane) are given by

$$\mathbf{r} = R\begin{pmatrix}\sin\theta\\ \cos\theta\end{pmatrix}\ [\mathrm{kpc}], \quad \mathbf{V} = \dot{\mathbf{r}} = V(R)\begin{pmatrix}\cos\theta\\ -\sin\theta\end{pmatrix}\ [\mathrm{km\,s^{-1}}],$$

where θ denotes the angle between the Sun and the object as seen from the Galactic center. From the geometry shown in Figure 36 it follows that

$$\mathbf{r} = \begin{pmatrix}D\sin\ell\\ R_0 - D\cos\ell\end{pmatrix}\ [\mathrm{kpc}].$$

If we now identify the two expressions for the components of $\mathbf{r}$, we obtain

$$\sin\theta = \left(\frac{D}{R}\right)\sin\ell\ [\text{dimensionless}], \quad \cos\theta = \left(\frac{R_0}{R}\right) - \left(\frac{D}{R}\right)\cos\ell\ [\text{dimensionless}].$$

If we disregard the difference between the velocities of the Sun and the LSR we get $\mathbf{V}_\odot \approx \mathbf{V} = (V_0, 0)$ in this coordinate system. Thus the relative velocity between the object and the Sun is, in Cartesian coordinates,

$$\Delta\mathbf{V} = \mathbf{V} - \mathbf{V}_\odot = \begin{pmatrix}V\left(\frac{R_0}{R}\right) - V\left(\frac{D}{R}\right)\cos\ell - V_0\\ -V\left(\frac{D}{R}\right)\sin\ell\end{pmatrix}\ [\mathrm{km\,s^{-1}}].$$

With the angular velocity defined as

$$\Omega \equiv \frac{V(R)}{R}\ [\mathrm{rad\,s^{-1}}]$$

we obtain for the relative velocity

$$\Delta\mathbf{V} = \begin{pmatrix}R(\Omega - \Omega_0) - \Omega D\cos\ell\\ -D\Omega\sin\ell\end{pmatrix}\ [\mathrm{km\,s^{-1}}],$$

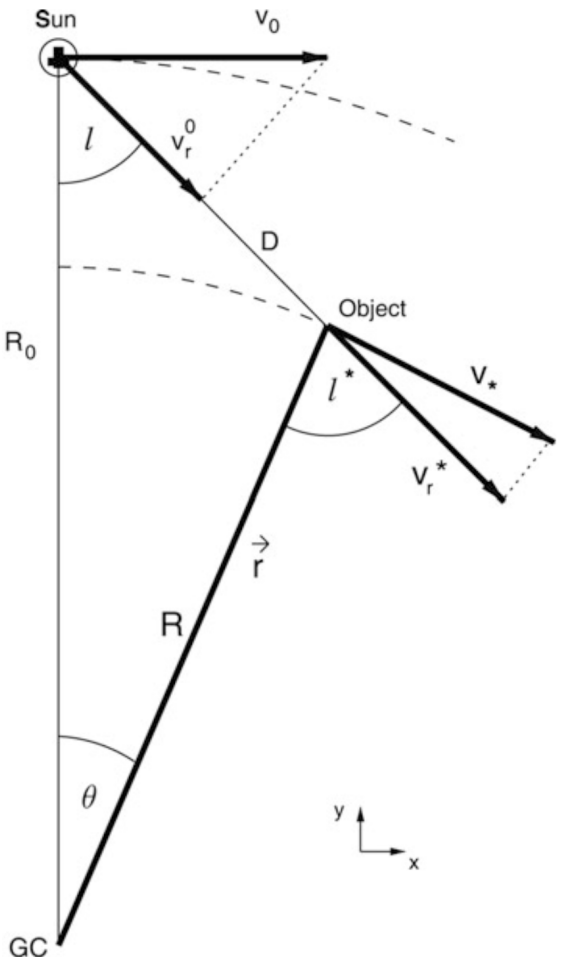

Figure 36: Geometric derivation of the formalism of differential rotation: $v_r = v_r^* - v_r^\odot = v_* \sin\ell^* - v_\odot \sin\ell$, $v_t = v_t^* - v_t^\odot = v_* \cos\ell^* - v_\odot \cos\ell$. One has: $\frac{\sin\ell}{R} = \frac{\sin(\pi-\ell^*)}{R_0} = \frac{\sin\ell^*}{R_0}$, $R\cos\ell^* + D = R_0\cos\ell$, which implies $v_r = R_0\left(\frac{v_*}{R} - \frac{v_\odot}{R_0}\right)\sin\ell$ $= (\Omega - \Omega_0)R_0 \sin\ell$, $v_t = R_0\left(\frac{v_*}{R} - \frac{v_\odot}{R_0}\right)\cos\ell - D\frac{v_*}{R}$ $= (\Omega - \Omega_0)R_0\cos\ell - \Omega D$. Figure taken from Schneider (2015).

where $\Omega_0 = V_0/R_0$ is the angular velocity of the Sun. The radial and tangential velocities of this relative motion then follow by projection of $\Delta\mathbf{V}$ along the direction parallel or perpendicular, respectively, to the separation vector,

$$v_r = \Delta\mathbf{V}\cdot\begin{pmatrix}\sin\ell\\-\cos\ell\end{pmatrix} = (\Omega-\Omega_0)R_0\sin\ell\ [\mathrm{km\,s^{-1}}],$$
$$v_t = \Delta\mathbf{V}\cdot\begin{pmatrix}\cos\ell\\\sin\ell\end{pmatrix} = (\Omega-\Omega_0)R_0\cos\ell - \Omega D\ [\mathrm{km\,s^{-1}}].$$

A purely geometric derivation of these relations is given in Figure 36.

Rotation curve near R_0, Oort constants: One can derive the angular velocity by means of measuring v_r from the equation above, but not the radius R to which it corresponds. Therefore, by measuring the radial velocity alone, $\Omega(R)$ cannot be determined. If one measures v_r and, in addition, the proper motion $\mu = v_t/D$ of stars, then Ω and D can be determined from the equations above, and from D and ℓ one obtains $R = \sqrt{R_0^2 + D^2 - 2R_0 D\cos\ell}$. The effects of extinction prohibits the use of this method for large distances D, since we have considered objects in the Galactic disk. For small distances $D \ll R_0$, which implies $|R - R_0| \ll R_0$, we can make a local approximation by evaluating the expressions above only up to first order in $(R - R_0)/R_0$. In this linear approximation we get

$$\Omega - \Omega_0 \approx \left(\frac{\mathrm{d}\Omega}{\mathrm{d}R}\right)_{R_0}(R - R_0)\ [\mathrm{km\,s^{-1}\,kpc^{-1}}],$$

where the derivative has to be evaluated at $R = R_0$. Hence

$$v_r = (R - R_0)\left(\frac{\mathrm{d}\Omega}{\mathrm{d}R}\right)_{R_0} R_0\sin\ell\ [\mathrm{km\,s^{-1}}],$$

and furthermore, with the definition of angular velocity $\Omega(R)$,

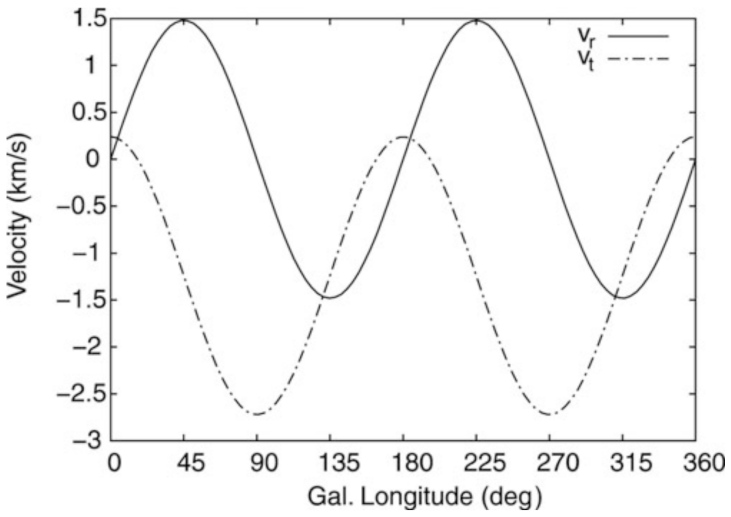

Figure 37: The radial velocity v_r of stars at a fixed distance D is proportional to $\sin 2\ell$; the tangential velocity v_t is a linear function of $\cos 2\ell$. From the amplitude of the oscillating curves and from the mean value of v_t the Oort constants A and B can be derived. Figure taken from Schneider (2015).

$$R_0 \left(\frac{\mathrm{d}\Omega}{\mathrm{d}R}\right)_{R_0} = \frac{R_0}{R}\left[\left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{R_0} - \frac{V}{R}\right] \approx \left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{R_0} - \frac{V_0}{R_0}\ [\mathrm{km\,s^{-1}\,kpc^{-1}}],$$

in zeroth order in $(R - R_0)/R_0$. Combining the last two equations yield

$$v_r = \left[\left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{R_0} - \frac{V_0}{R_0}\right](R - R_0)\sin\ell\ [\mathrm{km\,s^{-1}}],$$

in analogy to this, we obtain for the tangential velocity

$$v_t = \left[\left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{R_0} - \frac{V_0}{R_0}\right](R - R_0)\cos\ell - \Omega_0 D\ [\mathrm{km\,s^{-1}}].$$

For $|R - R_0| \ll R_0$ it follows that $(R_0 - R) \approx D\cos\ell$; if we insert this into v_r and v_t we get

$$v_r \approx AD\sin 2\ell\ [\mathrm{km\,s^{-1}}], \quad v_t \approx AD\cos 2\ell + BD\ [\mathrm{km\,s^{-1}}],$$

where A and B are the **Oort constants**:

$$A \equiv -\frac{1}{2}\left[\left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{R_0} - \frac{V_0}{R_0}\right]\ [\mathrm{km\,s^{-1}\,kpc^{-1}}],$$

$$B \equiv -\frac{1}{2}\left[\left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{R_0} + \frac{V_0}{R_0}\right]\ [\mathrm{km\,s^{-1}\,kpc^{-1}}].$$

The radial and tangential velocity fields relative to the Sun show a sine curve with period π, where ${}_v t$ and v_r are phase-shifted by $\pi/4$. This behavior of the velocity field in the Solar neighborhood is indeed observed (see Figure 37). By fitting the data for $v_r(\ell)$ and $v_t(\ell)$ for stars of equal distance D one can determine A and B, and thus

$$\Omega_0 = \frac{V_0}{R_0}\ [\mathrm{rad\,s^{-1}}] = A - B\ [\mathrm{km\,s^{-1}\,kpc^{-1}}], \quad \left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{R_0} = -(A + B)\ [\mathrm{km\,s^{-1}\,kpc^{-1}}].$$

The Oort constants thus yield the angular velocity of the Solar orbit and its derivative, and therefore the local kinematical information. If our Galaxy was rotating rigidly so that Ω was independent of the radius, $A = 0$ would follow. But the Milky Way rotates differentially (i.e., the angular velocity depends on the radius). Measurements yield the following values for A and B,

$$A = (14.8 \pm 0.8)\ [\mathrm{km\,s^{-1}\,kpc^{-1}}]$$
$$B = (-12.4 \pm 0.6)\ [\mathrm{km\,s^{-1}\,kpc^{-1}}].$$

Galactic rotation curve for $R < R_0$; tangent point method: To measure the rotation curve for radii that are significantly smaller than R_0, one has to turn to large wavelengths due to extinction in the disk. Usually the 21 cm emission line of neutral hydrogen is used, which can be observed over large distances, or the emission of CO in molecular gas. These gas components are found throughout the disk and are strongly concentrated towards the plane. Furthermore, the radial velocity can easily be measured

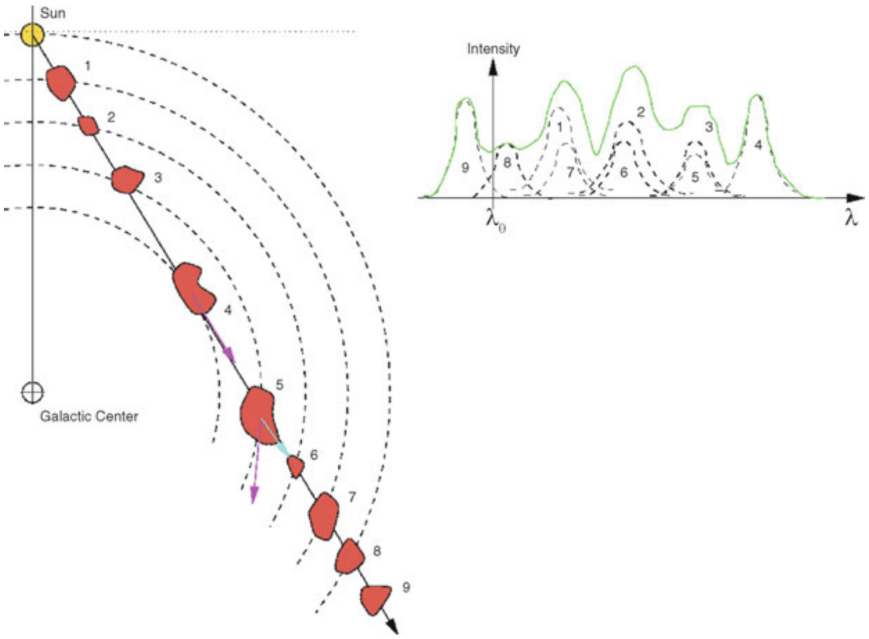

Figure 38: The ISM is optically thin for 21 cm radiation, and thus we receive the 21 cm emission of HI regions from everywhere in the Galaxy. Due to the motion of an HI cloud relative to us, the wavelength is shifted. This can be used to measure the radial velocity of the cloud. With the assumption that the gas is moving on a circular orbit around the Galactic center, one expects that for the cloud in the tangent point (cloud 4), the full velocity is projected along the line-of-sight so that this cloud will therefore have the largest radial velocity. If the distance of the Sun to the Galactic center is known, the velocity of a cloud and its distance from the Galactic center can then be determined. Adopted from B.W. Carroll & D.A. Ostlie 1996, Introduction to Modern Astrophysics, Addison-Wesley. Figure taken from Schneider (2015).

from the Doppler effect. However, since the distance to a hydrogen cloud cannot be determined directly, a method is needed to link the measured radial velocities to the distance of the gas from the Galactic center. For this purpose the tangent point method is used.
Consider a line-of-sight at fixed Galactic longitude ℓ, with $\cos\ell > 0$ (thus 'inwards'). The radial velocity v_r along this line-of-sight for objects moving on circular orbits is a function of the distance D, as found previously using the Oort constants. If $\Omega(R)$ is a monotonically decreasing function, v_r attains a maximum where the line-of-sight is tangent to the local orbit, and thus its distance R from the Galactic center attains the minimum value $R_{\min}$. This is the case at

$$D = R_0 \cos\ell \text{ [kpc]}, \quad R_{\min} = R_0 \sin\ell \text{ [kpc]}$$

(see Figure 38). The maximum radial velocity there, to equations derived above with the Oort constants, is

$$v_{r,\max} = [\Omega(R_{\min} - \Omega_0)]R_0 \sin\ell = V(R_{\min}) - V_0 \sin\ell \text{ [km s}^{-1}\text{]},$$

so that from the measured value of $v_{r,\max}$ as a function of direction ℓ, the rotation curve inside R_0 can be determined,

$$V(R) = \left(\frac{R}{R_0}\right) V_0 + v_{r,\max}(\sin\ell = R/R_0) \text{ [km s}^{-1}\text{]}.$$

In the optical regime of the spectrum this method can only be applied locally, i.e., for small D, due to extinction. This is the case if one observes in a direction nearly tangential to the orbit of the Sun, i.e., if $0 < \pi/2 - \ell \ll 1$ or $0 < \ell - 3\pi/2 \ll 1$, or $|\sin\ell| \approx 1$, so that $R_0 - R_{\min} \ll R_0$. In this case we get, to first order in $(R_0 - R_{\min})$,

$$V(R_{\min}) \approx V_0 \left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{R_0} (R_{\min} - R_0) \tag{1}$$

$$= V_0 - \left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{R_0} R_0(1 - \sin\ell) \text{ [km s}^{-1}\text{]}, \tag{2}$$

Applying this to our equation for $v_{r,\max}$,

$$v_{r,\max} = \left[V_0 - \left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{R_0} R_0\right](1 - \sin\ell)$$
$$= 2AR_0(1 - \sin\ell)\ [\mathrm{km\,s^{-1}}],$$

This relation can also be used for determining the Oort constant A.
To determine $V(R)$ for smaller R by employing the tangent point method, we have to observe in wavelength regimes in which the Galactic plane is transparent, using radio emission lines of gas. In Figure 39, a typical intensity profile of the 21 cm line along a line-of-sight is sketched; according to the Doppler effect this can be converted directly into a velocity profile using $v_r = (\lambda - \lambda_0)\lambda_0$. It consists of several maxima that originate in individual gas clouds. The radial velocity of each cloud is defined by its distance R from the Galactic center (if the gas follows the Galactic rotation), so that the largest radial velocity will occur for gas closest to the tangent point, which will be identified with $v_{r,\max}(\ell)$. Figure 2.28 shows the observed intensity profile of the ^{12}CO line as a function of the Galactic longitude, from which the rotation curve for $R < R_0$ can be read off.

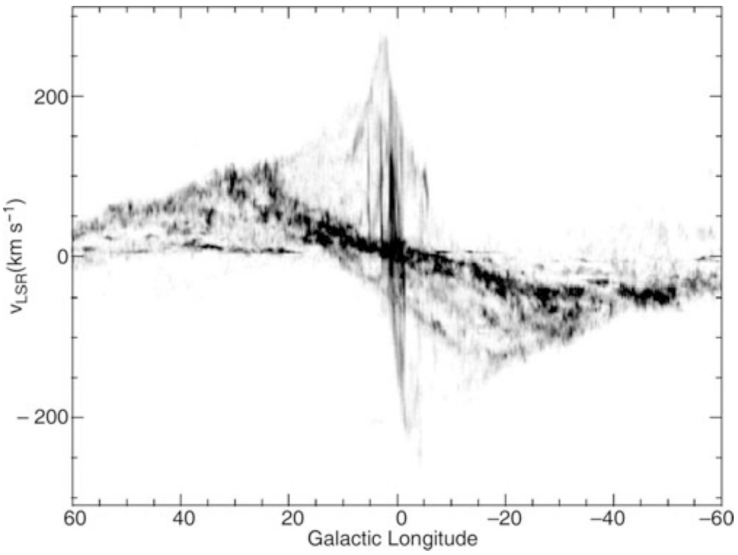

Figure 39: ^{12}CO emission of molecular gas in the Galactic disk. For each ℓ, the intensity of the emission in the $\ell - v_r$ plane is plotted, integrated over the range $-2° \leq b \leq 2°$ (i.e., very close to the middle of the Galactic plane). Since v_r depends on the distance along each line- of-sight, characterized by ℓ, this diagram contains information on the rotation curve of the Galaxy as well as on the spatial distribution of the gas. The maximum velocity at each ℓ is rather well defined and forms the basis for the tangent point method. Source: P. Englmaier & O. Gerhard 1999, Gas dynamics and large-scale morphology of the Milky Way galaxy, MNRAS 304, 512, p. 514, Fig. 1. Figure taken from Schneider (2015).

With the tangent point method, applied to the 21 cm *line of neutral hydrogen or to radio emission lines of molecular gas, the rotation curve of the Galaxy inside the Solar orbit (i.e., for $R < R_0$) can be measured.*
Rotation curve for $R > R_0$: The tangent point method cannot be applied for $R > R_0$ because for lines-of-sight at $\pi/2 < \ell < 3\pi/2$, the radial velocity v_r attains no maximum. In this case, the line-of-sight is nowhere parallel to a circular orbit.
Measuring $V(R)$ for $R > R_0$ requires measuring v_r for objects whose distance can be determined directly (e.g., Cepheids, for which the period-luminosity relation is used, or O- and B-stars in HII-regions). With ℓ and D known, R can then be calculated which allows us to obtain $\Omega(R)$ or $V(R)$, respectively. Any object with known D and v_r thus contributes one data point to the Galactic rotation curve. Since the distance estimates of individual objects are always affected by uncertainties, the rotation curve for large values of R is less accurately known than that inside the Solar circle. Recent measurements of blue horizontal-branch stars within the outer halo of the MW by SDSS yielded an estimate of the rotation curve out to $r \sim 60\,\mathrm{kpc}$. The situation will improve dramatically once the results from Gaia will become available: Gaia will measure distances via trigonometric parallaxes, and proper motions of many star outside the Solar circle.
It turns out that the rotation curve for $R > R_0$ does not decline outwards (see Figure 35) as we would expect from the distribution of visible matter in the MW. Both the stellar density and the gas density of the Galaxy decline exponentially for large R. This steep radial decline of the visible matter density should imply that $M(R)$, the mass inside R, is nearly constant for $R \gtrsim R_0$, from which a velocity profile like $V \propto R^{-1/2}$ would follow, according to Kepler's law. However, this is not the case: $V(R)$ is virtually constant for $R > R_0$, indicating that $M(R) \propto R$. In fact, a small decrease to about $180, \mathrm{km\,s^{-1}}$ at $R = 60\,\mathrm{kpc}$ was estimated, corresponding to a total mass of $(4.0 \pm 0.7) \times 10^{11}\,\mathrm{M}_\odot$ enclosed within the inner 60 kpc, but this decrease is much smaller than expected from Keplerian rotation. In order to get an almost constant rotational velocity of the Galaxy, much more matter has to be present than we observe

in gas and stars.
The MW contains, besides stars and gas, an additional component of matter that dominates the mass at $R \gtrsim R_0$. Its presence is known only by its gravitational effect, since it has not been observed directly yet, neither in emission nor in absorption. Hence, it is called ***dark matter***. This is a common phenomenon: the rotation curves of spiral galaxies are flat up to the maximum radius at which they can be measured; spiral galaxies contain dark matter. A better way of phrasing is would be to say that the visible galaxy is embedded in a dark matter halo, since the total mass of the MW (and other spiral galaxies) is dominated by dark matter.
Rotation curves and dark matter: The rotation curves of other spiral galaxies are easier to measure than that of the MW because we are able to observe them 'from outside'. These measurements are achieved by utilizing the Doppler effect, where the inclination of the disk (i.e., its orientation with respect to the line-of-sight) has to be accounted for. The inclination angle is determined from the observed axis ratio of the disk, assuming that disks are intrinsically axially symmetric (except for the spiral arms). Mainly the stars and HI gas in the galaxies are used as luminous tracers, where the observable HI disk is in general significantly more extended than the stellar disk. Therefore, the rotation curves measured from the 21 cm line typically extend to much larger radii than those from optical stellar spectroscopy.
Like our MW, other spirals also rotate considerably faster in their outer regions than one would expect from Kepler's law and the distribution of visible matter (see Figure 40).

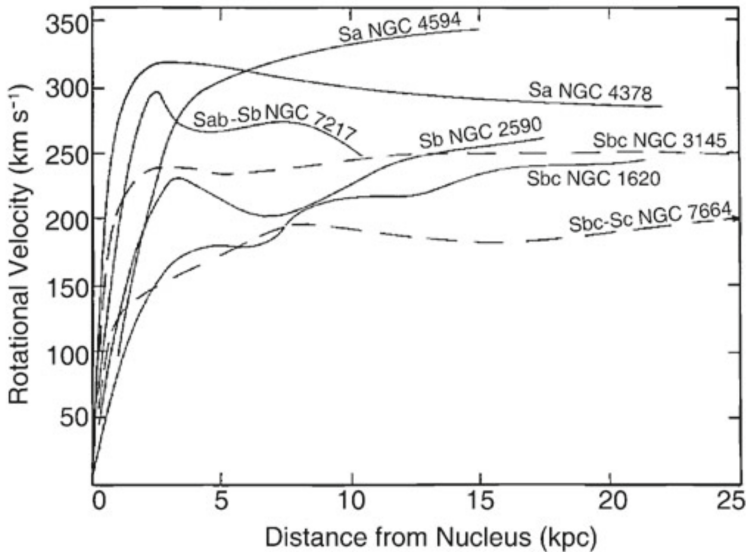

Figure 40: Examples of rotation curves of spiral galaxies. They are all flat in the outer region and do not behave as expected from Kepler's law if the galaxy consisted only of luminous matter. Also striking is the fact that the amplitude of the rotation curve is higher for early-type than for late-type spirals. Source: V. Rubin et al. 1978, Extended rotation curves of high-luminosity spiral galaxies. IVSystematic dynamical properties, SA through SC, ApJ 225, L107, p. L109, Fig. 3. Figure taken from Schneider (2015).

The rotation curves of spirals do not decrease for $R \geq h_R$, as one would expect from the light distribution, but are basically flat. We therefore conclude that spirals are surrounded by a halo of dark matter. The density distribution of this dark halo can be derived from the rotation curves.
To see how the density distribution of the dark matter can be derived from the rotation curves, we employ the force balance between gravitation and centrifugal acceleration, as described by the Kepler rotation law,

$$v^2(R) = \frac{GM(R)}{R} \; [\mathrm{km^2\,s^{-2}}],$$

from which one directly obtains the mass $M(R)$ within a radius R. The rotation curve expected from the visible matter distribution is[1]

$$v^2_{\mathrm{lum}} = \frac{GM_{\mathrm{lum}}(R)}{R} \; [\mathrm{km^2\,s^{-2}}].$$

$M_{\mathrm{lum}}(R)$ can be determined by assuming a plausible value for the mass-to-light ratio $M/L4$ of the luminous matter. This value is obtained either from the spectral light distribution of the stars, together with knowledge of the properties of stellar populations, or by fitting the innermost part of the rotation curve (where the mass contribution of dark matter can presumably be neglected), assuming that M/L is independent of radius for the stellar population. From this estimate of the mass-to-light ratio, the

[1]This consideration is strongly simplified insofar as the given relations are only valid in this form for spherical mass distributions. The rotational velocity produced by an oblate (disk-shaped) mass distribution is more complicated to calculate; for instance, for an exponential mass distribution in a disk, the maximum of v_{lum} occurs at $\sim 2.2h_R$, with a Kepler decrease, $v_{\mathrm{lum}} \propto R^{-1/2}$, at larger radii.

discrepancy between $v^2_{\rm lum}$ and v^2 yields the distribution of the dark matter, $v^2_{\rm dark} = v^2 - v^2_{\rm lum} = GM_{\rm dark}/R$, or

$$M_{\rm dark}(R) = \frac{R}{G}[v^2(R) - v^2_{\rm lum}(R)]\ [\mathrm{M}_\odot].$$

An example of this decomposition of the mass contributions is shown in Figure 41.

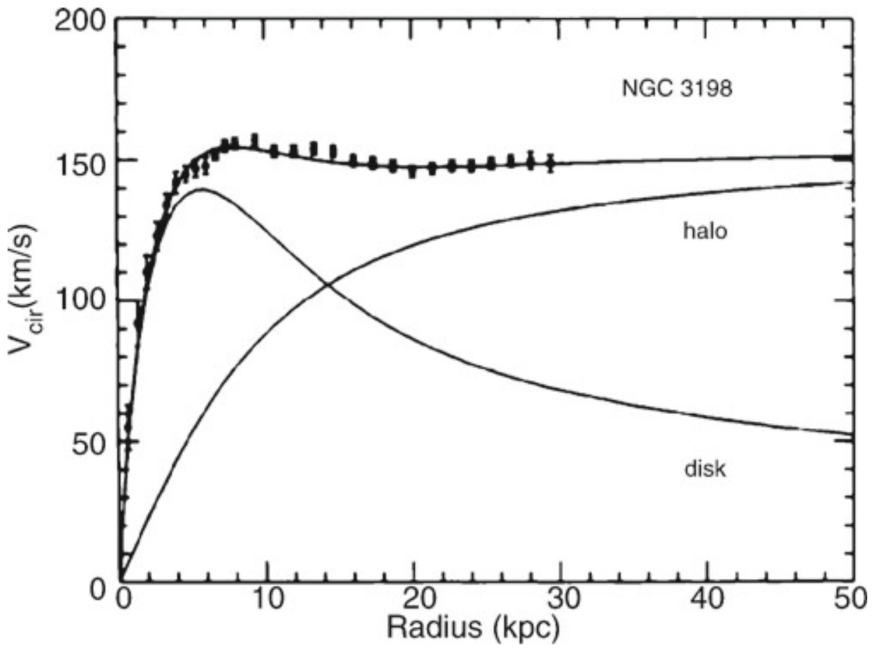

Figure 41: The flat rotation curves of spiral galaxies cannot be explained by visible matter alone. The example of NGC 3198 demonstrates the rotation curve which would be expected from the visible matter alone (curve labeled 'disk'). To explain the observed rotation curve, a dark matter component has to be present (curve labeled 'halo'). However, the decomposition into disk and halo mass is not unambiguous because for it to be so it would be necessary to know the mass-to-light ratio of the disk. In the case considered here, a 'maximum disk' was assumed (i.e., it was assumed that the innermost part of the rotation curve is produced solely by the visible matter in the disk). Source: T.S. van Albada et al. 1985, Distribution of dark matter in the spiral galaxy NGC 3198, ApJ 295, 305, p. 309, Fig. 4. Figure taken from Schneider (2015).

The corresponding density profiles of the dark matter halos seem to be flat in the inner region, and decreasing as R^{-2} at large radii. It is remarkable that $\rho \propto R^{-2}$ implies a mass profile $M \propto R$ (i.e., the mass of the halo increases linearly with the radius for large R). As long as the extent of the halo is undetermined the total mass of a galaxy will be unknown. Since the observed rotation curves are flat out to the largest radius for which 21 cm emission can still be observed, a lower limit for the radius of the dark halo can be obtained, $R_{\rm halo} \gtrsim 30h^{-1}\,\mathrm{kpc}$. Inside the optical radius of a disk, the dark matter comprises about 2/3 of the total mass.
To derive the density profile out to even larger radii, other observable objects in an orbit around the galaxies are needed. Potential candidates for such luminous tracers are satellite galaxies – companions of other spirals, like the Magellanic Clouds are for the MW. Because we cannot presume that these satellite galaxies move on circular orbits around their parent galaxy, conclusions can be drawn based only on a statistical sample of satellites. These analyses of the relative velocities of satellite galaxies around spirals still give no indication of an 'edge' to the halo, leading to a lower limit for the radius of $R_{\rm halo} \gtrsim 100h^{-1}\,\mathrm{kpc}$.

1.14.3 Follow-up Questions

- What assumptions are made in deriving the $1/r^2$ profile?

1.15 Question 15

What thermal phases are postulated to exist in the interstellar medium? Describe the dominant mechanism of cooling for each phase.

1.15.1 Short answer

Answer.

1.15.2 Additional context

Interstellar matter accounts for $\sim 10-15\%$ of the total mass of the Galactic disk. It tends to concentrate near the Galactic plane and along the spiral arms, while being very inhomogeneously distributed at small scales. Roughly half the interstellar mass is confined to discrete clouds occupying only $\sim 1-2\%$ of the interstellar volume. These interstellar clouds can be divided into three types: the **dark clouds**, which are essentially made of very cold ($T \sim 10-20\,\mathrm{K}$) molecular gas and block off the light from background stars, the **diffuse clouds**, which consist of cold ($T \sim 100\,\mathrm{K}$) atomic gas and are almost transparent to the background starlight, except at a number of specific wavelengths where they give rise to absorption lines, and the **translucent clouds**, which contain molecular and atomic gases and have intermediate visual extinctions. The rest of the interstellar matter, spread out between the clouds, exists in three different forms: **warm (mostly neutral) atomic**, **warm ionized**, and **hot ionized**, where warm refers to a temperature of $10^4\,\mathrm{K}$ and hot to a temperature of $10^6\,\mathrm{K}$ (see Table 3).

Component	T (K)	n (cm^{-3})	$\Sigma_\odot$ ($M_\odot\,\mathrm{pc}^{-2}$)	$\mathcal{M}$ ($10^9 M_\odot$)
Molecular	10–20	10^2–10^6	~2.5	~1.3[a]–2.5[b]
Cold atomic	50–100	20–50	~3.5	} ≳6.0
Warm atomic	6000–10 000	0.2–0.5	~3.5	
Warm ionized	~8000	0.2–0.5	~1.4	≳1.6
Hot ionized	$\sim 10^6$	~0.0065		

[a]Adapted from Bronfman *et al.*, 1988.
[b]Adapted from Clemens *et al.*, 1988.

Table 3: Descriptive parameters of the different components of the interstellar gas. T is the temperature, n is the true (as opposed to space-averaged) number density of hydrogen nuclei near the Sun, $\Sigma_\odot$ is the azimuthally averaged mass density per unit area at the solar circle, and $\mathcal{M}$ is the mass contained in the entire Milky Way. Both $\Sigma_\odot$ and $\mathcal{M}$ include 70.4% hydrogen, 28.1% helium, and 1.5% heavier elements. All values were rescaled to $R_\odot = 8.5\,\mathrm{kpc}$. Figure taken from Ferrière (2001).

Molecular gas: The H_2 molecule itself is not directly observable at radio wavelengths: because it possesses no permanent electric dipole moment and has a very small moment of inertia, all its permitted transitions lie outside the radio domain. The CO molecule, for its part, has a $J = 1 \to 0$ rotational transition at a radio wavelength of 2.6 mm; the corresponding emission line, which was first observed a few months before the detection of CO in UV absorption, has become the primary tracer of molecular interstellar gas.
Neutral atomic gas: Neutral atomic hydrogen, usually denoted by HI (as opposed to HII for ionized hydrogen), is not directly observable at optical wavelengths. Under most interstellar conditions, particle collisions are so infrequent that nearly all hydrogen atoms have their electron in the ground energy level $n = 1$. It turns out that all the electronic transitions between the ground level and an excited state (forming the Lyman series) lie in the UV, with the Lyman α (Lyα) transition between the ground level and the first excited state $n = 2$ occurring at a wavelength of 1216Å.
The breakthrough event that opened the era of radio-astronomical observations of interstellar HI was the detection of the interstellar 21 cm line emission. The existence of the 21 cm line results from the "hyperfine" structure of the hydrogen atom. In brief, the interaction between the magnetic moment of the electron and that of the proton leads to a splitting of the electronic ground level into two extremely close energy levels, in which the electron spin is either parallel (upper level) or anti-parallel (lower level) to the proton spin. It is the "spin-flip" transition between these two energy levels that corresponds to the now famous 21 cm line. The major advantage of 21 cm photons resides in their ability to penetrate deep into the ISM, thereby offering a unique opportunity to probe the interstellar HI gas out to the edges of the MW. On the other hand, the highly forbidden spin-flip transition is intrinsically so rare (Einstein A coefficient $A_{21} = 2.85 \times 10^{-15}\,\mathrm{s}^{-1}$) that very long paths are needed for the 21 cm line to be detectable.
The 21 cm absorption spectra generally look quite different from emission spectra taken in a nearby direction: while the emission spectra contain both distinct narrow peaks and much broader features, only

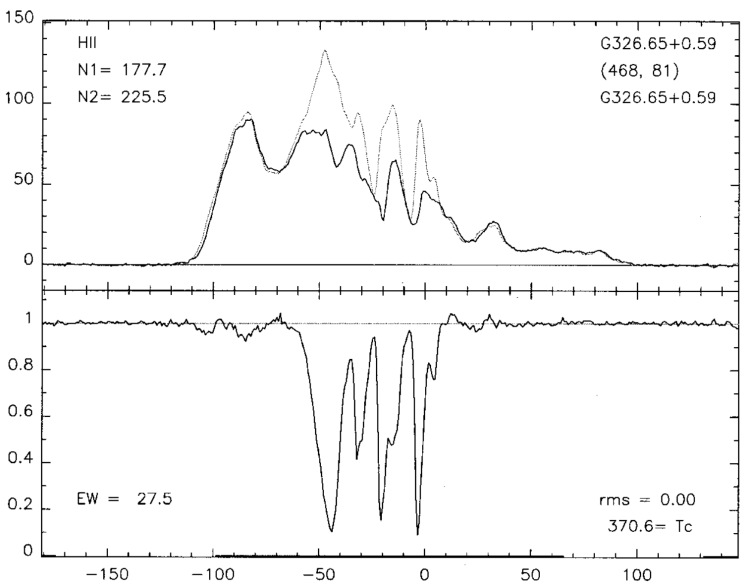

Figure 42: 21 cm spectra of HI gas. The x axis is gas velocity along the line of sight (in $\mathrm{km\,s^{-1}}$) and the y axis is intensity in arbitrary units. **(Top:)** Emission line profile nearby to the direction toward an HII region. **(Bottom:)** The absorption spectrum, which can represent only relatively cool gas exactly along the line of sight to the HII region. Figure taken from Ferrière (2001).

the narrow peaks are present in the absorption spectra (see Figure 42). The conventional interpretation of this difference is that the narrow peaks seen in emission and in absorption are produced by discrete cold ($T \sim 5 - 100\,\mathrm{K}$) H I clouds, whereas the broader features seen in emission only are due to a widespread HI gas that is too warm to give rise to detectable 21 cm absorption. The estimated temperature of the warm HI component is $T \sim 6000 - 10000\,\mathrm{K}$.

Warm ionized gas: O and B stars, the hottest and most massive stars in the MW, emit a strong UV radiation, which, below a wavelength of 912Å (corresponding to an energy of 13.6 eV), is sufficiently energetic to ionize hydrogen atoms. As a result, these stars are surrounded by a so-called "**HII region**" within which hydrogen is almost fully ionized. Given that the ionizing UV photons are promptly absorbed by neutral hydrogen, the transition between the HII region and the ambient ISM is rather abrupt. Inside the HII region, ions and free electrons keep recombining before being separated again by fresh UV photons from the central star. Thus the HII region grows until the rate of recombinations within it becomes large enough to balance the rate of photoionizations. In a uniform medium, this balance occurs when the radius of the HII region reaches the value of the **Strömgren radius**,

$$r_s = 30 \left(\frac{N_{48}}{n_{\mathrm{H}} n_e} \right)^{1/3} \ [\mathrm{pc}],$$

where N_{48} is the number of ionizing photons emitted per unit time by the central star, in $10^{48}\,\mathrm{s^{-11}}$ (e.g., $N_{48} \approx 34$ for an O5V star and $N_{48} \approx 1.7$ for a B0V star), and n_{H} and n_e are the free-proton and free-electron number densities in the HII region, in $\mathrm{cm^{-3}}$.

The process of photoionization is accompanied by a net heating of the interstellar gas, as the ionizing photons transfer a fraction of their energy (the excess with respect to the ionization potential) to the ejected electrons. The equilibrium temperature, set by a balance between photoelectric heating and radiative cooling, has a typical value $\sim 8000\,\mathrm{K}$, depending on density and metallicity. This theoretical estimate turns out to be in good agreement with observational determinations based on measurements of the radio continuum radiation and on studies of emission-line ratios from HII regions.

The radio continuum radiation of an HII region arises from the **bremsstrahlung or** "**free-free**" emission generated as free electrons are accelerated in the Coulomb field of positive ions ($\mathrm{H^+}$,$\mathrm{He^+}$,$\mathrm{He^{2+}}$). Emission lines, found at optical, IR, and radio wavelengths, are primarily due to radiative recombination of hydrogen and helium ions with free electrons and to radiative de-excitation of collisionally excited ionized metals. Of special importance are the optical hydrogen **Balmer lines** produced by electronic transitions from an excited state $n > 2$ to the first excited state $n = 2$. Because each recombination of a free proton with a free electron into an excited hydrogen atom leads sooner or later to the emission of one Balmer photon, and because the rate per unit volume of recombinations into an excited hydrogen atom is proportional to $N_{\mathrm{H^+}} + n_e \propto n_e^2$, the integrated intensity of the Balmer lines is directly proportional to the **emission measure**

$$\mathrm{EM} = -\int_{\text{source}}^{\text{observer}} n_e^2 \mathrm{d}s \ [\mathrm{pc\,cm^{-6}}],$$

where ds is the length element along the line of sight through the HII region. For future reference, let us specify that the hydrogen Balmer transition between the electronic energy levels $n = 3$ and $n = 2$ is usually referred to as the **H**α transition and has a wavelength of 6563 Å.

Hot ionized gas: The notion that hot interstellar gas exists in the MW dates back to Spitzer's (1956) paper on a possible Galactic corona, made of hot rarefied gas, which would provide the necessary pressure to confine the observed high-altitude interstellar clouds. The presence of such a hot gas was borne out almost two decades later by two independent types of observations: (1) the Copernicus satellite detected, in the spectrum of several bright stars, broad UV absorption lines of high-stage ions that form only at elevated temperatures, and (2) the observed soft-x-ray background radiation was found to be most likely due to thermal emission from a hot interstellar plasma.

Amongst the high-stage ions accessible to UV observations, OVI (five-times-ionized oxygen, with a doublet at (1032Å, 1038Å)) and NV (four-times-ionized nitrogen, with a doublet at (1239Å, 1243Å)) are the best tracers of hot collisionally ionized gas, insofar as their high ionization potential makes them difficult to produce by photoionization.

The soft-x-ray background radiation around 0.25 keV appears to arise predominantly from the Local Bubble. It is very likely that 0.25 keV x-ray-emitting regions exist throughout the MW, but because their radiation is efficiently absorbed by the intervening cool interstellar gas, the majority of them must escape detection. On the other hand, a number of bright features have been observed in the intermediate energy band $0.5 - 1.0$ keV, which is less affected by photoelectric absorption. Most of these features were shown to be associated either with individual supernova remnants (produced by isolated supernova explosions) or with "**superbubbles**" (produced by the joint action of stellar winds and supernova explosions in a group of massive stars), and their x-ray radiation was attributed to thermal emission from a hot plasma at a temperature of a few 10^6 K.

It is now widely accepted that the hot interstellar gas is generated by SN explosions and, to a lesser extent, by the generally powerful winds from the progenitor stars.

Figures 44, 45, and 46 show the various components of the ISM within the Galactic disk, about 300 pc of the midplane, and larger scale structure of the Galactic atmosphere, respectively. In these figures, purple indicates dark molecular clouds; solid green-cold HI clouds; hatched greenwarm HI; hatched green on yellow background-diffuse warm HII; orange-hotter gas bearing OVI; red-material hot enough to emit X rays; and (in Figure 44 only) black lines with arrowheads-the magnetic field.

Effects of Supernovae on the ISM: The dynamical state of the ISM in the MW and other galaxies is strongly affected by SN explosions. The light emitted by them is spectacular, but it is the high-velocity ejecta that have the dominant effect on the ISM. Depending on the SN type, the ejecta mass can range from $\sim 1.4\,M_\odot$ (a Type Ia SNe, produced by explosion of a white dwarf near the Chandrasekhar limit) to perhaps $\sim 10 - 20\,M_\odot$ for Type II SNe following core collapse in massive stars. The ejecta will have a range of velocities, with the outermost material moving the fastest.

This velocity is far greater than the sound speed in the surrounding material, and the expanding ejecta will therefore drive a fast shock into the circumstellar medium. We will refer to all of the matter interior to this shock surface as the **supernova remnant**, or SNR. In the first days after the explosion, the density of the ejecta far exceeds the density of the circumstellar medium, and the ejecta continue to expand ballistically at nearly constant velocity – this is referred to as the **free expansion phase**. At these early times, there is only one shock of interest – the shock wave propagating outward into the ambient medium.

As the density of the expanding ejecta drops (as t^{-3}), the pressure of the shocked circumstellar medium soon exceeds the thermal pressure in the ejecta, and a reverse shock is driven into the ejecta. The remnant now contains two shock fronts: the original outward-propagating shock (the **blastwave**) expanding into the circumstellar/interstellar medium, and the **reverse shock** propagating inward, slowing and shock-heating the ejecta (which had previously been cooled by adiabatic expansion). The reverse shock becomes important when the expanding ejecta material has swept up a mass of circumstellar or interstellar matter comparable to the ejecta mass.

Once the reverse shock has reached the center of the remnant, all of the ejecta are very hot, and the free-expansion phase is over. The pressure in the supernova remnant is far higher than the pressure in the surrounding medium. The hot gas has been emitting radiation, but if the densities are low, the radiative losses at early times are negligible.

The SNR now enters a phase that can be approximated by idealizing the problem as a **point explosion** injecting energy into a uniform-density zero-temperature medium: we neglect (1) the finite mass of the ejecta, (2) radiative losses; and (3) the pressure in the ambient medium. The hot gas interior to the shock front is, of course, radiating energy. When the radiative losses become important the blastwave will enter a **radiative** phase, where the gas in the shell just interior to the shock front is now able to cool to temperatures much lower than the temperature at the shock front.

At this point, cooling causes the thermal pressure just behind the shock to drop suddenly, and the shock wave briefly stalls. However, the very hot gas in the interior of the SNR has not cooled, and its outward pressure forces the SNR to continue its expansion. The blastwave now enters what is called the **snowplow phase**, with a dense shell of cool gas enclosing a hot central volume where radiative cooling is unimportant. This is called the snowplow phase because the mass of the dense shell increases as it "sweeps up" the ambient gas.

For typical interstellar parameters, the shock speed at the beginning of the snowplow phase is $\sim 150\,\mathrm{km\,s^{-1}}$, which results in a very strong shock when propagating through interstellar gas with $T \lesssim 10^4\,\mathrm{K}$. However, the shock front gradually slows, and the shock compression declines. This proceeds until the shock speed approaches the effective sound speed in the gas through which the blastwave is propagating, at which point the compression ratio approaches unity, and the shock wave turns into a sound wave.

SNe blastwaves propagates more rapidly in a low density medium. The lowest density phase in the ISM is the HIM, with temperatures $T \approx 10^6\,\mathrm{K}$, a substantial volume filling factor, and a density HIM of $(n_\mathrm{H})_\mathrm{HIM} \approx 0.005\,\mathrm{cm^{-3}}$. A typical Sn blastwave will be expanding into such gas, with the blastwave passing around any high-density clouds that may be present. If the clouds were rigid, and had a small filling factor, they would have little effect on the propagation of the blastwave, but of course real are compressible, and can be heated and "evaporated" by contact with hot gas. The shock passing through the cloud will set the cloud material into motion, but with velocity gradients in the shocked material. If the cloud is not self-gravitating, these velocity gradients can act to "shred" the cloud. However, the magnetic field that is almost certainly already present in the cloud may be able to oppose these shearing effects.

After the blastwave has passed over the cloud, the cloud finds itself engulfed in hot gas resulting from the shock-heating of the low density inter-cloud medium. Thermal conduction will transport heat from the hot plasma into the cool cloud. If the cloud is sufficiently small, this thermal conduction can lead to **evaporation** of cloud material, resulting in mass loss from the cloud, and an increase in the mass density of the shocked inter-cloud medium. The increase in the mass density will be accompanied by a drop in the temperature, as the thermal energy is shared by more particles. The combined effects of increased density and lowered temperature act to promote radiative cooling.

Three-phase model of the ISM: An initially uniform ISM consisting of warm HI would be transformed by SNRs into a medium consisting of low-density hot gas and dense shells of cold gas. This transformation would take place in just a few Myr. McKee & Ostriker (1977) developed a model of the ISM that took into account the effects of these blastwaves. They envisaged an ISM consisting of three distinct phases: cold gas: the **cold neutral medium** (CNM); warm gas: the **warm neutral medium** (WNM) and **warm ionized medium** (WIM); and hot gas: the **hot ionized medium** (HIM). A SNR blastwave expands into this composite medium, as illustrated in Figure 43.

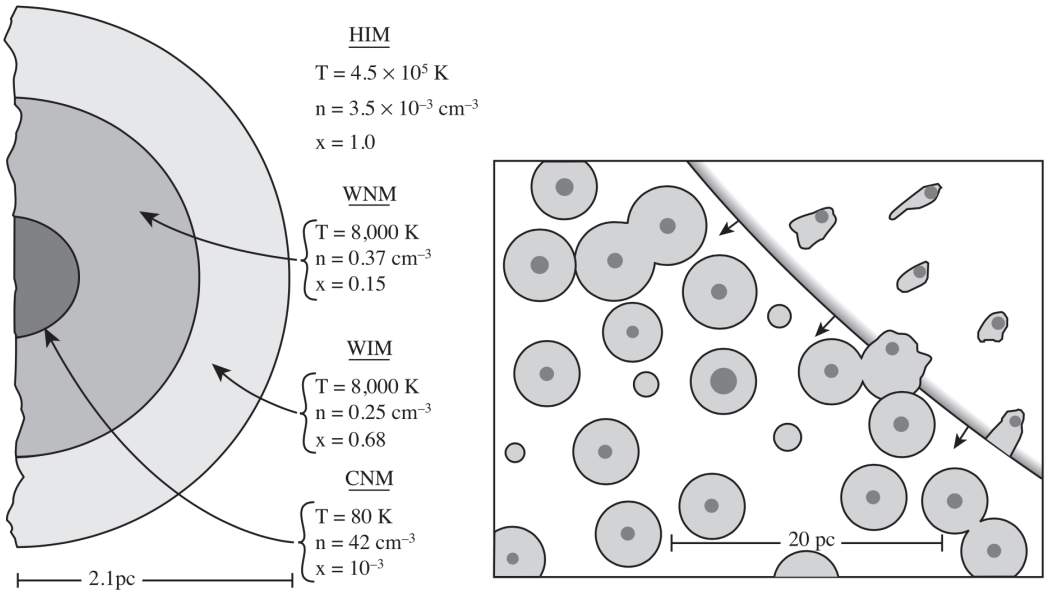

Figure 43: **Left:** Structure of a typical cold cloud in the three-phase model of McKee & Ostriker (1977). **Right:** Close-up of a supernova blastwave. From McKee & Ostriker (1977). Figure taken from Draine (2011).

McKee & Ostriker (1977) argued that the pressure in the ISM was maintained by SNe – if initially the ISM had a low pressure, then SNRs would expand to large radii before "fading," with resulting overlap.

The pressure in the ISM will rise until the SNRs tend to overlap just as they are fading, at which point the pressure in the ISM is the same as the pressure in the SNR. According to this argument, SNRs can be used to predict the pressure in the ISM. This is a remarkable result: given (1) the observed SN rate/volume; (2) the observed kinetic energy per supernova; and (3) the atomic physics of the cooling function (using observed abundances) – from these alone they were able to predict the interstellar pressure!
This model envisaged three phases of the ISM: CMM, WNM, WIM, and HIM. They did not explicitly consider molecular gas, because it occupied a very small volume filling factor, and can be considered part of the CNM. Our current view of the ISM continues to identify these as major phases, and follows the central ideas of the this model:

- Pressurization of the ISM by SNRs.
- Mass exchange between the phases: cold clouds "evaporated" and converted to diffuse gas, and diffuse gas swept up by SN blastwaves and compressed in the high-pressure shells of radiative SNRs.
- Injection of high-velocity clouds by fragmentation of the dense shell present in radiative SNRs.

The principal shortcoming of the this model is the failure to predict the substantial amount of warm HI that is present in the ISM. The model parameters have only 4.3% of the HI mass in the warm phase (WNM and WIM). However, 21 cm line observations indicate that more than 60% of the HI within 500 pc of the Sun is actually in the warm phase.

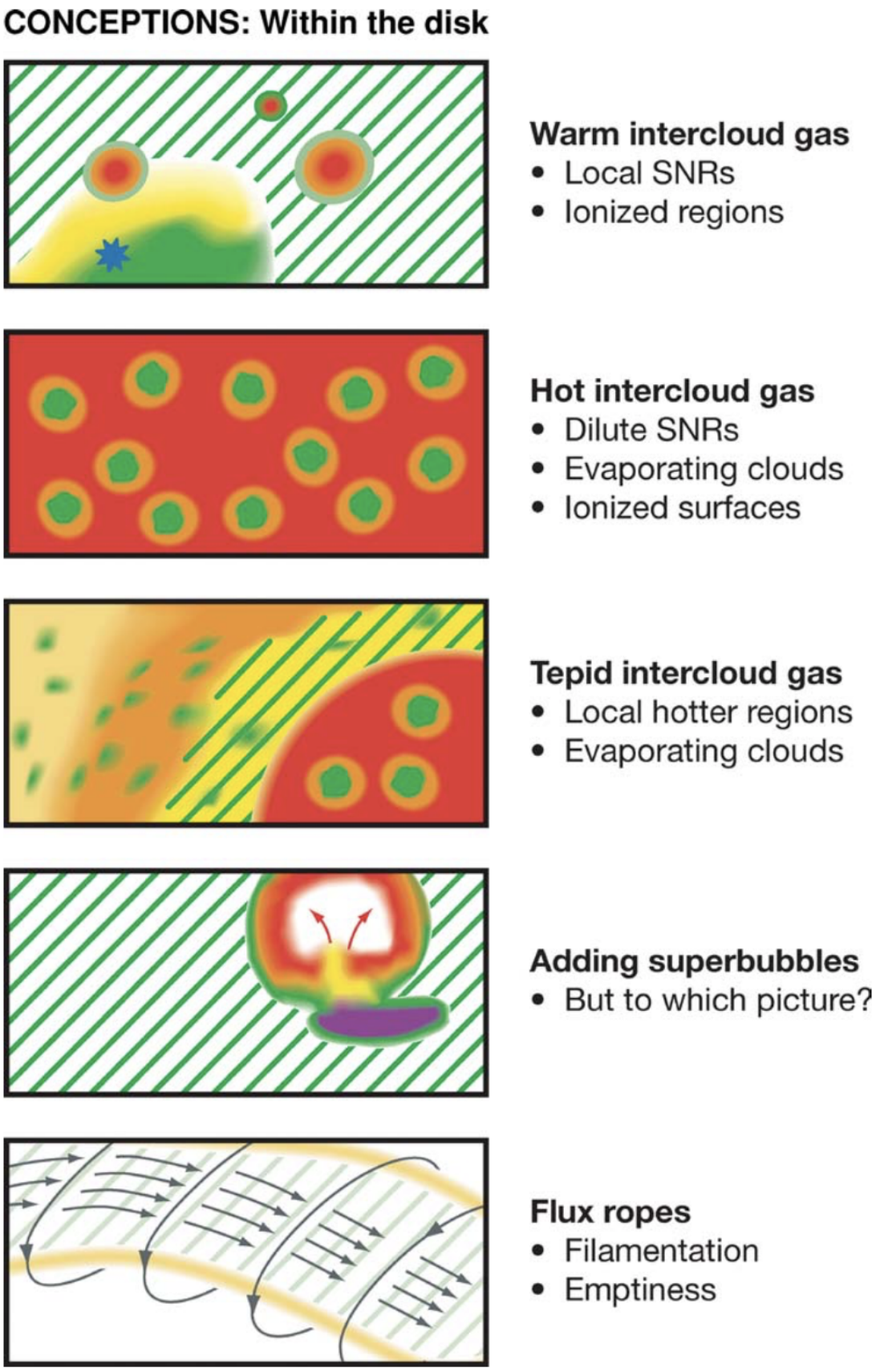

Figure 44: Various conceptions of the ISM within the disk. In this figure, purple indicates dark molecular clouds; solid greencold HI clouds; hatched greenwarm HI; hatched green on yellow backgrounddiffuse warm HII; orange-hotter gas bearing OVI; redmaterial hot enough to emit X rays; and (in this figure only) black lines with arrowheadsthe magnetic field. A blue star in the top panel is shown contributing to the ionization of the diffuse gas in its vicinity. None show the correct nature of the inhomogeneity. Figure taken from Cox (2005).

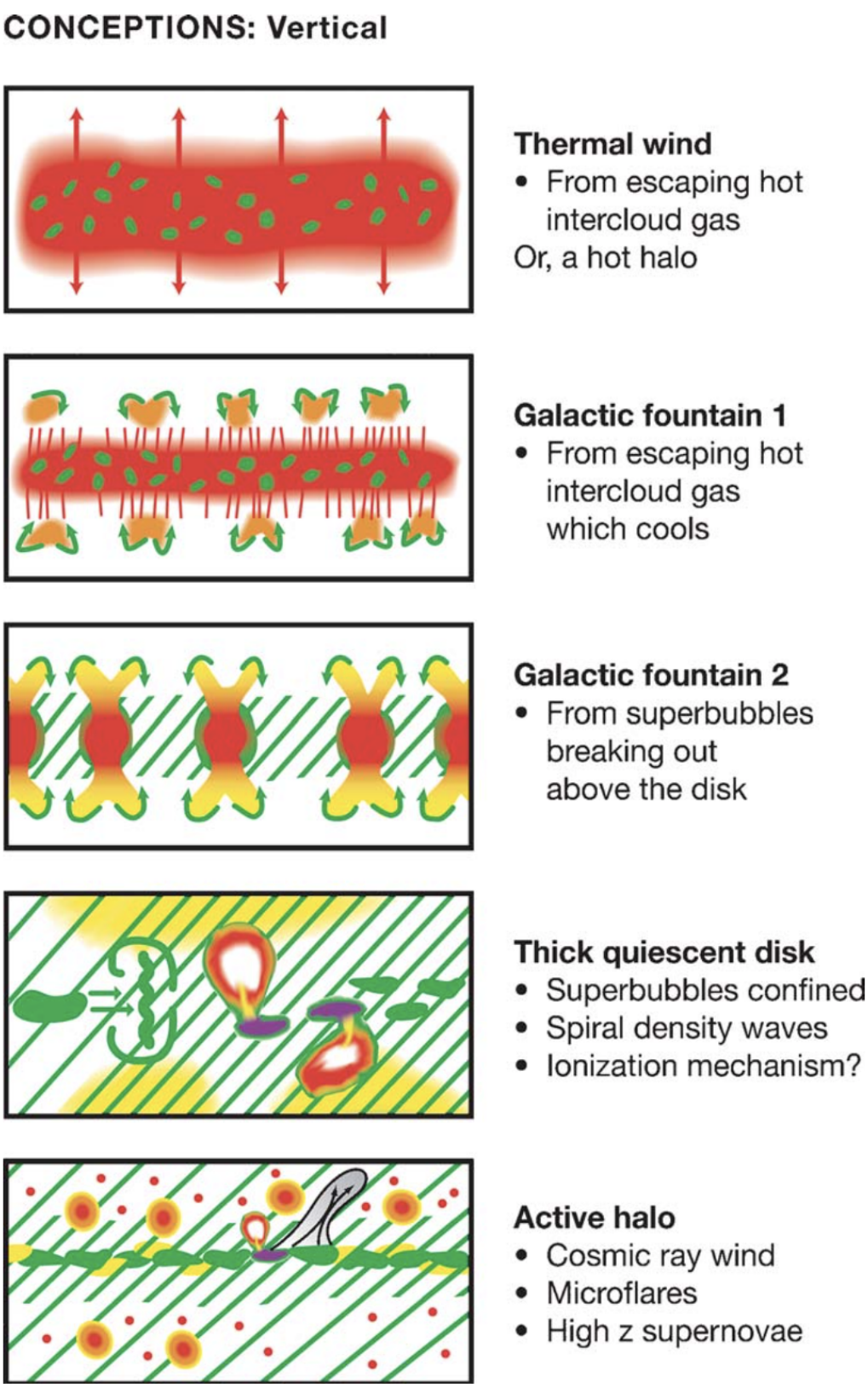

Figure 45: Various conceptions of the ISM within about 300 pc of the midplane. In this figure, purple indicates dark molecular clouds; solid greencold HI clouds; hatched greenwarm HI; hatched green on yellow backgrounddiffuse warm HII; orangehotter gas bearing OVI; and redmaterial hot enough to emit X rays. The green arrows in the fourth panel indicate material flowing into a spiral arm and encountering a combination shock and hydrualic jump (shown as a vertical squiggly green line with arcs enclosing it at high $|z|$). The red dots in the bottom panel represent microflares, releases of magnetic energy via localized reconnection events; whereas the gray region is a column of escaping cosmic rays and the associated distortions of the magnetic field. None show the correct nature of the inhomogeneity. Figure taken from Cox (2005).

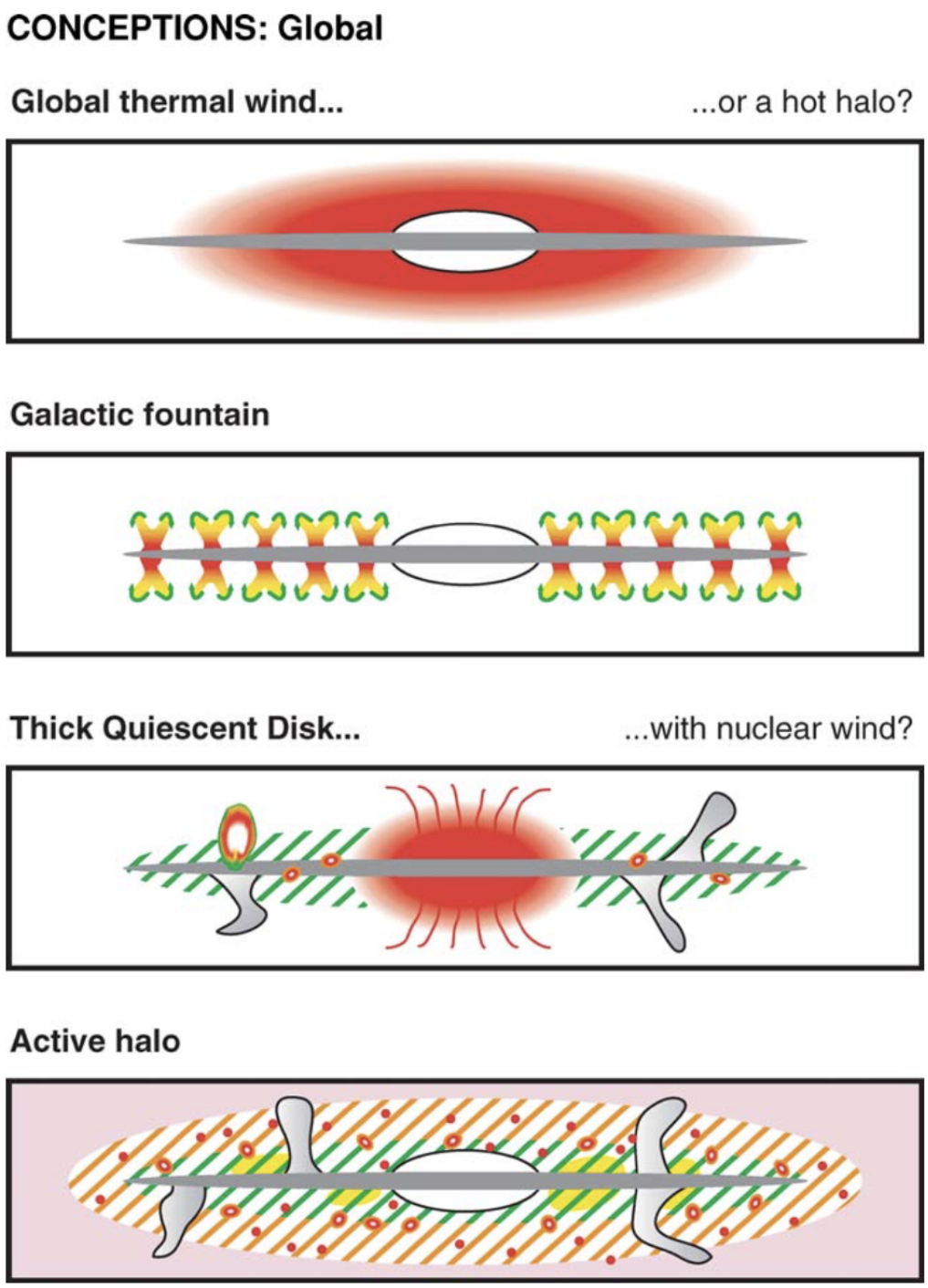

Figure 46: Various conceptions of the larger scale structure of the Galactic atmosphere. In this figure, hatched green indicates warm HI; hatched green on yellow backgrounddiffuse warm HII; orangehotter gas bearing OVI; redmaterial hot enough to emit X rays; grayplumes of escaping cosmic rays; and red dots-microflares. The lower two panels contain some elements of potentially greater realism. Figure taken from Cox (2005).

1.15.3 Follow-up Questions

- Write down typical temperatures and densities for each phase.
- Where do you find each of these phases?
- Why don't we see molecular gas (H_2) in all of these phases?
- Describe what each of these regions might looks like.
- How do constituents change between the different thermal phases?

1.16 Question 16

Characterize the stellar populations in the following regions: i) the Galactic bulge ii) the Galactic disk, outside of star clusters iii) open star clusters iv) globular clusters v) the Galactic halo vi) a typical elliptical galaxy.

1.16.1 Short answer

Answer.

1.16.2 Additional context

i) The Galactic bulge: Stars in the Galactic bulge (and halo) have smaller metallicities than those in the disk; this includes population II (pop II) that are metal-poor ($Z \sim 0.001$)
ii) The Galactic disk: Stars in the disk have a higher metallicity than those in the bulge and halo; this includes population I (pop I) stars which have a Solar-like metallicity ($Z \sim 0.02$). Obviously, star formation in our MW takes place mainly in the spiral arms within the disk.
iii) Open star clusters:
iv) Globular clusters:
v) The Galactic halo: Stars in the Galactic halo (and bulge) have smaller metallicities than those in the disk; this includes population II (pop II) that are metal-poor ($Z \sim 0.001$)
vi) Typical elliptical galaxy:

1.17 Question 17

How can one determine the temperature of an HII region?

1.17.1 Short answer

The temperature of an HII region can be determined using three techniques:

- Nebular diagnostics.
- The Balmer jump in the emission spectrum of the recombining hydrogen.
- Emission lines that follow dielectronic recombination.

1.17.2 Additional context

Hot stars photoionize the gas around them; the photoionized gas is referred to as an HII region, because the hydrogen is predominantly ionized. The dominant heating process is photoionization: ionizing photons have energies larger than the ionization threshold, and the resulting photoelectron will have nonzero kinetic energy, adding to the thermal energy of the gas. At the same time, recombination processes (primarily radiative recombination) are removing electrons from the plasma, along with the kinetic energy that they possessed just before the recombination event, and thermal energy is also lost when electron collisions excite ions from lower to higher energy levels, followed by emission of photons. The temperature of the gas is thus determined by a balance between the heating and cooling processes.

Nebular diagnostics: The populations of excited states of atoms and ions depend on the local density and temperature. Therefore, if we can determine the level populations from observations, we can use atoms and ions as probes of interstellar space. Nebular diagnostics allow us to probe the density and temperature of photoionized gas ("**emission nebulae**") in the temperature range $3000 \lesssim T \lesssim 3 \times 10^4\,\mathrm{K}$. To be a useful probe, an atom or ion must be sufficiently abundant to observe, must have energy levels that are at suitable energies, and must have radiative transitions that allow us to probe these levels, either through emission lines or absorption lines.

There are two principal types of nebular diagnostics. The first type of diagnostic uses ions with two excited levels that are both "energetically accessible" at the temperatures of interest, but with an energy difference between them that is comparable to $k_B T$, so that the populations of these levels are sensitive to the gas temperature. The level populations are normally observed by their line emission.

The second type of diagnostic uses ions with two or more "energetically accessible" energy levels that are at nearly the same energy, so that the relative rates for populating these levels by collisions are nearly independent of temperature. The ratio of the level populations will have one value in the low-density limit, where every collisional excitation is followed by spontaneous radiative decay, and another value in the high-density limit, where the levels are populated in proportion to their degeneracies. If the relative level populations in these two limits differ (as, in general, they will), then the relative level populations (determined from observed emission line ratios) can be used to determine the density in the emitting region.

Once the temperature of the gas has been determined, abundances of emitting species can be estimated using the strengths of collisionally excited emission lines relative to the emission from the ionized hydrogen. When fine-structure emission lines are used, the inferred abundances are quite insensitive to uncertainties in the temperature determination.

We will discuss two types of temperature diagnostics using collisionally excited optical/UV lines: np^2 and np^4 ions, and np^3 ions. But first, let's give a short overview of atomic structure.

Brief overview of atomic structure: Atoms consist of three subatomic particles: protons, neutrons, and electrons. A proton has a positive charge and a neutron has no charge. Both protons and neutrons are found in the densely packed, positively charged nucleus. The nucleus contains essentially all of the mass of the atom. Electrons carry a negative charge and are found in electron shells surrounding the nucleus. The mass of the electron is considered to be negligible.

All atoms have an atomic number, Z, and a mass number, A. The atomic number, Z, represents its number of protons and its mass number, A, represents its number of protons plus its number of neutrons. All atoms of a particular element will have the same atomic number, however they may differ in their number of neutrons and electrons. Two atoms with the same atomic number but different mass numbers (different numbers of neutrons) are referred to as **isotopes**. The average atomic mass found on the periodic table represents a weighted average of the naturally occurring isotopes for a particular element. Atoms with unequal numbers of protons and electrons produce charged atoms or **ions**.

Electrons within an atom are found in particular **orbitals**. An atomic orbital is a mathematical function that describes the wave-like behavior of either one electron or a pair of electrons in an atom. This function

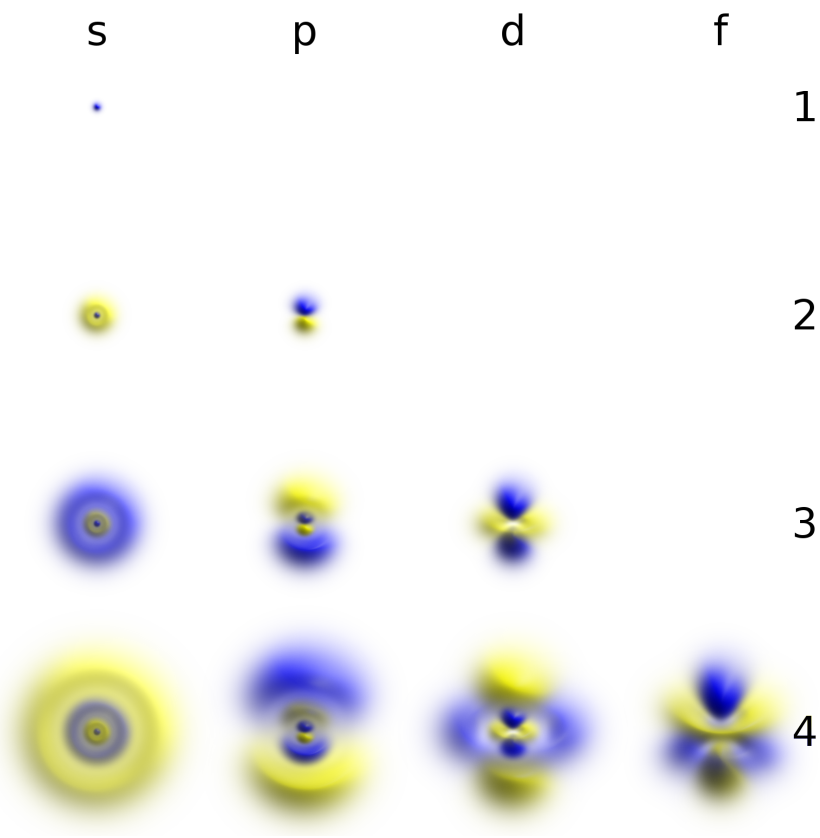

Figure 47: 3D views of some hydrogen-like atomic orbitals. Figure taken from Wikipedia.

can be used to calculate the probability of finding any electron of an atom in any specific region around the atom's nucleus. The term atomic orbital may also refer to the physical region or space where the electron can be calculated to be present, as defined by the particular mathematical form of the orbital.
Electrons within an atom can be assessed according to the shell, subshell, and orbital to which they are assigned. These assessments are based on the quantum mechanical model. **Shells** are numbered as $n = 1$, 2, 3, 4, etc. and increase in size and energy as they get further away from the nucleus. Shells can be subdivided into subshells.
The maximum number of **subshells** is equivalent to the shell number. For example, when $n = 1$ (first shell), only one subshell is possible and when $n = 2$ (second shell), two subshells are possible. There are four different types of subshells. These various types of subshells are denoted by the letters s, p, d, and f. Each subshell has a maximum number of electrons which it can hold: s: 2 electrons, p: 6 electrons, d: 10 electrons, and f: 14 electrons. The s subshell is the lowest energy subshell and the f subshell is the highest energy subshell. As was mentioned previously, the shell number is equal to the possible number of subshells. Thus, when $n = 1$, the only subshell possible is the $1s$ subshell. When $n = 2$, two subshells are possible: the $2s$ and $2p$. When $n = 3$, three subshells are possible: the $3s$, $3p$, and $3d$. This means that in the first shell only two electrons are possible and they would be found in the $1s$ (2 electrons) subshell. In the second shell, 8 electrons are possible and would be found in the $2s$ (2 electrons) and the $2p$ (6 electrons) subshells.
Each subshell is further divided into **orbitals**. An orbital is defined as a region of space in which an electron can be found. *Only two electrons are possible per orbital.* Thus, the s subshell may contain only one orbital and the p subshell may contain three orbitals. Each orbital has its own distinct shape. An s orbital found in an s subshell is spherical, p orbitals found in p subshells are two-lobed, and d orbitals found in d subshells are four-lobed. Since there are three possible orbitals per p subshell, each orbital adopts its own orientation. The p_x orbital lies along the x axis, the p_y orbital lies along the y axis, and the p_z orbital lies along the z axis.
When writing electron configurations for atoms, the shorthand symbol for a subshell followed by a superscripted number which represents the number of electrons in that subshell is used. For example, a Carbon atom with 6 electrons would have the electron configuration: $1s^2 2s^2 2p^2$. A possible arrangement within the $2p$ subshell would be to find one of the electrons in the $2p_x$ orbital and the second electron in the $2p_y$ orbital.
Figure 47 shows hydrogen-like atomic structure for s, p, d, and f orbitals.

Configuration	Ion	$n_{\rm crit}(e)$ at $T = 10^4$ K				
		$^3P_0$	$^3P_1$	$^3P_2$	$^1D_2$	$^1S_0$
$1s^2 2s^2 2p^2$	C I	–	7.37×10^0	1.21×10^1		
	N II	–	1.67×10^2	2.96×10^2	7.68×10^4	1.23×10^7
	O III	–	1.74×10^3	3.79×10^3	6.40×10^5	2.78×10^7
	Ne V	–	3.19×10^5	3.48×10^5	1.44×10^8	9.58×10^8
$1s^2 2s^2 2p^4$	O I	3.11×10^3	2.87×10^4	–	1.62×10^6	4.04×10^8
	Ne III	3.02×10^4	2.76×10^6	–	9.47×10^6	1.37×10^8
	Mg V	4.36×10^6	4.75×10^7	–	1.07×10^9	8.07×10^9
$1s^2 2s^2 2p^6 3s^2 3p^2$	Si I	–	7.72×10^2	1.92×10^3		
	S III	–	4.22×10^3	1.31×10^4	7.33×10^5	1.52×10^7
	Ar V	–	1.09×10^7	1.16×10^7	3.65×10^8	2.49×10^8
$1s^2 2s^2 2p^2 3s^2 3p^4$	S I	1.04×10^5	1.55×10^5	–	4.12×10^7	1.38×10^9
	Ar III	2.49×10^5	2.67×10^6	–	1.26×10^7	4.54×10^8

Table 4: Critical electron density $n_{\rm crit}(e^-)[\ {\rm cm}^{-3}]$ for selected np^2 and np^4 ions. Figure taken from Draine (2011).

Configuration	Ion	$n_{\rm crit}(e)$ at $T = 10^4$ K			
		$^2D^o_{3/2}$	$^2D^o_{5/2}$	$^2P^o_{1/2}$	$^2P^o_{3/2}$
$1s^2 2s^2 2p^3$	N I	2.18×10^4	1.19×10^4	7.11×10^7	3.15×10^7
	O II	4.49×10^3	3.31×10^3	5.30×10^6	1.03×10^7
	Ne IV	1.40×10^6	4.66×10^5	4.17×10^8	2.79×10^8
$1s^2 2s^2 2p^6 3s^2 3p^3$	S II	1.49×10^4	1.57×10^3	1.49×10^6	1.91×10^6
	Ar IV	1.35×10^6	1.55×10^4	1.06×10^7	1.81×10^7

Table 5: Critical electron density $n_{\rm crit}(e^-)[\ {\rm cm}^{-3}]$ for selected np^3 ions. Figure taken from Draine (2011).

np² and np⁴ ions: Atoms or ions with six electrons have $2p^2$ as their lowest configuration: the ground state term is ^3P, and the first two excited terms are ^1D and ^1S. If the ^1S term is at a low enough energy ($E/k_B \lesssim 70000\,{\rm K}$), so that the rate for collisional excitation in gas with $T \approx 10^4\,{\rm K}$ is not prohibitively slow, and the abundance of the ion itself is not too low, then the ion can produce observable line emission from both the ^1D and ^1S levels. Because these levels are at very different energies, the relative strengths of the emission lines will be very sensitive to the temperature; the measured intensity ratio can be used to determine the temperature in the nebula.
Candidate $2p^2$ ions are CI, NII, OIII, FIV, NeV, and so on. CI is easily photoionized, and will have very low abundance in an HII region. The ionization potentials of FIV, NeV, and so on exceed 54.4 eV, and we do not expect such high ionization stages to be abundant in H II regions excited by main-sequence stars with effective temperatures $k_B T_{\rm eff} \lesssim 5\,{\rm eV}$. This leaves NII and OIII as the only $2p^2$ ions that will be available in normal HII regions.
Systems with eight electrons will have $2p^4$ configurations that will also have ^1D and ^1S as the first two excited terms. For OI, FII, and NeIII, the ^1S term is at $E/k_B < 70000\,{\rm K}$.
Similar considerations for systems with 14 or 16 electrons in HII regions photoionized by main-sequence stars leave PII and SIII as the only $3p^2$ ions, and ClII, ArIII, and KIV as the only $3p^4$ ions, that can be used for temperature determination by comparison of emission lines from the ^1D and ^1S levels.
It is easy to calculate what happens in the limit of very low densities, in which case essentially all of the NII and OIII ions will be in the ground state $^3P_0$. Let C_{03} and C_{04} be the rates for collisional excitation from the ground state to the $^1D_2$ and $^1S_0$ excited states. At low densities, every collisional excitation will be followed by radiative decays returning the ion to the ground state, with branching ratios that are determined by the Einstein coefficients A_{ul}. For example, after excitation of level 4, the probability of a $4 \rightarrow 3$ radiative transition is $A_{43}/(A_{41} + A_{43})$. Thus the power radiated per unit volume in the $4 \rightarrow 3$ and $3 \rightarrow 2$ transitions is

$$P(4 \rightarrow 3) = E_{43}[n_0 C_{04}] \frac{A_{43}}{A_{43} + A_{41}},$$
$$P(3 \rightarrow 2) = E_{32} \left[n_0 C_{03} + n_0 C_{04} \frac{A_{43}}{A_{43} + A_{41}} \right] \frac{A_{32}}{A_{32} + A_{31}},$$

where

$$C_{lu} = 8.629 \times 10^{-8} T_4^{-1/2} \left(\frac{\Omega_{lu}}{g_{lu}} \right) e^{-E_{ul}/k_B T} n_e \ [{\rm cm}^3\,{\rm s}^{-1}].$$

Thus, in the limit $n_e \rightarrow 0$, the **emissivity ratio**

$$\frac{j(4\to 3)}{j(3\to 2)} = \frac{A_{43}E_{43}}{A_{32}E_{32}} \frac{(A_{32}+A_{31})\Omega_{04}e^{-E_{43}/k_BT}}{(A_{43}+A_{41})\Omega_{03}+A_{43}\Omega_{04}e^{-E_{43}/k_BT}} \text{ [dimensionless].}$$

Therefore, in the low-density limit, the emissivity ratio depends only on the atomic physics (A_{ul}, E_{ul}, Ω_{ul}) and the gas temperature T. If the atomic physics is known, the observed emissivity ratio can be used to determine T. The low-density limit applies when the density is below the critical density for both $^1D_2$ and $^1S_0$. The critical densities for NII and OIII are given in Table 4.

The steady-state level populations have been calculated as a function of T for NII and OIII, and the ratios of emission lines from the $^1S_0$ and $^1D_2$ levels are shown in Figure 48. We see that if $n_e \ll n_{\rm crit}$ for the $^1D_2$ level ($n_{\rm crit} = 8\times 1^0 4\,{\rm cm}^{-3}$ for NII, and $6\times 10^5\,cm^{-3}$ for OIII), the line ratio is independent of the density, and depends only on the temperature. Fortunately, these values of $n_{\rm crit}$ are high enough so that these temperature diagnostics are useful in many ionized nebulae (e.g., the Orion Nebula, with $n_e \approx 3000\,{\rm cm}^{-3}$).

Note that the fine-structure excited states of the ground ^3P term have values of $n_{\rm crit}$ that are considerably lower than $n_{\rm crit}$ for $^1D_2$ and $^1S_0$ levels, because the fine-structure levels of the ground term have radiative lifetimes that are much longer than the excited terms.

However, when **L-S coupling** is a good approximation, quantum-mechanical calculations of the collision strengths Ω_{ul} for different fine-structure levels l within a single term (e.g., $^3P_{0,1,2}$) have $\Omega_{ul}/\Omega_{ul'} \approx g_l/g_{l'}$. When this is true, the collisional rate coefficients for excitation out of the different fine-structure levels will be nearly the same for the different fine-structure levels, so it does not matter whether these levels are populated thermally or whether the only level occupied is the ground state $^3P_0$.

Ions with 14 electrons (SIII is an example) have $1s^22s^22p^63s^23p^2$ configurations with the same term structure as $1s^22s^22p^2$, and therefore can be used for temperature determination in the same way. Figure 48 shows how the ratio [SIII]6313.8/[SIII]9533.7 serves as a temperature diagnostic.

A fundamental assumption is that the levels producing the observed lines are populated only by collisional excitation. The $^1D_2$ level is at sufficiently high energy that as the temperature T is lowered below $\sim 5000\,{\rm K}$, the rate of collisional excitation becomes very small. This means that the line becomes very weak and difficult to observe; it also means that if the next ionization state (NIII, OIV, SIV) has an appreciable abundance, radiative recombination with electrons may make a significant contribution to population of the $^1D_2$ level. As a result, the observed line ratios may not be suitable for temperature determination when the intensity $I(^1D_2 \to ^1S_0) \lesssim 10^{-3} I(^1S_0 \to ^3P_J)$.

np³ ions: Atoms or ions with seven electrons have $1s^22s^22p^3$ as their lowest configuration: the ground term is $^4S^0_{3/2}$, and the first two excited terms are $^2D^0_{3/2,5/2}$ and $^2P^0_{1/2,3/2}$. Candidate ions are NI, OII, FIII, NeIV, and so on. NI will be photoionized in HII regions, leaving OII, FIII, and NeIV as the $2p_3$ ions suitable for observation in HII regions.

Atoms or ions with 15 electrons have $1s^22s^22p^63s^23p^3$ as their lowest configuration. Just as for $2p^3$, the ground term is $^4S^0_{3/2}$, and the first two excited terms are $^2D^0_{3/2,5/2}$ and $^2P^0_{1/2,3/2}$. Candidate ions are PI, SII, ClIII, and ArIV. PI is easily photoionized, leaving SII, ClIII, and ArIV as the $3p^3$ ions that will be present in regions with $h\nu > 13.6\,{\rm eV}$ radiation extending possibly up to 54.4 eV.

The ratio of the intensities of lines emitted by the $^2P^0$ term to lines from the $^2D^0$ term is temperature-sensitive. Figure 48 shows [OII](7322 + 7332)/(3730 + 3727) and [SII](6718 + 6733)/(4070 + 4077) as functions of temperature. For these two ions, the critical density for $^2D^0$ is relatively low (see Table 5), so that these T-sensitive line ratios are also sensitive to n_e for $n_e \gtrsim 300\,{\rm cm}^{-3}$. Because of this sensitivity, the np^3 ions are only useful if n_e is known, or is known to be $\leq 10^2\,{\rm cm}^{-3}$.

Balmer jump: It is possible to determine the temperature from the strengths of the discontinuities in the recombination continuum relative to the strengths of recombination lines. The most commonly used discontinuity is the "**Balmer jump**" at $\lambda = 3645.1$Å:

$$\text{BJ} \equiv I_\lambda(\lambda_{\rm BJ,blue}) - I_\lambda(\lambda_{\rm BJ,red})\ [{\rm erg\,s^{-1}\,cm^{-2}\,Hz^{-1}\,sr^{-1}}],$$

where $\lambda_{\rm BJ,blue}$ is chosen to be just blueward of the jump, and $\lambda_{\rm BJ,red}$ is chosen to be slightly redward of the jump, and to be located between H recombination lines. For example, $\lambda_{\rm BJ,red} = 3682.6$Å would fall between the H20 and H21 lines.

The "jump" discontinuity is produced by recombining electrons with zero kinetic energy, and is therefore proportional to the electron energy distribution at $E = 0$, and therefore BJ $\propto$ EM $\times\, T^{-3/2}$, where EM is the **emission measure**.

The strength of a recombination line such as the H11 line ($n = 11 \to 2$ at $\lambda = 3769.7$Å) is proportional to rates of radiative recombination to levels $n \geq 11$. The effective recombination rate coefficient for emitting H11 will vary approximately as $T^{-0.8}$ near the temperatures of interest, and the intensity of the recombination line I(H11) $\propto$ EM $\times\, T^{-0.8}$. Thus we expect BJ/I(H11) $\propto T^{-0.7}$: the dependence on T is strong enough that this is a useful diagnostic.

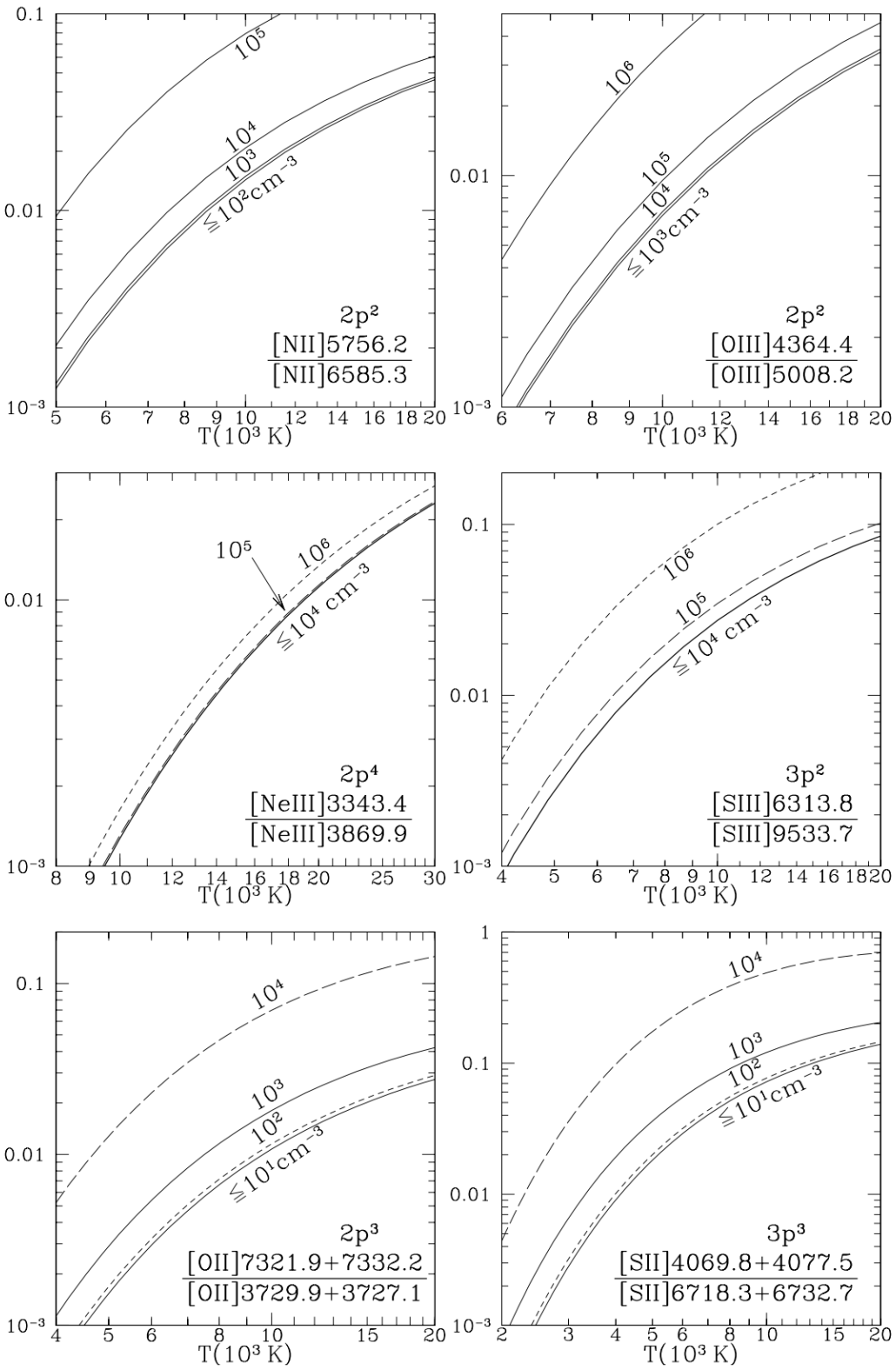

Figure 48: Critical electron density $n_{\rm crit}(e^-)[\ {\rm cm}^{-3}]$ for selected np^2 and np^4 ions. Figure taken from Draine (2011).

Allowance must be made for the contribution from helium: doubly ionized helium recombining to level $n = 4$ contributes to the observed Balmer jump, and HeII $n = 22 \rightarrow 4$ coincides with H11.
This method has been used to determine the electron temperature in H II regions and planetary nebulae. In a sample of 23 planetary nebulae, temperatures $T_{e,\rm BJ}$ derived from the Balmer jump are generally lower than the temperature Te, [OIII] determined from collisional excitation of [OIII] optical lines, with $T_{e,\rm BJ}/T_{e,\rm CE} \approx 0.75 \pm 0.25$. The reason for the discrepancy is unclear; some suggests that cool, dense, metal-rich knots may be present in planetary nebulae.

Dielectronic recombination: For some ions, it is possible to observe both collisionally excited lines and lines emitted following dielectronic recombination. For example, electrons colliding with CIV can produce collisionally excited levels of CIV, but can also produce excited levels of CIII by **dielectronic recombination**. Because the rate coefficients for collisional excitation and for dielectronic recombination will have different temperature dependences, the ratio of dielectronic lines to collisionally excited lines will be temperature-sensitive, and therefore useful as a temperature diagnostic. Examples of useful line ratios are [CIII]2297/[CIV]1549, [OII]4705/[OIII]1665, and [CII]4267/[CIII]1909.

1.18 Question 18

What is the G-dwarf problem in the solar neighborhood?

1.18.1 Short answer

Answer.

1.18.2 Additional context

When forming a stellar system out of gas, supernovae explosions, planetary nebulae, and quieter mass-loss enrich the remaining gas so that succeeding generations of stars are more metal-rich than their ancestors. Uncomplicated analytic models are helpful for exploring the basics of the chemical evolution problem, and the most uncomplicated of them all is termed the **Simple model**, with capitalization intact. The Simple model assumes:

1. that the galaxy is represented by one zone, initially full of metal-free gas, eventually full of stars,
2. that it is a closed box with no inflow or outflow,
3. that enrichment occurs immediately upon forming new stars (the Instantaneous Recycling Approximation),
4. and that the stars produce a constant yield of heavy elements returned to the ISM.

Data from the solar vicinity has been compared to analytic models, and the Simple model in particular. Even after correcting the purely local data to include stars at high Z in the solar cylinder, the Galactic data still displays a paucity of metal-poor stars compared to the Simple model. This is referred to as the **G Dwarf problem** after the unevolved tracer stars used for the abundance determinations. The G Dwarf problem can be alleviated in a variety of ways, including gas infall, gas outflow, prompt initial enrichment, consideration of a spatially inhomogeneous metal abundance pattern, or of a variable stellar yield. In addition, many modelers now consider many-zone models, many track individual elements (as opposed to one overall heavy-element enhancement), and many drop the Instantaneous Recycling Approximation. It has been established in different regions of the Galaxy: the solar neighbourhood and, to a lesser extent, the halo and the bulge. In addition, a G-dwarf problem has been detected in both bulge-dominated and disk-dominated galaxies, which is consistent with the idea that the G-dwarf problem is universal.
Until recently, the solar vicinity was the only region in which chemical evolution models could be applied. The metal-poor dwarf spheroidals have small spreads in metallicity based on the width of the red giant branch (RGB) much like Galactic globulars, the vast majority of which are tightly constrained to a single metallicity throughout. In 1988 the distribution of metallicities appeared to match the Simple model fairly well, but subsequent recalibration of the metallicity scale narrows the distribution somewhat by pushing the very metal rich stars back toward solar metallicity. Adjustments both for differences of RGB lifetime with [Fe/H] and for the fact that metal-rich K giants are underrepresented because they become M giants further narrow the abundance distribution. Recently, other galaxies have begun to be studied. A relatively deep HST color-magnitude diagram of individual stars in local group elliptical M32 has been decomposed into a metallicity distribution that is significantly more peaked than the solar neighborhood, along with a field in the outer disk of M31 which shows a distribution almost as narrow. Some have analyzed HST images in the halo of M31, finding an unexpectedly metal-rich environment poor in metal-deficient stars compared to our own halo.

1.18.3 Follow-up Questions

- Is it reasonable to assume that the IMF changes over time? Why?
- How much does the mean molecular weight change over cosmological timescales?
- What is an appropriate value for the mean molecular weight mu? (i.e., 2 for molecular hydrogen setting limits on formation masses.)
- Do we talk about upper mass limits because more massive stars can't exist, or because they don't exist?

1.19 Question 19

Describe the general characteristics of spiral structure in galaxies.

1.19.1 Short answer

Short answer.

1.19.2 Additional context

The spiral arms are the bluest regions in spirals and they contain young stars and HII regions. For this reason, the brightness contrast of spiral arms increases as the wavelength of the (optical) observation decreases. In particular, the spiral structure is very prominent in a blue filter, as is shown impressively in Figure 49. The spiral structure itself varies from galaxy to galaxy, even within a Hubble type. A classification system based on the regularity of the spiral arms was introduced in the 1960s by Sidney van den Bergh and revised in the 1980s by Debra and Bruce Elmegreen. Approximately 10% of spiral galaxies contain only a single, bisymmetric spiral that extends in a grand design from the edge of the galactic bulge to the outer limit of the perceptible disk. Most galaxies have a multiple-arm spiral structure, or even a highly chaotic or **flocculent** spiral-like structure. Multiple arm galaxies represent 60% of the early and intermediate Hubble types with bars and intermediate types without bars. Flocculent spirals represent 60% of early Hubble types without bars.

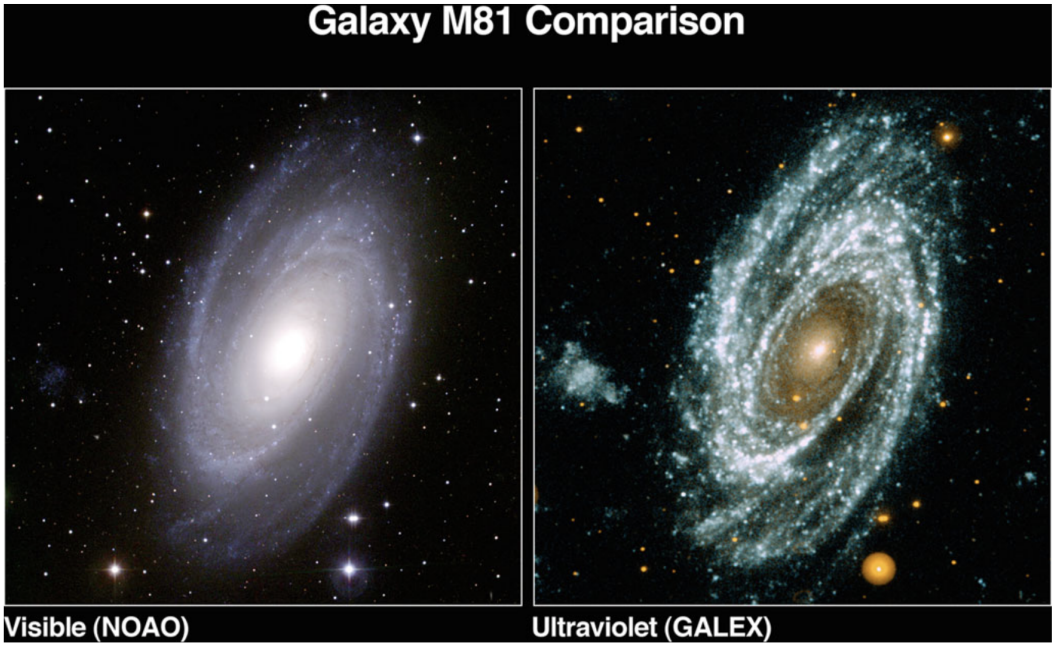

Figure 49: The galaxy M81 in optical light (left) and the UV (right). The spiral arms are much more prominent in the UV than in optical light, showing that star formation occurs almost exclusively in spiral arms. Note the absence of any visible UV-emission in the center of the galaxy, indicating the lack of hot stars there. Credit: NASA/JPL-Caltech/NOAO. Figure taken from Schneider (2015).

Flocculent galaxies contain many more, and much shorter, spiral arms than multiple arm galaxies. Such short arms could form by the same mechanisms as the longer arms in multiple arm galaxies, with the larger number of arms causing the chaotic appearance, or they could differ because of a lack of global wave stimulation, or an inability to amplify global waves in flocculent galaxies. Bars and oval distortions may drive spiral waves in some galaxies. *Among galaxies with early Hubble types, those with bars are twice as likely to have symmetric spirals in their outer disks as those without bars. Among intermediate and late-type galaxies, however, those with bars have approximately the same proportions of grand design, multiple arm, and flocculent spirals as those without bars.* This dependence of the bar-spiral correlation on Hubble type may result from a variation of bar length and strength with Hubble type, that is, the bars in early-type galaxies tend to be larger and stronger than the bars in late-type galaxies.
Seventy percent of galaxies in the nearby Universe are characterized by a disk with prominent spiral arms, but our understanding of the origin of these patterns is incomplete, even after decades of theoretical study. Probably the most obvious answer would be that they are material structures of stars and gas, rotating around the galaxy's center together with the rest of the disk. However, this scenario cannot explain

spiral arm structure since, owing to the differential rotation, they would wind up much more tightly than observed within only a few rotation periods.

Several ideas have been proposed to explain the formation of spiral arms. One model posits that these features are large-scale density waves, regions of higher density (possibly $10 - 20\%$ higher than the local disk environment), continuing to propagate in a differentially rotating disk. In particular, this theory argues that the matter in the galaxy (stars and gas) can maintain a density wave through gravitational interactions even in the presence of shear. This density wave remains at least quasi-stationary in a frame of reference rotating around the center of the galaxy at a fixed angular speed, identified with the pattern speed of the spirals, and covers the entire disk. If the gas, on its orbit around the center of the galaxy, enters a region of higher density, it is compressed, and this compression of molecular clouds results in an enhanced star formation rate. This accounts for the blue color of spiral arms. Since low-mass (thus red) stars live longer, the brightness contrast of spiral arms is lower in red light, whereas massive blue stars are born in the spiral arms and soon after explode there as SNe. Indeed, only few blue stars are found outside spiral arms.

An alternative theory proposes that spiral arms are stochastically produced by local gravitational amplification in a differentially rotating disk. The mechanism behind this process is known as **swing amplification** and it can be seeded either by preexisting leading waves or else by the response of a disk to the presence of a co-rotating overdensity, such as a giant molecular cloud. This dynamical response takes the form of **wakelets** in the surrounding medium, each amplified by its own self-gravity through the swinging of leading features into trailing ones owing to the shear. According to this second theory, spiral arms would fade away in one or two galactic years if the driving perturbations were removed, in contrast to the quasi-steady nature of the arms in the first proposed model. Thus, a continuous source of perturbations would be required for these fluctuating spiral patterns to be maintained throughout the lifetime of a galaxy. Indeed, by mimicking the effects of dissipative infall of gas, it has been shown that the addition of fresh particles on circular orbits could cause such spiral patterns to recur, and some have demonstrated that almost any mechanism of dynamical cooling can maintain spiral activity of a similar kind.

However, the pioneering work on density waves in the 1960s was based on analytic theory and invoked various assumptions in order for solutions to be found. For example, the swing amplification analysis involved linear approximations to the equations of motion. The theoretical emphasis since that time has been on identifying driving mechanisms that can sustain the wave in spite of the damping which would cause it to decay in this picture. Indeed, whereas observations show that spiral arms might be density waves, N-body experiments have not yielded long-lived spiral structures, as predicted by the stationary density wave theory. Simulations of cool, shearing disks always exhibit recurrent transient spiral activity and this situation has not changed over the past several decades as computational power has increased.

Some work showed that spiral patterns fade away in numerical simulations of stellar disks if the effects of gas dissipation are not included; the reason is that the disk becomes less responsive as random motions rise owing to particle scattering by the spiral activity and GMCs. Moreover, the debate about the longevity of the arms practically ceased two decades ago because the available computational power did not permit definitive tests of some of the predictions of the theories and also because observations at that time were not sufficiently detailed to discriminate between the two main competing views.

In the past decade, some studies have argued that the continuous infall of substructures in the dark matter halos of galaxies could induce spiral patterns in disks by generating localized disturbances that grow by swing amplification. According to such simulations, the main agent producing transient features would be satellite passages through the inner part of a disk. Because the tidal effects of the satellites are generally small, this process is distinct from interactions thought to be responsible for **grand-design spirals** like M51. However, there are indications that dark matter substructures orbiting in the inner regions of galaxy halos would be destroyed by dynamical processes such as disk shocking, and hence would not be able to seed the formation of spiral structure.

1.20 Resources

- Galaxy Formation; Longair (2008)
- Galaxies in the Universe; Sparke & Gallagher (2007)
- The Galactic Bulge: A Review; Minniti & Zoccali (2007)
- The Interstellar Environment of Our Galaxy; Ferrier (2001)
- Galactic Dynamics; Binney & Tremaine (2011)
- Galaxies: Interactions and Induced Star Formation; Kennicutt, Schweizer & Barnes (1996)
- Stellar Populations; Greggio & Renzini (2011)
- Physics of the Interstellar and Intergalactic Medium; Draine (2011)
- Astrophysics of the Interstellar Medium; Maciel (2013)
- Notes on Star Formation; Krumholz (2015)
- Principles of Star Formation; Bodenheimer (2011)
- Stellar Evolutionary Effects on the Abundances of Polycyclic Aromatic Hydrocarbons and Supernova-Condensed Dust in Galaxies; Galliano et al. (2008)
- Some insights on the dust properties of nearby galaxies, as seen with Herschel; Galliano (2017)
- The Three Phase Interstellar Medium Revisited; Cox (2005)
- Extragalactic Astronomy and Cosmology; Schneider (2015)
- The G-dwarf problem in the Galaxy; Caimmi (2007)
- The G Dwarf Problem Exists in Other Galaxies; Worthey et al. (1996)
- Self-Perpetuating Spiral Arms in Disk Galaxies; D'Onghia et al. (2013)