Dark Matter

Justin Read
justin.inglis.read@gmail.com
http://justinread.net
University of Surrey

Abstract

I discuss the observational and theoretical basis for “dark matter”: an invisible but dominant non-baryonic matter component in the Universe. I show that dynamics, lensing and cosmology all point towards dark matter and suggest that it behaves dynamically like a collisionless fluid. This presents a challenge for “alternative gravity” explanations for dark matter, and lends support to the idea that dark matter is comprised of some new fundamental particle that remains to be discovered. I discuss the latest probes of the nature of this particle, in particular its ‘temperature’ and self-interaction cross section, showing that the latest data are consistent with a non-relativistic ‘cold’ non-self interacting particle. I discuss ‘small scale puzzles’ that challenge this ‘cold dark matter’ model on the scale of individual galaxies, and I show how these are solved if the dark matter fluid is ‘heated up’ at the centres of galaxies due to feedback from massive stars. I conclude with a look to the future and where the field will go next.

Notation

Vectors are denoted in bold $\mathbf{v}$; time derivatives are denoted by a dot $\dot{x} = \frac{dx}{dt}$; and spatial derivatives are denoted by a dash $y' = \frac{dy}{dx}$. We will typically use units of kiloparsecs, Solar masses, kilometres per second and gigayears: L=kpc, M=$\mathrm{M}_\odot$, V=km/s and T=Gyrs, unless otherwise stated. For reference, unit conversions to S.I. values are given in appendix A. We use the usual notation for different coordinate systems, Cartesian: (x, y, z), cylindrical polars: (R, ϕ, z), and spherical polars: (r, θ, ϕ).

Reading list

Suggested further reading for the course:

- Binney and Tremaine 2008
- Peacock 1999
- Peebles 1980
- Weinberg 2008
- Lecture notes on GR by Sean Carroll: `http://preposterousuniverse.com/grnotes/`.
- For more astronomical background: Shu 1982, esp. Part III (Chap. 11-16) Galaxies and Cosmology.

Contents

Lecture 1

Observables

In this lecture we discuss what astronomers can see in the Universe and how. We discuss time and distance scales to get a feel for how huge the Universe is and how long, typically, we have to wait for things to happen.

1.1 What's out there?

Before embarking on the course proper, it's worth a brief summary of what the Universe is made up of; this is summarised in Figure 1.1. The scales are very difficult to grasp. The typical human can comprehend the difference in size between a grain of salt and a giant cathedral. This is a dynamic range of about 10^5. While impressive, this only just about allows us to imagine just how far away from the *Moon* we are! The Universe is a very big place!

1.2 Measuring starlight

Most of what astronomers see in the Universe is star light. Individual stars emit a spectrum remarkably close to that of a perfect black body radiator, and this is shown in Figure 1.2. The total power output from our own star – the Sun – is called its *luminosity*, and is given by: $L_\odot = 3.83 \times 10^{26}\,\mathrm{W}$[1]. (The symbol, $\odot$, will be used a lot throughout this course and just means 'of our sun'.)

The solar luminosity, $L_\odot$, is really the *bolometric* luminosity: the total rate of energy output integrated over all wavelengths. More usually, in astronomy, we use the luminosity output in a particular *waveband* (range of wavelengths). This is of more practical value since astronomical instruments are usually sensitive only over some limited range of wavelengths (an optical telescope, for example). Many such wavebands are used by astronomers. Most common are the Visual band centred on $\lambda = 550\,\mathrm{nm}$; the Blue band centred on $\lambda = 440\,\mathrm{nm}$; and the Ultraviolet band centred on $\lambda = 365\,\mathrm{nm}$. These are marked on Figure 1.2. The U, B, V labels stand for something sensible, like 'visual', but this is not the case for all bands (the infrared bands are labelled I, J, K). Even more confusing is the fact that the exact definition of these bands has evolved along with the instruments and telescopes which astronomers use: not every instrument has the same sensitivity, and their wavelength filters can differ, sometimes by quite a lot! Fortunately, if you are ever confused, there is an excellent review by Fukugita *et al.* 1995, which pretty much covers all of the wavebands you are ever likely to need, and how to convert between them.

1.2.1 Absolute magnitude

Luminosities span an enormous dynamic range in astronomy and it makes sense to use a logarithmic scale. This is called the *absolute magnitude*, and is given by:

[1] Many astronomers still use the *erg* as the unit of energy. For completeness this is defined in appendix A; I will not use this unit in this course.

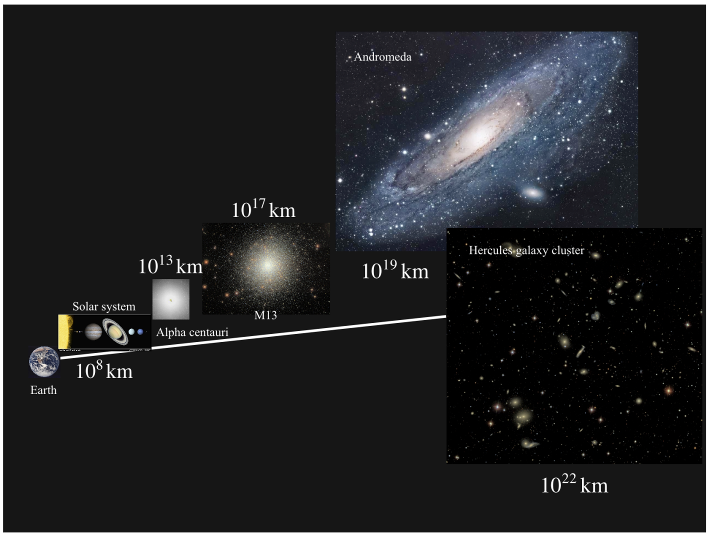


Figure 1.1: The Universe: a very very big place. Alpha Centauri – the nearest star to us – is some 10^{13} km away; that's very close compared to the extent of our Galaxy (called the *Milky Way*): 10^{18} km, or the distance to the Hercules cluster of galaxies some 10^{22} km away. Also marked is the Andromeda galaxy (M31) – the nearest large spiral to our own Galaxy, and the star cluster, M13, which orbits within our Galaxy.

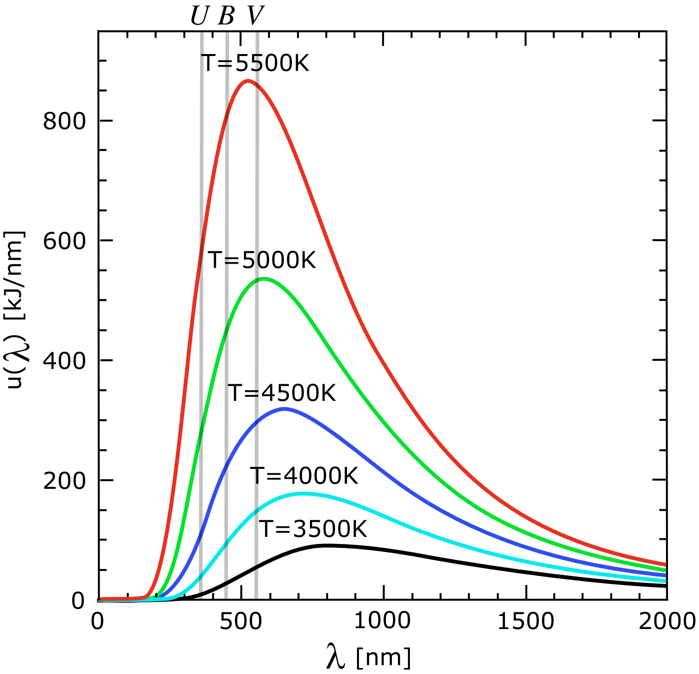


Figure 1.2: Stars emit a near perfect black body spectrum of radiation. Marked on the plot are lines of different black body temperature, T, also known as the *effective temperature* of a star, T_{eff} and the U, B, V wavebands.

$$M \equiv -2.5 \log_{10}\left(\frac{L}{L_\odot}\right) + \text{const.} \tag{1.1}$$

where the constants are chosen separately for each waveband. In the B and V bands, for example, the constants are chosen such that the solar absolute magnitudes are:

$$M_{\odot,B} = 5.48, M_{\odot,V} = 4.83 \tag{1.2}$$

This choice of normalisation is just historical. The system of logarithmic magnitudes comes originally from the fact that luminosities were measured just by eye: the human eye responds on a logarithmic scale.

1.2.2 Flux and apparent magnitude

Flux in astronomy – the actual number of photos arriving per unit area per unit time – is measured using *apparent* magnitudes. The flux is given by: $f = L/(4\pi d^2)$, where d is the distance to the source; the apparent magnitude is given by:

$$m \equiv -2.5 \log_{10}\left[\frac{L}{L_\odot}\frac{(10\,\text{pc})^2}{d^2}\right] + \text{const.} = M + 5\log_{10}(d/10\,\text{pc}) \tag{1.3}$$

again, the choice of normalisation: 10 pc, is historical. The constant is *the same* as in equation 1.1.

1.2.3 Other observed properties of stars

Some other useful definitions are:

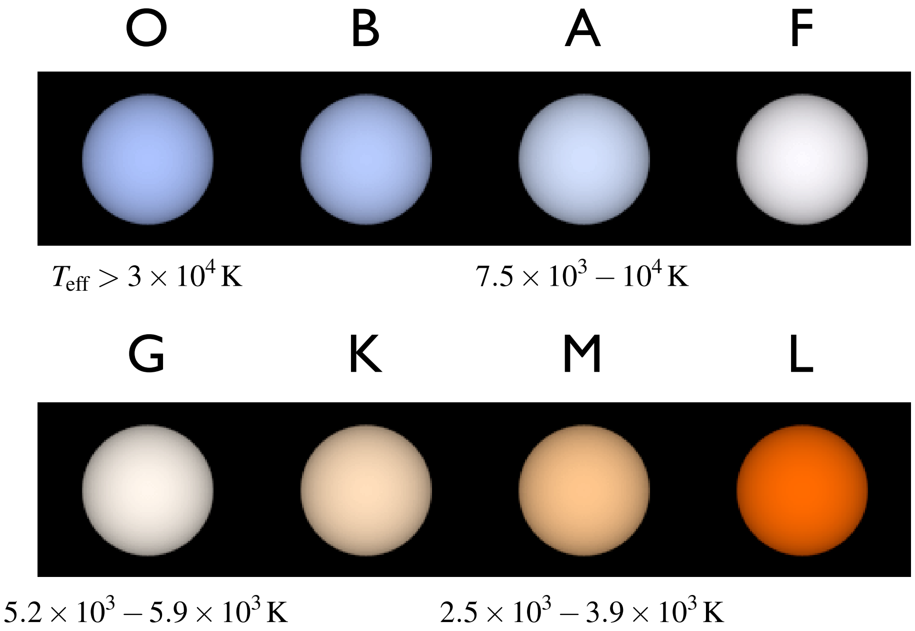
- The *distance modulus*: $m - M = 5\log_{10}(d/10\,\text{pc})$.
- The *color* of a star: L_V/L_B, or $M_B - M_V = m_B - m_V = B - V$. This is useful because it is independent of distance, and because stars are approximately black body radiators. Thus, the colour gives a measure of the star surface temperature.
- The *effective temperature* of a star: the temperature it would have were it a black body radiator (which stars nearly are). Thus the Stefan-Boltzmann law gives: $L = 4\pi R_*^2 \sigma T_{\text{eff}}^4$, defining T_{eff}. R_* is the radius of the star, and $\sigma = 5.670 \times 10^{-8}\,\text{J}\,\text{K}^{-4}\,\text{m}^{-2}\,\text{s}^{-1}$ is the Stefan-Boltzmann constant.
- The *spectral class* of a star. This is a measure of the star's surface temperature (T_{eff}). The historical labels, in order of decreasing temperature are: O, B, A, F, G, K and M (see Figure 1.3); each class is divided into a subclass numbered 0-9 (e.g. B0 is slightly cooler than O9). Our Sun is a G2 star with $T_{\text{eff}} = 5770\,\text{K}$.

An example of some real stellar spectra is given in Figure 1.4. Notice the *absorption features* due to elements in the stellar atmospheres. Measuring these lines allows for much better spectral classification of stars; while measuring their *Doppler shift*[2] allows a determination of the *radial velocity* of a star – its speed towards or away from us. The two absorption lines at just below 500 nm and around 650 nm are Hβ and Hα respectively and are caused by ionised hydrogen.

1.2.4 The Hertzsprung-Russel (HR) diagram

As we have shown, most stars are very nearly perfect black body radiators. They are well defined, observationally, by just two numbers: a colour (which is equivalent to a surface temperature), and a luminosity. A plot of colour v.s. luminosity is called a Hertzsprung-Russel, or HR, diagram and is shown in Figure 1.5. Notice that, as mentioned previously, the colour may be determined from the difference in just two wavebands – in this case $B - V$. This is because stars are so close to black body radiators that only two points along the black body curve are required to define a temperature

[2] The Doppler shift is the shift in the wavelength of the emitted star light due to the motion of the source. Sources moving away from us are redshifted, those moving towards us are blueshifted.


O
B
A
F
$T_{\rm eff} > 3 \times 10^4$ K
$7.5 \times 10^3 - 10^4$ K
G
K
M
L
$5.2 \times 10^3 - 5.9 \times 10^3$ K
$2.5 \times 10^3 - 3.9 \times 10^3$ K

Credit: *Prof. Richard Pogge*

Figure 1.3: Stellar spectral classification.

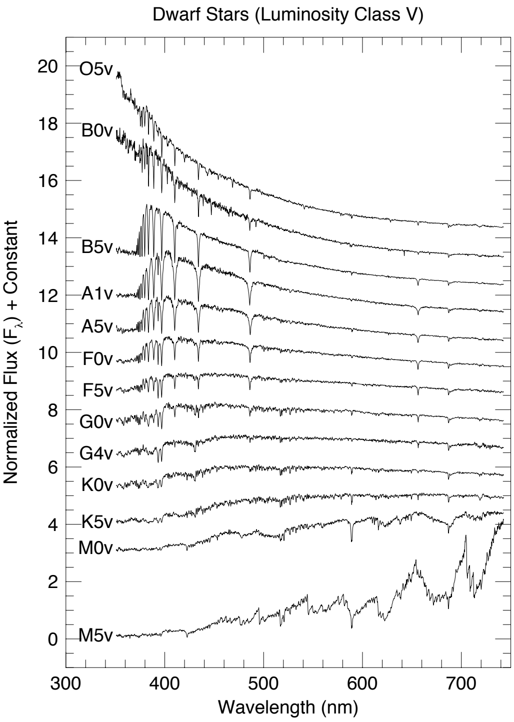


Figure 1.4: Examples of real stellar spectra for O through to M type stars. The wavelength is *rest frame* and has already been corrected for the motion of the stars. The fluxes are offset from one another for clarity.

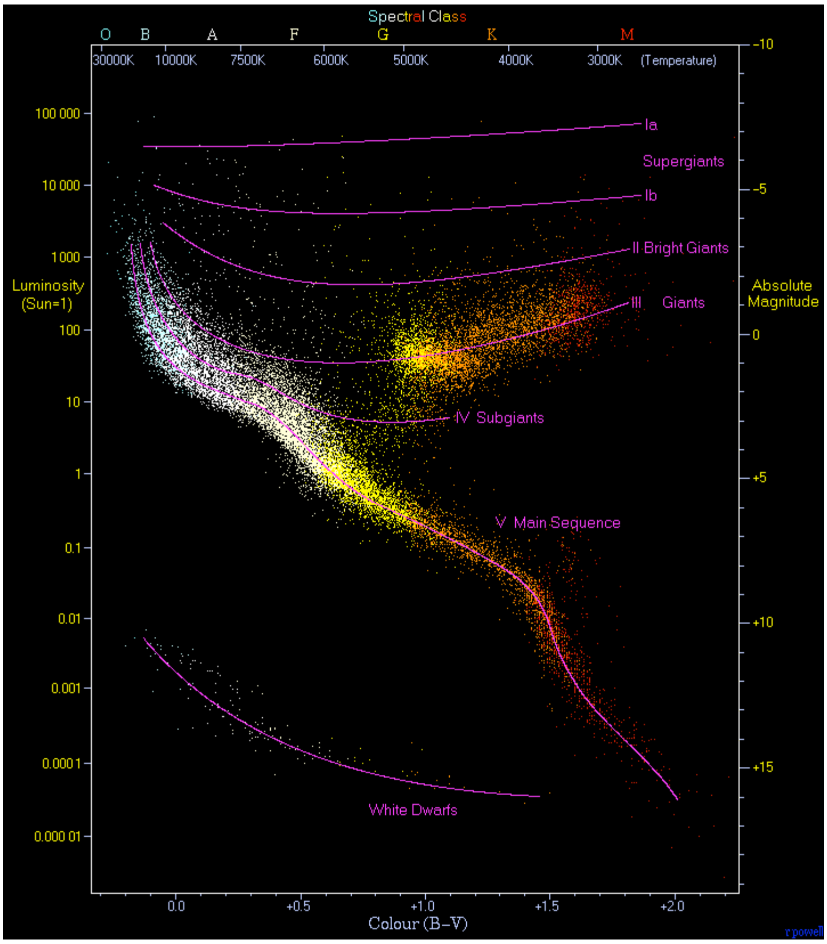


Figure 1.5: The Hertzsprung-Russel (HR) diagram.

(see Figure 1.2). The luminosity may also be represented instead on the logarithmic scale of absolute magnitudes (see equation 1.1). In Figure 1.5, the luminosity is shown along the left axis, and the absolute magnitude is shown along the right. Similarly, colour is shown along the bottom axis and the equivalent surface temperature and associated spectral type is shown along the top axis. Notice that most stars lie along the *main sequence*. These stars are called (once again for historical reasons) *dwarf stars* and are denoted by a V. Depending on their initial mass and chemical composition (known as their *metallicity*[3]), stars are born somewhere along the main sequence. They *do not* evolve along the main sequence. Instead when a proto-star ignites 'burning' hydrogen[4], it moves onto the main sequence. Once stars use up all of their hydrogen fuel, they evolve off of the main sequence and enter a *giant* phase. At the end of their lives, stars eject most of their remaining mass. For the more massive stars, this will be in the form of a *supernovae explosion*. After this mass loss phase, only the very core of the stars remains. Low mass stars become *white dwarfs*, while the most massive stars will end up as *neutron stars* or *black holes*. For a much more detailed account of the lives of stars see Phillips 1999.

[3] It is worth making an important point here. Stars can be almost completely characterised by just three numbers: their total mass, their age, and finally, what astronomers call their *metallicity*. Metallicity is a measure of the chemical composition of stars. Zero metallicity means the stars are composed entirely of hydrogen and helium. Anything heavier than helium is (confusingly) what an astronomer calls a 'metal'. Note that this means we can expect stars to be degenerate on the H-R diagram, which only contains two pieces of information per star.

[4] The use of the verb 'burning' here is standard in astronomy. It refers to, of course, nuclear fusion of the hydrogen into helium and other by-products.

1.2.5 Integrated starlight

In practise, stars are often so far away that they are *unresolved*. This means that in a distant star cluster or galaxy, we really measure the integrated light from many stars. In this case the light from the galaxy is really just the sum of many different stellar spectra such as those shown in Figure 1.4.

1.3 Gas

Astronomers see more than just start light. Useful information about the Universe also comes from gas seen either in absorption or emission. We briefly review some relevant observations for this course:

- HI – pronounced 'H one' refers to observations of *atomic hydrogen* – one proton and one electron. There is a very sharp line in the atomic hydrogen emission/absorption spectrum in the radio at 21cm. It is caused by the transition between the spin states of the proton and electron being aligned and anti-aligned. The rarity of such a highly forbidden transition (once every $\sim 10^7$ yrs) means we do not ever observe it on Earth. In space, however, where the number of hydrogen atoms is astronomical (hehe), such transitions are common. The line is useful because it is naturally so narrow in energy[5]. This means that any broadening of the line observed, must be due to Doppler shifts – the line is useful for measuring *gas kinematics*[6] .

- Hα refers to observations of *excited hydrogen*. The observed photons come from Balmer-α emission from the $n = 3$ to the $n = 2$ transition (recall that $n = 1$ is the ground state). Hα has a wavelength of 656.3 nm (optical light). It is a good tracer of *ionised* gas because it requires little more energy to ionise hydrogen than to excite an electron from $n = 1$ to $n = 3$. Since ionised gas is hot, Hα is often a good indicator of star formation. Where ionised gas exists Hα can be used to trace kinematics.

- CO refers to observations of the roto-vibrational lines from *carbon monoxide* molecules. These emit photons at wavelengths of a few microns. It is a good tracer of cold gas and can also be used for kinematics.

Note that each of the above methods probes *different gas* which may have quite different kinematics, even in the same galaxy – see Figure 1.6; data taken from Simon *et al.* 2003.

1.4 The parsec and the distance ladder

Since astronomers only really see star light, it is notoriously difficult to measure distances: is an object faint and close, or distant and bright? In this section we present the standard measures of distance in astronomy which comprise the 'distance ladder'.

The standard unit of length in astronomy is the "parsec", which stands for **par**allax of one arc**sec**ond[7]. Unlike some astronomical units and conventions, the parsec is of practical, rather than just historic value. It is derived from the *parallax* method of distance measurement, which is first attributed to Hipparchus of Rhodes. As the Earth moves around the Sun, the angular separation of a given nearby star with respect to the very distant background stars (effectively infinitely far away) changes. Knowing the Earth-Sun distance (1 a.u. $= 1.49597892(1)\times10^{11}$ m; determined from radar ranging), then defines the parsec as $1\,\text{pc} = 3.08567802(2)\times10^{16}$ m. This is shown in Figure 1.7.

With the launch of the *Hipparcos* satellite, we can now use the parallax distance measure out to about 1000 pc[8]. This is no mean feat, but as we have seen it is barely an eighth of the distance from

[5] Recall that *isolated* forbidden transitions occur through quantum fluctuations and that the uncertainty principle gives us: $\Delta E \Delta t \geq \hbar/2$. This means that the 21cm line, which has a very long lifetime, must have a very narrow energy. We emphasise *isolated* here because on Earth forbidden transitions proceed through collisional de-excitation. In space, however, the extremely low gas densities make this unlikely.

[6] It is worth making an important point here. *Kinematics* refers to velocity measurements. By contrast *dynamics* involves accelerations. It is very rare in astronomy that we can ever really measure accelerations since changes in velocity occur on such long timescales. This is why we need dynamics. Our dynamical model allows us to calculate accelerations given the positions and velocities of the particles (the measurements). We can then calculate what happened in the past, and what will happen in the future.

[7] 1 arcsec $= 2\pi/360/60/60$ radians. See also appendix A.

[8] At a stretch! *Accurate* determinations are at more like 150 pc.

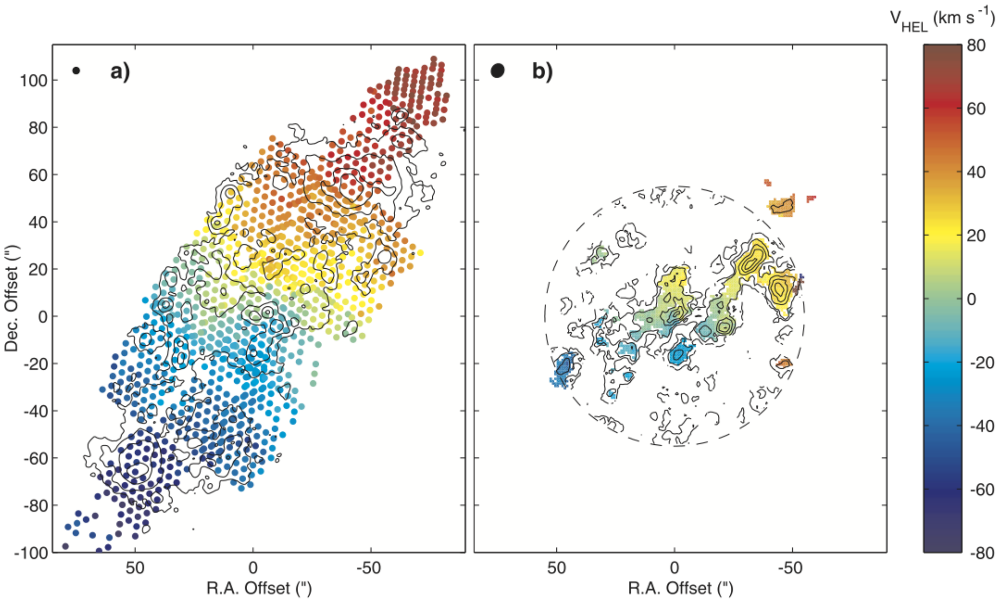


Figure 1.6: (a) Hα velocity field of the dwarf spiral galaxy NGC 2976. The contours show Hα intensity. (b) CO velocity field of the same galaxy; the contours show intensity. The angular resolution of each data set is shown by the filled circles in the top left. Notice how Hα and CO observations trace very different regions of the galaxy. However, the kinematics are similar.

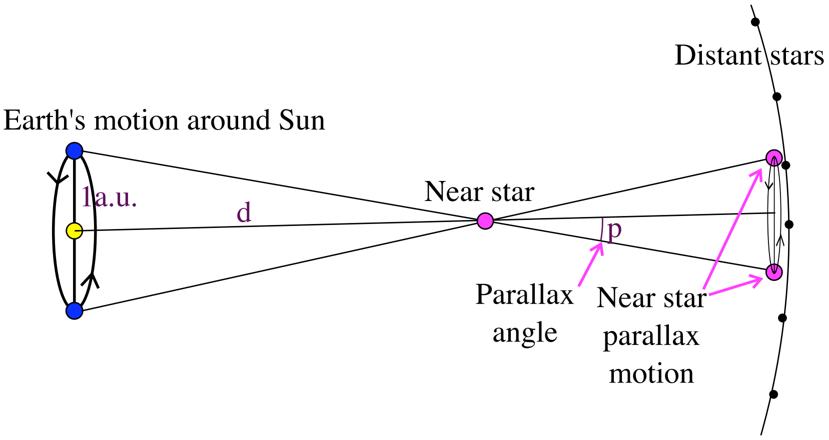


Figure 1.7: Geometric definition of the parsec: $dp = 1$ a.u.; $d = 1$ pc if $p = 1$ arcsec.

Object	*Typical distance*	*Method*
Sun, Solar system	10^{-6} pc	Radar
Hyades star cluster	40 pc	Hipparcos
Galaxy	10^4 pc	Cepheid variable stars
Andromeda	10^5 pc	Cepheid variable stars
Virgo cluster	10^7 pc	Cepheid variable stars
Beyond	$> 10^8$ pc	Hubble expansion: redshift

Table 1.1: The distance ladder. Note that beyond the Virgo cluster, even very bright stars like Cepheids become unresolved and we see only the integrated light from galaxies. Further away than this, we must determine distances using the *redshift* of galaxies.

our Sun to the centre of the Galaxy; not very far in astronomical terms. To measure greater distances, a number of other methods are adopted by astronomers. These are calibrated, first by reference to the parallax distance, and then later to each other, building what is known as the *distance ladder*. For example, Cepheid variable stars are a type of star which pulsate in a periodic fashion related to their luminosity. By calibrating their period-luminosity relation using parallax distance (Hipparcos can *just* about do this), they can be used to measure distances reliably out to the Virgo cluster of galaxies, some 10^7 pc away! The distance ladder is summarised in Table 1.1.

1.5 Measuring velocity

We have already seen that the *radial* velocity of stars and gas can be measured by the Doppler shift of absorption lines in their spectra. The angular velocity of an object on the sky is called its *proper motion* and can be measured for nearby stars, star clusters and galaxies if we are very patient. The idea is very simple: we measure the position of an object relative to a bright distant background source – like a Quasar.[9] Then we return about five years later and measure it again. Even very small movements can be detected if we have high signal to noise. The object may only move 1/200th of a pixel on a CCD, and yet its motion can be detected because the *relative flux* in each pixel will change.

1.6 Timescales

In this section we discuss some important *timescales* in astrophysics. In our Solar system, the relevant timescale is the time it takes planets to orbit around the Sun. This is the *orbit* time, and it gives us a measure of how rapidly things progress on the scale of our Solar system. The orbit time is a relevant quantity on much larger scales in the Universe too. The orbits of stars within a star cluster, of stars within a galaxy and of galaxies within clusters of galaxies all have meaningful orbit times.

Other relevant timescales are the interaction time between the stars within these self-gravitating systems. This governs whether or not a system is *collisional* or *collisionless*. These two regimes lead to very different dynamics. We will focus in this course mainly on collisionless systems.

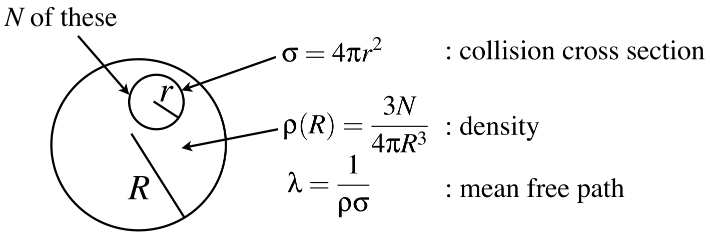
A summary of timescales as a function of scale is given in Table 1.2.

1.6.1 The orbit time

The *orbit* time is the orbital timescale for a particle at radius, r. Using gravitational constant, G, and enclosed mass M, this gives:

$$t_{\rm orb} = \frac{2\pi r}{v_{\rm circ}}; \qquad v_{\rm circ} = \sqrt{\frac{GM}{r}}; \qquad \Rightarrow t_{\rm orb} = 2\pi\sqrt{\frac{r^3}{GM}} \tag{1.4}$$

[9] A Quasar is an extremely bright unresolved galaxy which is typically very far away. They are believed to be so bright because they contain a super-massive black hole at the centre which is consuming a large amount of gas very rapidly and emitting a large amount of energy in the process.



N of these
r
R
$\sigma = 4\pi r^2$
: collision cross section
$\rho(R) = \frac{3N}{4\pi R^3}$
: density
$\lambda = \frac{1}{\rho\sigma}$
: mean free path

Figure 1.8: Calculating t_{coll}: the typical timescale between direct collisions in a self-gravitating system.

1.6.2 The crossing time

The *crossing* time the time taken for a particle to cross the system (galaxy, star cluster, Solar system etc.). The *typical velocity* of the particle is given by $v_{\text{typ}} \simeq v_{\text{circ}} = \sqrt{GM/r}$, where r is the radius of the system and M the mass. The crossing time is then:

$$t_{\text{cross}} = \frac{r}{v_{\text{typ}}} = \sqrt{\frac{r^3}{GM}} \tag{1.5}$$

1.6.3 The dynamical time

The *dynamical* time is the time taken for a particle to fall from a radius r to the centre of a constant density sphere. This is given by:

$$t_{\text{dyn}} = \sqrt{\frac{3\pi}{16G\rho}} \tag{1.6}$$

Given that, for a constant density sphere, $M = 4/3\pi r^3\rho$, the above three timescales are identical to within some small pre-factors. For this reason, they are often used interchangeably.

1.6.4 The [direct] collision time

The direct collision time is the timescale over which direct collisions within an *equilibrium* self-gravitating system occur. Consider a system of size $\sim R$, containing $\sim N$ bodies, each of size r. This is shown in Figure 1.8. Each body has a cross sectional area for collision of $\sigma = 4\pi r^2$. Note that this is *not* the surface area of a sphere; it is a cross sectional area of radius $2r$ – collisions occur when two stars, each of radius r collide. Thus we have $\sigma = \pi(2r)^2$.

The density of bodies is $\rho(R) = \frac{3N}{4\pi R^3}$. The mean free path of each body is $\lambda = \frac{1}{\rho\sigma}$, and the typical velocity of a body within the system is given by $v_{\text{typ}}^2 = \frac{GM}{R}$. Putting this all together gives us:

$$t_{\text{coll}} = \frac{\lambda}{v_{\text{typ}}} = \left(\frac{R}{r}\right)^2 \frac{1}{3N}\frac{t_{\text{orb}}}{2\pi} \tag{1.7}$$

Notice that, for stars, typically $r \ll R$ and direct collisions (almost) never occur. Can you think of somewhere in the Universe where it might occur?

1.6.5 The relaxation time

Direct collisions almost never occur, but gravity is long range! Stars accumulate changes in velocity over time due to both long and short range gravitational interactions. Since such interactions are random in direction, in the mean, they produce no net effect. However, one can think of an individual star receiving velocity kicks from the surrounding stars and undergoing a 3D *random walk* in velocity space. As with the standard random walk, each kick is of random direction, yet over many kicks, a star can lie some way away from its initial velocity; this is shown in Figure 1.9. The *relaxation time*

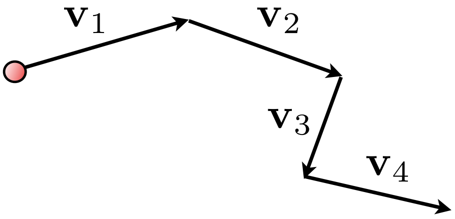


Figure 1.9: A random walk in velocity space. A star (marked by the red circle) starts initially with zero velocity. It receives successive velocity kicks of random direction $\mathbf{v}_1 ... \mathbf{v}_n$. The final velocity is then given by $|\mathbf{v}_t|^2 = |\sum_{i=1}^n \mathbf{v}_i|^2 = \sum_{i=1}^n |\mathbf{v}_i|^2$. The last equality follows because of the random direction of each kick. Notice that it is the root mean squared (r.m.s.) sum of kicks which determines the final velocity magnitude, not the mean of velocity kicks.

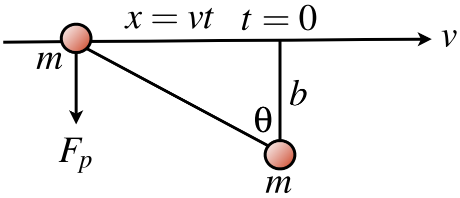


Figure 1.10: Calculating $t_{\rm relax}$: the timescale over which accumulated gravitational interactions turn a star through 45 degrees.

is the time over which these accumulated gravitational interactions *on average* turn a star through 45 degrees[10]. Imagine the interaction between two stars, each of mass m, as shown in Figure 1.10.

By symmetry, we need only consider the perpendicular force on one of the stars, F_p. This is given by:

$$F_p = \frac{Gm^2}{r^2}\cos(\theta) \simeq \frac{Gm^2}{b^2}\left(1 + \left(\frac{vt}{b}\right)^2\right)^{-3/2} \tag{1.8}$$

where b is the *impact parameter*: the perpendicular distance of closest approach between the two stars; and the approximation sign is there to remind us that we have assumed *straight line* trajectories: $x = vt$.

Using Newton's laws: $F_p = m\dot{v}_p$, this gives a change in perpendicular velocity of the star given by:

$$\delta v_p \simeq \int_{-\infty}^{\infty} \frac{Gm}{b^2}\left(1 + \left(\frac{vt}{b}\right)^2\right)^{-3/2} dt = \frac{2Gm}{bv} \tag{1.9}$$

The above is a reasonable approximation provided that $\delta v_p \ll v \Rightarrow b_{\rm min} \gg \frac{Gm}{v_{\rm typ}^2}$.

That was one encounter. Let us assume that the star travels across the system once. If the system is of size $\sim b_{\rm max}$, then the number of other stars it will encounter is given by:

$$dn = \frac{N}{\pi b_{\rm max}^2} 2\pi b db \tag{1.10}$$

where b is the impact parameter as before.

Now, over many encounters, $\overline{\Delta v_p} = 0$, but $\overline{\Delta v_p^2} \neq 0$; this is illustrated in Figure 1.9. Thus the change in the perpendicular velocity of the star when it crosses the system once is given by:

[10] This is *one* of many definitions of the relaxation time. It will suffice for our order-of-magnitude calculation here.

Object	t_{orb}	t_{relax}
Solar system	~ 1 year[11]	–
Hyades open star cluster	$\sim 4 \times 10^6$ yrs	140×10^6 yrs
M13 globular cluster	$\sim 2 \times 10^8$ yrs	5×10^9 yrs
Milky Way Galaxy	$\sim 2 \times 10^8$ yrs	2×10^{16} yrs
Virgo galaxy cluster	$\sim 3 \times 10^9$ yrs	10^{10} yrs
Hercules galaxy cluster	$\sim 6 \times 10^9$ yrs	10^{10} yrs

Table 1.2: Timescales in astronomy.

$$\Delta v_p^2 = \int_{b_{\mathrm{min}}}^{b_{\mathrm{max}}} \delta v_p^2 dn \simeq 8N \left(\frac{Gm}{b_{\mathrm{max}} v_{\mathrm{typ}}}\right)^2 \ln\left(\frac{b_{\mathrm{max}}}{b_{\mathrm{min}}}\right) \tag{1.11}$$

where $v_{\mathrm{typ}} = \sqrt{\frac{GM}{b_{\mathrm{max}}}}$ is the typical stellar velocity (recall that M is the mass of the whole system interior to b_{max}, while m is the mass of one star).

The relaxation time is the time over which accumulated gravitational interactions *on average* turn a star through 45 degrees. This occurs when the star has crossed the system $n_{\mathrm{cross}} = v_{\mathrm{typ}}^2/\Delta v_p^2$ times. Since each crossing takes $t_{\mathrm{cross}} \sim b_{\mathrm{max}}/v_{\mathrm{typ}} = t_{\mathrm{orb}}/(2\pi)$, we find:

$$t_{\mathrm{relax}} = n_{\mathrm{cross}} t_{\mathrm{cross}} \sim \frac{N}{16\pi \ln\Lambda} t_{\mathrm{orb}} \tag{1.12}$$

where $\ln\Lambda = \ln(b_{\mathrm{max}}/b_{\mathrm{min}})$ is known as the *Coulomb logarithm*. Notice that it is set by the *dynamic range* in the system. Since $b_{\mathrm{max}}/b_{\mathrm{min}} = [10, 10000]$ gives $\ln\Lambda = [2.3, 9]$, it is reasonable to assume $\ln\Lambda \sim 10$ for most back of the envelope calculations.

The relaxation time is particularly important. It determines whether or not a self-gravitating system can be thought of as *collisionless*: $t_{\mathrm{relax}} > t_{\mathrm{universe}}$ or *collisional*: $t_{\mathrm{relax}} < t_{\mathrm{universe}}$. Collisionless systems are much easier to model and we will deal almost exclusively with these.

A summary of timescales in astronomy is given in Table 1.2. Notice how slowly things typically orbit; our Sun, for example, cannot have made more than ~ 50 revolutions about the centre of our Galaxy over the entire lifetime of the Universe (~ 14 Gyrs). This means we are very unlikely to actually see anything happen in the Universe. Even when dynamical times are very short, like at the very centre of our Galaxy, we can usually only hope to see stars move enough to measure their transverse velocity across the sky. In general, the Universe must be viewed as a snapshot. By measuring the positions and velocities of stars (using their relative Doppler shifts or proper motions), we can then use dynamics to work out where those stars were in the recent past, and where they will be in the future.

As a final point, have a think about the relaxation times for the clusters of galaxies. We derived the relaxation time in a relatively crude way assuming that all of the objects undergoing relaxation have the *same mass*. Is this likely to be true for galaxies in a cluster of galaxies?

[11] Using this and $r = 1$ a.u. means we can now weigh the Sun!

Lecture 2

Classical evidence for dark matter

In this lecture, I present the classical evidence for dark matter, starting with Fritz Zwicky in the 1930's.

2.1 Coma and the virial theorem

Coma Cluster of Galaxies
Hubble Heritage
NASA, ESA, and The Hubble Heritage Team (STScI/AURA) • Hubble Space Telescope ACS • STScI-PRC08-24

Figure 2.1: The Coma cluster of galaxies as seen by the Hubble space telescope (**Credit:** NASA, ESA, and the Hubble Heritage Team (STScI/AURA)).

In the 1930's, Fritz Zwicky was interested in the Coma cluster of galaxies (Figure 2.1; Zwicky 1933; Zwicky 1937). In these two papers he estimated the mass of the cluster in several different ways. We consider two of these here. First, let us add up all the stellar light we can see. Assuming the 1000 then visible galaxies in Coma make up most of the mass, Zwicky estimated the total visible mass as: $M_{\mathrm{vis}} \sim 10^{12}\Gamma\,\mathrm{M}_\odot$, where $\Gamma \sim 1$ is the mass to light ratio of the stars within the galaxies[1]. Now, let us estimate the cluster mass instead using the *virial theorem*. There are several ways to derive this (we will encounter a different one later on). For now, a simple derivation follows from Newton's laws of motion. Consider the total force $\mathbf{F}_i$ acting on a galaxy i[2] within the cluster:

$$m_i\ddot{\mathbf{x}}_i = \mathbf{F}_i \tag{2.1}$$

where $\mathbf{x}_i$ is the distance to the galaxy relative to the centre of mass of the cluster. Now let us multiply through by the vector $\mathbf{x}_i$. This gives (after some algebra):

$$\frac{1}{2}\frac{d^2}{dt^2}\left(m_i x_i^2\right) = \mathbf{F}_i \cdot \mathbf{x}_i + m_i\dot{x}_i^2 \tag{2.2}$$

where $x_i = |\mathbf{x}_i|$ and similar. Now let us sum over all of the galaxies:

$$\frac{1}{2}\frac{d^2 I}{dt^2} = 2T + V \tag{2.3}$$

where $I = \sum_i m_i x_i^2$ is the moment of inertia of the cluster, $T = \sum_i m_i \dot{x}_i^2$ is the total kinetic energy, and $V = \sum_i \mathbf{F}_i \cdot \mathbf{x}_i$ is the total potential energy of the cluster. For clusters in equilibrium, the second time derivative of the total moment of inertia I should be zero, and thus we derive the *scalar virial theorem*:

$$2T + V = 0 \tag{2.4}$$

[1] Actually, Zwicky found a number much lower than this because he didn't have the right Hubble constant. Why do you think the Hubble constant enters into the analysis?

[2] Zwicky called galaxies *nebulae*, which was the accepted terminology of the time.

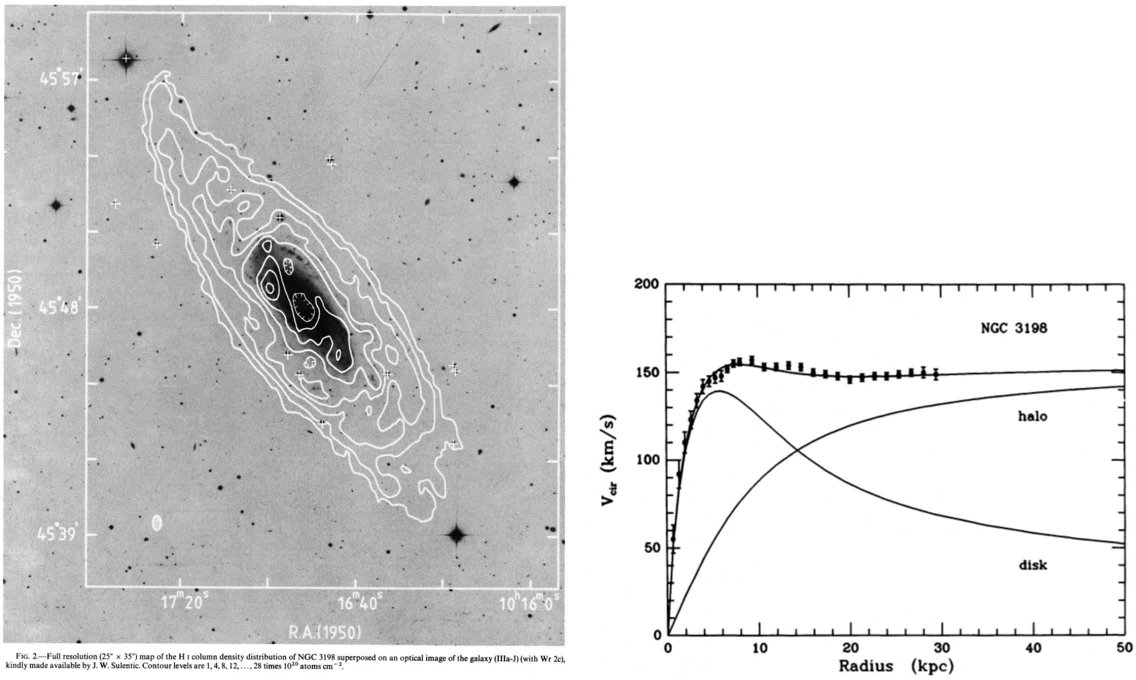


FIG. 2.—Full resolution (25″ × 35″) map of the H I column density distribution of NGC 3198 superposed on an optical image of the galaxy (IIIa-J) (with Wr 2c), kindly made available by J. W. Sulentic. Contour levels are 1, 4, 8, 12, . . . , 28 times 10^{20} atoms cm^{-2}.

Figure 2.2: NGC3198 viewed in HI (left) and its 'rotation curve' (right).

We may then crudely write the cluster 'kinetic energy' as $T = \frac{1}{2}M3\sigma^2$, where σ is the line of sight velocity dispersion: $\sigma^2 = \overline{v^2} - \overline{v}^2$ and M is the mass. Similarly, the potential energy can be estimated as $V = -GM^2/R$, where R is some scale that defines the rough 'size' of the cluster. (This is all slightly hand-wavy at this stage; we will return to mass modelling more carefully in later lectures.) Thus, from the virial theorem and the above estimates, the cluster mass is estimated as:

$$M \sim 3\sigma^2 R/G \tag{2.5}$$

Zwicky used data from Edwin Hubble for the doppler shifts of the galaxies in Coma to estimate $\sigma^2 \sim 1000\,\mathrm{km/s}$ (Hubble 1936). Using a cluster radius of $R \sim 1000\,\mathrm{kpc}$, we have: $M \sim 6 \times 10^{14}\,\mathrm{M}_\odot$ which is not too far off the modern value.

Thus, we arrive at Zwicky's puzzling result: there is apparently far more mass than light in the Coma cluster: 'dark matter' is born! Unfortunately, Zwicky had trouble convincing his colleagues of the importance of these findings and it was not until rotation curves of galaxies showed the same result some forty years later, that the idea of dark matter really took off. (Note that in Zwicky 1937, Zwicky also advanced the idea of using gravitational lensing to measure cluster masses. We will discuss this in detail later on, but Zwicky was clearly well ahead of his time!)

2.2 Galaxy rotation curves

Early evidence for missing matter in galaxies came from Babcock 1939, Volders 1959 and Freeman 1970. However, the evidence became irrefutable after later studies by Bosma (e.g. Bosma and van der Kruit 1979), Rubin (Rubin *et al.* 1980) and van Albada (van Albada *et al.* 1985) that collected much larger samples of galaxies. In Figure 2.2, I show the data from van Albada *et al.* 1985 for the galaxy NGC3198. Like Zwicky, these authors used kinematical tracers of galaxies to measure their total mass. However, disc galaxies are easier to mass model than galaxy clusters because they have large amounts of HI gas in a disc. This is useful for two reasons. Firstly, it is straightforward to measure the velocity of this HI gas by using the relative doppler shifts of the 21cm hydrogen line. Secondly, we can expect this gas to move on near-*circular orbits*, since this is the lowest energy state of the system. A simple proof of this follows using Lagrangian mechanics (a refresher of this is given in Appendix G). Assuming that the disc is axisymmetric, the classical Lagrangian is given by:

$$L = T - V = \frac{1}{2}m_i(\dot{R}_i^2 + R_i^2\dot{\phi}_i^2 + \dot{z}_i^2) - m_i\Phi(R_i, z_i) \tag{2.6}$$

where Φ is the axisymmetric gravitational potential and, as before, T is the kinetic and V is the potential energy. Application of the Euler-Lagrange equations:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial\dot{\phi}}\right) - \frac{\partial L}{\partial\phi} = 0 \tag{2.7}$$

then derives the familiar result that the z-component of the angular momentum is conserved:

$$\frac{d}{dt}\left(m_i R_i^2\dot{\phi}_i\right) = 0 = \frac{d}{dt}\left(m_i J_{z,i}\right) \tag{2.8}$$

where $J_{z,i}$ is the (conserved) specific z-angular momentum of an orbiting element i.

Now substituting $J_{z,i}$ into the energy equation, we have:

$$E = T + V = \frac{1}{2}m_i(\dot{R}_i^2 + \frac{J_{z,i}^2}{R_i^2} + \dot{z}_i^2) + m_i\Phi(R_i, z_i) \tag{2.9}$$

and it is now clear that E is minimised for $\dot{R}_i = 0$ and $\dot{z}_i = 0$: a planar circular orbit. QED.

Now a circular orbit has the nice property that, balancing the centripetal force and gravity, the mass is simply derived:

$$\frac{v_c^2}{R} = -\frac{\partial\Phi}{\partial R} \tag{2.10}$$

which for spherical symmetry[3] (using Newton's second theorem) gives a very simple form for the circular speed v_c as a function of R:

$$v_c^2 = \frac{GM(R)}{R} \tag{2.11}$$

where $M(R)$ is the mass enclosed within R.

Now, consider the circular speed, or 'rotation curve' shown in Figure 2.2. The data remain *flat* with radius R. From equation 2.11, $v_c^2 =$ const. implies that $M(R) \propto R$. But, the observed light distribution is falling off exponentially, with a scale length of just a few kiloparsecs! Thus, galaxies too must have a large amount of missing mass.

2.3 Beyond the classical evidence: gravitational lensing and cosmology

The above are the 'classical' evidences for dark matter: missing mass in galaxies and galaxy clusters. But if this were all the evidence that existed we should be skeptical. Perhaps these galaxies and clusters are not in equilibrium. Perhaps, the 'missing mass' is simply normal matter that is hard to see – cold gas, or faint stars. Perhaps, even, we have simply got gravity wrong on large scales. Each of these possibilities has been seriously explored over the past seventy years since missing mass was first discovered and we will discuss each in more detail in lecture 5.

To better understand the missing mass problem, it would be good to have genuinely independent probes of the mass distribution in the Universe. Luckily two are known: gravitational lensing, and cosmology. However, unlike the classical probes, above, that require only Newtonian dynamics, these latter two will require some understanding of Einstein's theory of gravity: general relativity. We discuss this necessary background in the next lecture (3). A treatment of cosmological probes, lensing, and a more detailed treatment of dynamics as a dark matter probe will then follow in 7, 4 and 11.

[3] Note that the assumption of spherical symmetry here may rightly be questioned. This is a disc galaxy after all! We will return to this when we perform more detailed mass modelling in later lectures.

Lecture 3

A brief primer on general relativity

Since many of you will have not covered (or only lightly covered) general relativity, I provide a very quick refresher here. I present the central concepts that lead to special and general relativity, derive the geodesic equation for GR and present the Einstein field equations. Finally, I discuss two solutions to the field equations: the Schwarzschild solution and the FLRW metric. I use the former to derive gravitational lensing – an important dark matter probe; the latter forms the basis of our current cosmological model.

3.1 What is wrong with good old Newton?

Newton himself understood that something is a bit fishy about Newtonian gravity. In a famous letter to Bently in 1693, Newton wrote[1]:

> *"It is inconceivable that inanimate brute matter should, without the mediation of something else which is not material, operate upon and affect other matter without mutual contact... That gravity should be innate, inherent, and essential to matter, so that one body may act upon another at a distance through a vacuum, without the mediation of anything else, by and through which their action and force may be conveyed from one to another, is to me so great an absurdity that I believe no man who has in philosophical matters a competent faculty of thinking can ever fall into it."*

However, the theory worked so spectacularly well that it was not until hundreds of years later – with the arrival of Albert Einstein[2] – that the worries returned. The problems can be simply understood:

1. Velocities are purely additive. Einstein understood that electromagnetism involves the speed of light. What happens then, he wondered, if we travel very rapidly? Is the speed of light some constant, plus the speed we are moving at? Einstein realised that such a theory would be unworkable because all velocities are *relative*. Without an absolute reference frame (and what would that be[3]?) we would be unable to even assign an unambiguous velocity to light. Physics would be ill-defined!

2. There are really *two* masses in Newtonian mechanics: inertial mass, and gravitational mass. It is truly remarkable that the two are identical as best we can tell (better than one part in 10^{11}; Will 1993). Surely this means something ...

3. Newtonian gravity implies instantaneous action at a distance. How do objects at the edge of the Universe know instantaneously that I'm jumping up and down?

[1] See e.g. `http://plato.stanford.edu/entries/newton-philosophy/`.

[2] If you have not read some of his original work, I can really recommend it (e.g. Einstein 1916). It is remarkable how little our pedagogical treatment of general relativity (at least for many physicists) has changed from Einstein's original exposition.

[3] Actually, the idea of an absolute reference frame was very popular in the late 1800's and Maxwell supposed that light travelled through an absolute *aether*. This appealing idea was, however, famously refuted by the Michaelson-Morley experiments in 1881 and 1887 (Michelson 1881;Michelson and Morley 1887).

The first point, as we shall see, leads us to Einstein's theory of special relativity. The second leads to general relativity. And the third, is something we will return to once armed with Einstein's general theory of relativity.

3.2 Special relativity

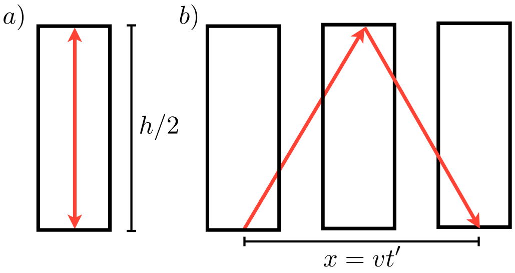


Figure 3.1: A schematic diagram of the 'light-clock' thought experiment. In panel a), the clock is stationary. The photon travels up and back in a time $t = h/c$, where c is the speed of light. In panel b), we watch the clock zoom past us at a speed v.

The first of the three considerations, above, led Einstein to assert that the speed of light *must be constant independent of the choice of inertial frame*[4]. This rather deep result leads to some remarkable conclusions. First, it implies that time must be relative. To arrive at this result, we can use a simple thought experiment: the light clock. Imagine I construct a clock so that in a time t a single photon of light travels upwards a distance $h/2$, bounces off a mirror, and travels back another distance $h/2$ to its original position. This is shown in Figure 3.1. In panel a), the clock is stationary. The photon travels up and back in a time $t = h/c$, where c is the speed of light. In panel b), we watch the clock zoom past us at a speed v. Since the speed of light is constant, to us the photon travel time is now:

$$t' = 2\frac{\left(\left(\frac{h}{2}\right)^2 + \left(\frac{v}{2}t'\right)^2\right)^{1/2}}{c} \tag{3.1}$$

where we now use t' to indicate that we are observing the clock from a different *inertial frame*. Rearranging, and after some simple algebra, the above gives:

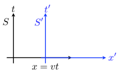
$$\frac{t'}{t} = \left(1 - \frac{v^2}{c^2}\right)^{-1/2} = \gamma \tag{3.2}$$

which defines the Lorentz factor γ.

For $v \ll c$, the above equation has almost no effect. But as we approach the speed of light, $t' > t$ and time becomes heavily *dilated*: moving clocks run slow!

The above derives a pure time transformation. But in general, we can transform the position coordinates between inertial frames too and the speed of light must remain invariant also in such situations. A general position transformation from a frame S to a frame S' can be written:

$$x' = a_1 x + a_2 y + a_3 z + a_4 ct + a_5 \tag{3.3}$$



Now, suppose that in the frame S', S is moving at speed a v along the x-axis. We may then define without loss of generality x' such that $x' = 0$ at $x = vt$ (see the margin figure). This gives:

$$x' = \gamma(x - vt) \qquad ; \qquad x = \gamma(x' + vt') \tag{3.4}$$

where the right equation follows by symmetry between the frames (in the frame S, S' moves at speed $-v$). In the Newtonian world view, we would then assert that $t = t'$ which derives the *Galilean* transformation, $\gamma = 1$. However, in special relativity, we have instead that c is a constant. Imagine, then, that we move a distance $x = ct$ in frame S. This must then correspond to $x' = ct'$ in frame S'. Substituting these relations into equations 3.4, we have:

$$ct' = \gamma(ct - vt) \qquad ; \qquad ct = \gamma(ct' + vt') \tag{3.5}$$

and we recover γ as in equation 3.2. Thus, we derive the full *Lorentz* transforms:

$$x' = \gamma\left(x - \frac{v}{c}ct\right) \tag{3.6}$$

[4] An inertial frame is one that experiences no accelerations.

$$ct' = \gamma\left(ct - \frac{v}{c}x\right) \tag{3.7}$$

where we have deliberately used the speed of light to give the time and position coordinates the same dimensions. We now see an important key result: the quantity:

$$ds^2 = c^2dt'^2 - dx'^2 = c^2dt^2 - dx^2 \tag{3.8}$$

is invariant. This is the fundamental 'length' – the Lorentz invariant – in special relativity. (Note that we have assumed up to now that $dy = dz = 0$. Putting these back in, the above generalises to: $ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2$.)

3.2.1 Introducing tensor notation

The key concept in special (and general) relativity is that physics must be independent of our choice of coordinate system – however crazy. To facilitate this, it is helpful to devise a mathematical framework that allows for arbitrary transformations, while maintaining physical properties like the Lorentz invariant. Consider the following "4-vector" position:

$$x^\mu = (cdt, x, y, z) \qquad ; \qquad \mu = 0, 1, 2, 3 \tag{3.9}$$

where we have used the speed of light, c, to make the time coordinates have the same dimensions of length as the other coordinates[5].

Now let us define some mathematics that returns the correct 'length' (the Lorentz invariant) when we take the product of this 4-vector with itself – independent of any coordinate transformation. To achieve this, let us first define the self product as:

$$x^\mu x_\mu = x^\mu g_{\mu\nu} x^\nu \tag{3.10}$$

where repeated indices are summed over, and we define the *metric* $g_{\mu\nu}$ as the object that transforms the *contravariant* form of x^μ to the *covariant*[6] form x_μ. In special relativity, the self product must produce the Lorenz invariant (equation 3.8). Thus:

$$ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2 = dx^\mu dx_\mu \tag{3.11}$$

which derives the metric for special relativity $g_{\mu\nu} = \eta_{\mu\nu} = \mathrm{diag}(1, -1, -1, -1)$. This is called *Minkowski spacetime.* As we will see later, the metric takes different forms in the presence of gravitational fields.

The Lorentz invariant (equation 3.8) also gives us the transformation laws that co- and contravariant 4-vectors must obey:

$$x'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} x^\nu \tag{3.12}$$

$$x'_\mu = \frac{\partial x^\nu}{\partial x'^\mu} x_\nu \tag{3.13}$$

where the above simply ensures that $x'^\mu x'_\mu = x^\mu x_\mu = \text{const.}$ by construction.

In special relativity, the coordinate transformation is defined by $\frac{\partial x'^\mu}{\partial x^\nu}$, often written as $\Lambda^\mu{}_\nu$, which is simply a matrix that defines the Lorentz transform (it is derived from the partial derivatives of the transformation equations 3.6 and 3.7):

$$\Lambda = \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3.14}$$

We may generalise such 4-vectors to higher dimensional beasts: *tensors*. Tensors – like $g_{\mu\nu}$ – are matrices that must obey transformation relations that ensure their effect is coordinate invariant:

[5] Note that c is only really required here because we use metres to measure distance, and seconds to measure time. We could instead adopt units where $c = 1$ and indeed this is often done in relativity books.

[6] A useful way to remember which is which is: "co- goes below".

$$A'^{\mu\nu} = \frac{\partial x'^{\mu}}{\partial x^{\alpha}} \frac{\partial x'^{\nu}}{\partial x^{\beta}} A^{\alpha\beta} \tag{3.15}$$

Tensors also have co- and contravariant (and now mixed) forms. It is straightforward to show that the above transformation rule ensures that the effect of $A'^{\mu\nu}$ acting on a 4-vector is coordinate independent.

Note that if we allow the metric to act on a tensor, we can lower the *dimensionality* of the tensor – a process call *contraction*:

$$g_{\mu\nu} A^{\mu\nu} = A_{\mu}{}^{\mu} = A \tag{3.16}$$

where A is a scalar contraction of the tensor $A^{\mu\nu}$.

The above is all very nice, but one thing remains a bit tricky. Suppose I want to take the *derivative* of a tensor. A first guess at a useful derivative operator would be the 4-derivative:

$$\partial_{\mu} = \frac{\partial}{\partial x^{\mu}} \equiv \left(\frac{\partial}{\partial ct}, \nabla \right) \tag{3.17}$$

However, if I were to use the 4-derivative then it is straightforward to show that I produce an object that is no longer a tensor. Writing:

$$Y_{a}{}^{b} = \partial_{a} X^{b} \tag{3.18}$$

I can then apply a coordinate transformation to X^b:

$$\begin{aligned} Y'^{\ b}_{a} &= \frac{\partial}{\partial x'^{a}} \left(\frac{\partial x'^{b}}{\partial x^{\nu}} X^{\nu} \right) \\ &= \frac{\partial x^{d}}{\partial x'^{a}} \frac{\partial}{\partial x^{d}} \left(\frac{\partial x'^{b}}{\partial x^{\nu}} X^{\nu} \right) \\ &= \frac{\partial x^{d}}{\partial x'^{a}} \frac{\partial x'^{b}}{\partial x^{\nu}} \partial_{d} X^{\nu} + \frac{\partial x^{d}}{\partial x'^{a}} \frac{\partial^{2} x'^{b}}{\partial x^{d} \partial x^{\nu}} X^{\nu} \end{aligned} \tag{3.19}$$

The first term on the left is the tensor transformation rule for $Y_{a}{}^{b}$, but the term on the right is an extra piece. Thus, we have proven that $Y_{a}{}^{b}$ is *not* a tensor. This is bad because we have developed this whole mathematical machinery in order to describe physics in a coordinate independent manner. We must therefore hunt for a derivative operator that produces tensors from tensors. There is more than one operator that can achieve this. Here, however, we will need only the *covariant* derivative operator:

$$\nabla_{c} X^{a} = \partial_{c} X^{a} + \Gamma^{a}_{bc} X^{b} \tag{3.20}$$

where $\Gamma^{a}_{bc} = \frac{\partial x'^{a}}{\partial x^{\alpha}} \frac{\partial^{2} x^{\alpha}}{\partial x'^{b} \partial x'^{c}}$ is called a *Christoffel symbol*.

It is straightforward to show that the addition of the Christoffel symbol here negates the effect of the extra piece we derived above for the non-tensorial 4-derivative and thus that the covariant derivative does indeed produce tensors from tensors (note that when deriving this result you must remember to also transform the Christoffel symbol).

The above mathematical trickery is useful. If we can phrase our physics in terms of such tensors and 4-vectors, then we will be coordinate independent by construction.

3.2.2 4-momentum, 4-force and all that ...

Although special relativity deals only with inertial frames, we can still happily watch other people accelerate. Thus, defining a four-momentum and four-force is still meaningful. Let us start by defining the proper time from the Lorentz invariant proper distance:

$$ds^{2} = c^{2} d\tau^{2} \tag{3.21}$$

Since the proper distance ds and the speed of light c are invariant, then the proper time $d\tau$ must be also. This is useful as it suggests that we can form invariant time derivatives of the 4-position by using τ. Thus suggests the following definition of 4-momentum:

$$P^\mu = m\frac{dx^\mu}{d\tau} \tag{3.22}$$

where m is an invariant mass – the 'rest mass' of a particle. It is straightforward to show that $\Delta P^\mu = 0$ then gives momentum-energy conservation laws that reduce to classical energy and momentum conservation in the low velocity limit. This suggests that the above choice is the right one. Similarly, we may then define a four-force:

$$F^\mu = \frac{dP^\mu}{d\tau} \tag{3.23}$$

which tends to the more usual 3-force in the non-relativistic limit.

The above equations form the basis of special relativistic dynamics that hopefully you have encountered before.

3.2.3 The clock hypothesis and general relativity

So far, we have discussed only how to deal with *inertial frames* that are not accelerating. How to deal with accelerations can be understood from the famous clock, or twin paradox. Imagine I have two twins on Earth. One flies away from the other for a time $t/2$ at a speed v, turns around, and then comes back at a speed $-v$. The twin on Earth sees his brother's clock run slow such that the total time elapsed is:

$$t' = \gamma(v)t/2 + \gamma(-v)t/2 = \gamma t \tag{3.24}$$

But now consider things from the view of the rocket-twin. Surely he sees *his* brother's clock running slow too such that $t = \gamma t'$! This is the 'paradox'. The solution, of course, is that the rocket-twin must accelerate to come back to Earth. And accelerations are not described in special relativity. Thus, the apparent symmetry of the problem is broken. On the other hand, however, we may assert that the Earth-twin *must* have the answer right since he does not accelerate and therefore special relativity in his frame is just fine: this is the *clock hypothesis* and is the basis for general relativity.

Let us take the above idea a little further. Suppose, then, we define a frame in which there are no accelerations. In such a frame special relativity must apply and the double proper time derivative of our 4-vector position (the 4-acceleration) must be zero:

$$\frac{d^2\epsilon^\mu}{d\tau^2} = 0; \qquad \epsilon^\mu = (ct, x, y, z) \tag{3.25}$$

Furthermore spacetime must be Minkowski:

$$c^2 d\tau^2 = \eta_{\alpha\beta} d\epsilon^\alpha d\epsilon^\beta \tag{3.26}$$

with $\eta_{\alpha\beta} = \text{diag}(1, -1, -1, -1)$.

Now, we can describe motion in *any* frame by simply transforming away from the above one. Inserting our general coordinate transform (equation 3.12) into equation 3.25, we obtain the general relativistic dynamics equations [exercise]:

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{\alpha\beta}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = 0 \tag{3.27}$$

with the metric equation:

$$c^2 d\tau^2 = g_{\alpha\beta} dx^\alpha dx^\beta \tag{3.28}$$

where the Christoffel symbols $\Gamma^\mu_{\alpha\beta}$ and new metric $g_{\alpha\beta}$ are defined by the transformation coefficients:

$$\Gamma^\mu_{\alpha\beta} = \frac{\partial x^\mu}{\partial \epsilon^\nu}\frac{\partial^2 \epsilon^\nu}{\partial x^\alpha \partial x^\beta} \tag{3.29}$$

$$g_{\alpha\beta} = \frac{\partial \epsilon^\mu}{\partial x^\alpha}\frac{\partial \epsilon^\nu}{\partial x^\beta}\eta_{\mu\nu} \tag{3.30}$$

And finally, it is possible to substitute for the metric inside the the Christoffel symbols, thus demonstrating that everything can be described purely by the metric:

$$\Gamma^{\alpha}_{\lambda\mu} = \frac{1}{2} g^{\alpha\nu} \left(\frac{\partial g_{\mu\nu}}{\partial x^{\lambda}} + \frac{\partial g_{\lambda\nu}}{\partial x^{\mu}} - \frac{\partial g_{\mu\lambda}}{\partial x^{\nu}} \right) \tag{3.31}$$

The metric itself simply describes how to define a length in some arbitrarily hideous spacetime. We can think of it as describing *curvature* and often people talk of curved spacetime in GR. But this is just one interpretation of the mathematics – it does not necessarily mean that spacetime is really curved.

3.2.4 The equivalence principle

We are now in a position to return to the troubling aspects of Newtonian gravity we set out at the start in §3.1. So long as I can keep transforming to a frame where there are no accelerations, to what extent can I be said to really be feeling any force? This insight led to Einstein's equivalence principle, which states that:

> *"Freely falling (infinitesimal) observers experience no gravitational effects – i.e. they can just keep transforming to Minkowski space."*

Or to put it another way – it is the *tidal gravitational field* that you feel, or the force caused by hitting something that gets in the way of your being a freely falling observer, not gravity per-se. The only gravity you feel is when bits of your body attempt to be freely falling observers that fall in different directions (ouch!). This is *very* different from the Newtonian world view. The even stronger version of the above is the *strong equivalence principle* that goes further:

> *"All laws of physics for freely falling observers are identical to those in the absence of gravity."*

The above has some profound implications. First, in order to be true the gravitational and inertial masses *must* be identical. This is because we require that the acceleration of our inertial frame be *exactly* equal to the gravitational acceleration (otherwise we cannot simply transform our frame and remove the force):

$$\mathbf{a} = \frac{m_G}{m_I} \mathbf{g} \tag{3.32}$$

where $\mathbf{g}$ is the gravitational acceleration, m_I is the inertial mass, and m_G is the gravitational mass. Thus we must have $m_I/m_G = 1$. And this then has a further implication:

> *"An observer in a windowless room cannot distinguish between being on the surface of the Earth, and being in a spaceship accelerating at 1g."*

3.2.5 The field equations

So far, we have derived the general relativistic equivalent of $\mathbf{F} = m\mathbf{a}$. To solve real problems, we must have a method to determine the accelerations and that means having a *field theory*. In classical Newtonian mechanics, gravity is a scalar field described by Poisson's equation, for example. What are the field equations in GR?

Up to now, our treatment has been completely general. We have simply demanded that physics remain invariant under transformations between inertial frames, and that special relativity apply in non-accelerating frames. Now, we enter a fuzzier and less satisfying realm where we have some choice in how to proceed. Given such choice, we will aim for the very simplest field equations possible. We will be guided by noting the following:

1. We may expect that the source of gravity should look something like (somewhat like in Newton's laws) mass.
2. We know that at least in classical mechanics mass, momentum and energy are conserved. In special relativity, this becomes the conservation of 4-momentum.

3. We know that we must find some *tensor* in these quantities to ensure coordinate invariance.

Suppose we just write down a tensor whose covariant derivative is nothing more than the 4-conservation laws for each component of the 4-momentum. This is the *energy-momentum* tensor $T^{\mu\nu}$ which satisfies:

$$\nabla_\nu T^{\mu\nu} = 0 \tag{3.33}$$

where T^{00} is the energy density, T^{12} is the x-component of the y-momentum current, etc., and the tensor must be *symmetric*. Since $T^{\mu\nu}$ is really defined only by a derivative (by conservation laws), its precise form must depend on the type of matter being considered. For a perfect fluid, for example, in the *rest frame*, $T^{\mu\nu}$ takes on a simple form:

$$T^{\mu\nu} = \mathrm{diag}(c^2\rho, p, p, p) \tag{3.34}$$

where ρ and p are the proper density and pressure, respectively. (Recall that pressure is just the flux of x-momentum in the x-direction[7].) Transforming to a general frame, this gives:

$$T^{\mu\nu} = \left(\rho + p/c^2\right) U^\mu U^\nu - p g^{\mu\nu} \tag{3.35}$$

where $U^\mu = \frac{dx^\mu}{d\tau}$ is the 4-velocity. (To use the above we must also specify an equation of state that links the pressure and density.)

We now have an equation for the *source terms* of our field, but not how to describe the response – how spacetime will 'curve' (i.e. what the metric will be) in response these sources. Again, we may be guided by some simple considerations:

1. We must construct something using the metric. However, we know from our dynamics equation 3.27 that simple coordinate transforms are described by *first derivatives* of the metric (these appear in the Christoffel symbols). Thus, we must use at least *second derivatives* of the metric if we want to describe physical spacetime distortions that cannot be simply transformed away.

2. As above, we must look for a tensor to ensure coordinate invariance for our field theory. Since we will equate it to the energy-momentum tensor, this should be a second rank tensor.

3. If our tensor – let us call it $G^{\mu\nu}$ – has a vanishing covariant derivative ($\nabla_\nu G^{\mu\nu} = 0$) then it must follow that $G^{\mu\nu} = kT^{\mu\nu}$, where k is some constant. Thus, we should hunt for something that has a vanishing covariant derivative.

It turns out that there is only one tensor that can be constructed that is linear in the second derivatives of the metric: the *Riemann tensor*:

$$R^a_{bcd} = \partial_c \Gamma^a_{bd} - \partial_d \Gamma^a_{bc} + \Gamma^e_{bd}\Gamma^a_{ec} - \Gamma^e_{bc}\Gamma^a_{ed} \tag{3.36}$$

which is a function only of the Christoffel symbols, and therefore a function of the metric and its derivatives (c.f. equation 3.30).

The Riemann tensor is a fourth rank tensor, but can be contracted using the metric to form a second rank tensor (the Ricci tensor), or contracted further to form a scalar. There is only one second rank tensor that can be constructed from the Riemann tensor and its contractions that has a vanishing covariant derivative: the Einstein tensor:

$$G^{\mu\nu} = R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R \tag{3.37}$$

where $R^{\mu\nu}$ is the Ricci tensor, and $R = g^{\mu\nu}R_{\mu\nu}$ is the curvature scalar.

And thus, we arrive at the Einstein field equations:

$$G^{\mu\nu} = kT^{\mu\nu} \tag{3.38}$$

[7] In case this is not clear, recall that pressure is just the force per unit area perpendicular to a surface. In the x-direction, for example, we may write: $P = \mathbf{F} \cdot \hat{\mathbf{x}}/A = \frac{d}{dt}(m\dot{\mathbf{v}}) \cdot \hat{\mathbf{x}}/A$, which is then clearly the flow of x-momentum per unit area along x (i.e. a momentum-current).

where the constant of proportionality k is determined by demanding that the field equations reproduce Newton's laws in the weak field limit (we will come to this shortly).

The above 'derivation' is rather sketchy and relies a bit on you having encountered this all before. But the sketched 'derivation' highlights an important point: general relativistic dynamics is on quite secure ground; the field equations are not. This will become important when we discuss alternative gravity theories in later lectures. To give you an idea now, though, of the remaining freedoms, consider the following. There is another tensor that should be familiar to you already whose covariant derivative is zero: the metric itself: $\nabla_\nu g^{\mu\nu} = 0$. Thus, we can generalise the field equations further to:

$$G^{\mu\nu} + \Lambda g^{\mu\nu} = kT^{\mu\nu} \tag{3.39}$$

where Λ is known as the *cosmological constant*.

Einstein himself was the first to propose adding the cosmological constant in order to create solutions where the Universe is static. He later called this his "greatest blunder"[8]. Lemaitre – following on from theoretical work by de Sitter and Friedmann – went on to demonstrate that the Universe is in fact expanding (Nussbaumer and Bieri 2011). But the cosmological constant has returned with recent observations that suggest that the Universe is *accelerating*: a phenomenon that could be explained by said cosmological constant (often referred to as 'dark energy'; of which more later).

To understand what the cosmological constant means physically, let's move it over to the right side of the field equations and add it to the energy-momentum tensor as if it were an additional source term:

$$T^{\mu\nu}_\Lambda = \frac{\Lambda}{k} g^{\mu\nu} \tag{3.40}$$

In the absence of any matter, $T^{\mu\nu} = 0$ and the above must represent the source terms coming from the vacuum itself. If that sounds strange, later on the in the course we will explain why there may be no such thing as genuinely 'empty' space. For now, consider what the above means dynamically. We are free to transform the metric into Minkowski space: $g^{\mu\nu} = \eta^{\mu\nu} = \mathrm{diag}(1, -1, -1, -1)$, which gives:

$$T^{\mu\nu}_\Lambda = \frac{\Lambda}{k}\mathrm{diag}(1, -1, -1, -1) = \mathrm{diag}(c^2\rho_{\rm vac}, p_{\rm vac}, p_{\rm vac}, p_{\rm vac}) \tag{3.41}$$

and if the above really represents the vacuum solution, then this Minkowski space solution must be just fine: all observers will see the same vacuum. The energy density of the vacuum must be encoded in $T^{00}_\Lambda = c^2\rho_{\rm vac}$ and thus we have derived that the vacuum pressure is *negative*:

$$p_{\rm vac} = -c^2\rho_{\rm vac} \tag{3.42}$$

Thus the vacuum – assuming such a thing exists – will behave like antigravity pushing the Universe apart. This is why the cosmological constant has been called 'dark energy' and evoked to explain the observed accelerating expansion of the Universe.

3.2.6 Energy conservation in general relativity

This is often a source of confusion and you will often hear (particularly when adding the cosmological constant) that energy is not conserved in general relativity. In fact, the issue is somewhat subtle. For starters, what *is* clearly conserved by construction is the energy-momentum tensor, and that is the fundamental coordinate invariant tensor that should be conserved. Furthermore, classical energy conservation is recovered in the weak-field limit. What is less clear is whether a scalar quantity like energy can be said to be conserved in general. Being a scalar, the energy is, of course, coordinate dependent and definition dependent and so you can arrive at the conclusion that 'energy' is conserved or not, depending on how you define it! Such issues should already be familiar to you by now from special relativity where it is the energy-momentum 4-vector that is conserved. In special relativity, energy is only conserved for inertial observers watching from a fixed frame. In general relativity, no observer can sit happily in a fixed frame anymore, and so a simple unambiguous scalar energy can no longer be defined.

[8] I believe the origin of this quote is George Gamow's autobiography: *"My World Line"*.

3.3 Solving the field equations

We now have the equations of motion and the field equations: in principle we are all set. In practice, finding solutions to the field equations is hard. Not least because coordinate transformations can fool us into thinking that two solutions are different when really they are the same but simply transformed!

Here, we will consider first weak field solutions to the equations that can be solved using perturbation theory. We then present two full solutions of interest for this course: the Schwarzshchild solution that is the relativistic equivalent of a point mass (and also happens to describe black holes), and the Friedmann, Lamaitre, Robertson, Walker (FLRW) metric that describes an infinite homogeneous and isotropic Universe which forms the backbone of our current cosmological model.

3.3.1 The Newtonian weak field limit

First let us consider 'Newtonian' weak field general relativity. This is defined by three things:

1. Objects move slowly, i.e. $\frac{1}{c}\frac{dx^\mu}{d\tau} \ll \frac{dt}{d\tau}$.
2. Gravity is weak such that spacetime is very close to Minkowski. In this case, we may write the metric as Minkowski plus some small perturbation, $h_{\mu\nu}$:

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \tag{3.43}$$

3. The metric is static and not a function of time.

From the first condition, the GR dynamics equation (3.27) becomes:

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{00}c^2\left(\frac{dt}{d\tau}\right)^2 = 0 \tag{3.44}$$

which, substituting our perturbed metric (equation 3.43) into the Christoffel symbols using equation 3.30 gives:

$$\frac{d^2\mathbf{x}}{dt^2} = -\frac{c^2}{2}\nabla h_{00} \quad ; \quad \frac{d^2t}{d\tau^2} = 0 \tag{3.45}$$

The term on the right tells us that in the Newtonian weak field limit, time dilation effects must be constant. But more interesting is the term on the left. This immediately tells us the meaning of the h_{00} term in the metric. It must correspond to the standard Newtonian gravitational potential as follows: $h_{00} = \frac{2\Phi}{c^2}$, thus recovering the familiar Newtonian dynamics equations. (Note that we are not specifying what the metric is here, but rather interpreting what it must mean.)

Furthermore, we may solve the Einstein field equations (3.38) for our perturbed metric. Using again that $h_{00} = \frac{2\Phi}{c^2}$, the left hand side reduces to:

$$R_{00} - \frac{1}{2}Rg_{00} = \frac{2\nabla^2\Phi}{c^2} \tag{3.46}$$

while the right hand side gives:

$$\frac{8\pi GT_{00}}{c^4} = \frac{8\pi G\rho}{c^2} \tag{3.47}$$

and thus we recover the familiar Poisson equation: $\nabla^2\Phi = 4\pi G\rho$. (This derives the constant of proportionality $k = 8\pi G/c^4$ in the Einstein field equations.) Note that some significant algebra is required to achieve the above results, and I would advise you to consult a good textbook on GR if embarking on the above derivations for the first time.

3.3.2 The weak field limit & gravitational waves

The Newtonian weak field limit helps us understand the connection between GR and Newton, but does not give us intuition as to how these two theories differ. Here, we consider instead a linear expansion of the GR field equations. In this case, we assume only that the field is weak such that the metric can be decomposed into Minkowski plus a small perturbation:

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \tag{3.48}$$

But we no longer assume slow moving particles or static fields. In this case, the field equations become (to linear order and in the 'harmonic gauge'):

$$\Box h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}\Box h = -\frac{16\pi G}{c^4}T_{\mu\nu} \tag{3.49}$$

where $\Box = \partial_{ct}^2 - \partial_x^2 - \partial_y^2 - \partial_z^2$ is called the *d'Alembert operator*.

If we consider then a vacuum ($T_{\mu\nu} = 0$), and define the 'trace-reversed' perturbation: $\overline{h}_{\mu\nu} = h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h$, then we have that:

$$\Box\overline{h}_{\mu\nu} = 0 \tag{3.50}$$

which is a wave equation! Thus, in general relativity – unlike in Newtonian gravity – gravitational perturbations drive gravitational waves through the vacuum. This solves the final of our our original problems with Newtonian gravity: instantaneous action at a distance[9]. In GR information about the gravitational tidal field is transmitted by gravitational waves at the speed of light.

3.3.3 The Schwarzschild solution

The *Schwarzschild* solution to the field equations was found within just a year of Einstein completing general relativity (Schwarzschild 1916), and penned while Schwarzschild was serving in the army during world war one[10]. It is a tragedy that he did not survive to contribute more to the field. The solution is, in fact, the *only* spherically symmetric vacuum solution to the Einstein field equations (i.e. with $T^{\mu\nu} = 0$) and has the following form:

$$c^2d\tau^2 = \left(1 - \frac{2GM}{c^2r}\right)c^2dt^2 - \frac{dr^2}{(1 - \frac{2GM}{c^2r})} - r^2(d\theta^2 + \sin^2\theta d\phi^2) \tag{3.51}$$

where r, θ, ϕ are the familiar spherical polar coordinates (in this context *Schwarzschild coordinates*). It is straightforward to verify that the above metric is indeed a solution of the field equations [exercise].

To understand the physical meaning of the Schwarzschild solution, it is instructive to consider the Newtonian weak field. As previously, we have:

$$\begin{aligned} \ddot{\mathbf{x}} &= -\frac{c^2}{2}\nabla g_{00} \\ &= -\frac{c^2}{2}\frac{2GM}{c^2r^2}\hat{\mathbf{r}} \\ &= -\frac{GM}{r^2}\hat{\mathbf{r}} \end{aligned} \tag{3.52}$$

where $\hat{\mathbf{r}}$ is a unit vector in the radial direction. Thus the Schwarzschild metric approaches the Newtonian solution for a point mass of mass M in the weak field limit. This is why it is often thought of as the GR equivalent of a point mass.

[9] See e.g. the excellent lecture notes on GR by Sean Carroll: `http://preposterousuniverse.com/grnotes/`.

[10] A quick search on the NASA Astronomy Abstract Service finds that this article has just 17 citations to date! Let this be a lesson, then, to us all that citations are not everything. It is also interesting to note that Schwarzschild was over 40 when he produced his most famous work. That busts another popular myth about science and age.

3.3.4 The FLRW metric and the cosmological model

Above, we wrote down the Scchwarzschild solution to the field equations that approaches a Newtonian point mass in the weak field limit. Here, we are interested in finding a solution that can describe the whole Universe. Out task is made significantly easier by the fact that observations of galaxies in the distant Universe (Wu *et al.* 1999; Yadav *et al.* 2005), and the cosmic microwave background radiation (the afterglow of the Big Bang – more on this later), suggest that the Universe is very close to being perfectly isotropic and homogeneous. Furthermore, it would be quite the coincidence if it appeared this way just from our perspective. Thus, it is reasonable to hunt for a Universe-metric that describes isotropic and homogeneous matter. As you will see on the problem sheet, such a metric due to its symmetries (rather like the Scchwarzschild metric) is unique. It is typically called the Friedmann, Lamaitre, Robertson, Walker (FLRW) metric after the various authors who discovered it:

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t)\left(\frac{dr^2}{1-kr^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2)\right) \tag{3.53}$$

where $R(t)$ is called the *scale factor*, and k is a parameter that measures the fundamental *curvature* of the spacetime. We will study dynamics in this metric further when we discuss cosmological probes of dark matter in lecture 7.

Lecture 4

Gravitational lensing basics

Here we derive the basic equations for our second non-classical dark matter probe: gravitational lensing.

4.1 Some history

Gravitational lensing is a key dark matter probe and one of the motivations for refreshing our knowledge of GR. Gravitational lensing was first proposed by Chwolson 1924, but the first real calculation came from Einstein 1936. At that time, with the exception of the incredible foresight of Zwicky 1937, it was largely though that such lensing effects would be unobservable. Only much later did Refsdal 1964 lay the groundwork for a modern theory of gravitational lensing. Refsdal's work lay largely unrecognised for years until suddenly in 1979, the first gravitational lens was actually discovered by Walsh *et al.* 1979. A modern image of this lens – Q0957+561 – is given in Figure 4.1, left panel. The right panel shows citations to Refsdal's pioneering lensing paper. Notice the sudden explosion after 1979 – lensing as a distinct research field was born.

4.2 The bending angle for a Schwarzschild lens

In this section, we will use the Scchwarzschild solution to derive gravitational lensing – one of our key dark matter probes. To do that, we need to solve for the dynamics in the Schwarzschild metric. One way to achieve this is to simply plug the metric in to the GR dynamics equation (3.27), but this is not the most elegant. An alternative way to derive GR dynamics is through a *principle of least action*. Recall that the action S is defined by the path integral:

$$S = \int_{\mathbf{x}_1,\lambda_1}^{\mathbf{x}_2,\lambda_2} L(\mathbf{x},\dot{\mathbf{x}},\lambda)d\lambda \tag{4.1}$$

where L is called the *Lagrangian* and λ is an *affine parameter* that describes the motion along the path (you can think of it as the proper time τ, and indeed it is related via a simple transformation $\lambda = a\tau + b$). For a suitably defined Lagrangian, the dynamics follow from extremising the action $\delta S = 0$, which derives the Euler Lagrange equations (see Appendix G):

$$\frac{d}{d\lambda}\left(\frac{\partial L}{\partial \dot{x}^\mu}\right) - \frac{\partial L}{\partial x^\mu} = 0 \tag{4.2}$$

where $\dot{x}^\mu = \frac{dx^\mu}{d\lambda}$ now refers to derivatives with respect to the affine parameter λ. Test particles that move along extremum paths are said to follow *geodesics*.

Since we are extremising the *path length*, a GR Lagrangian immediately suggests itself. Notice that we can write:

$$S = \int \frac{ds}{d\lambda} d\lambda \tag{4.3}$$

where ds is the proper length. Comparing this with equation 4.1, we derive a Lagrangian F:

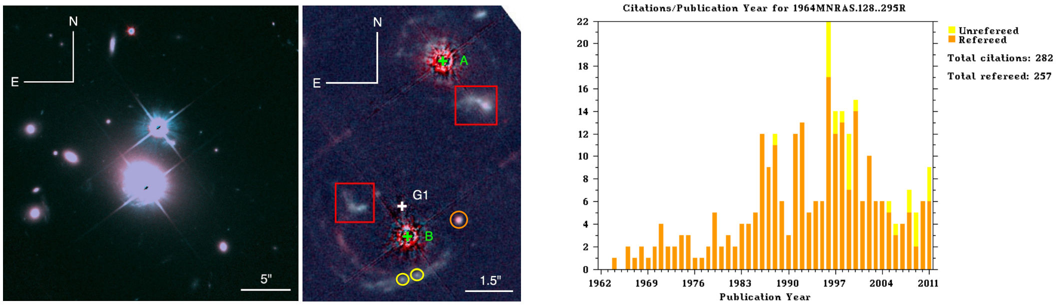


Figure 4.1: **Left & Middle:** A modern image of the very first gravitational lens discovered in the Universe: Q0957+561. The images A and B mark the two main quasar images. The yellow circles mark additional faint images that have recently been found. The red squares mark probable complex lensing features. The orange circle marks what is most likely a faint foreground Milky Way halo star. The cross, G1, marks the galaxy lens. **Right:** A history of citations to Refsdal's seminal paper on gravitational lensing. Notice how interest suddenly shot up right after the discovery of the first lens in the Universe in 1979.

$$F = \frac{ds}{d\lambda} = \sqrt{g_{\mu\nu}\frac{dx^{\mu}}{d\lambda}\frac{dx^{\nu}}{d\lambda}} \tag{4.4}$$

however, this is difficult to work with because of the square root. An extremely useful trick is to note that:

$$\delta\int_{\mathbf{x}_1,\lambda_1}^{\mathbf{x}_2,\lambda_2} F^2 d\lambda = \delta\int_{\mathbf{x}_1,\lambda_1}^{\mathbf{x}_2,\lambda_2} 2F d\lambda \tag{4.5}$$

Thus, we may work just as well with $L = \frac{1}{2}F^2$ and avoid the square root. Using special relativistic 'Cartesian' coordinates $x^\mu = (ct, x, y, z)$, it is straightforward to show that application of the Euler Lagrange equations (4.2) using the above Lagrangian recovers the GR dynamics equation (3.27) – also known as the *geodesic equation*. More useful, however, is that we can now directly derive the dynamics for the Schwarzschild metric using spherical polar coordinates $x^\mu = (ct, r, \theta, \phi)$ and the Euler Lagrange equations. The Lagrangian is given from the metric by:

$$L = \frac{1}{2}\left[\left(1 - \frac{2GM}{c^2 r}\right)c^2\dot{t}^2 - \frac{\dot{r}^2}{(1-\frac{2GM}{c^2 r})} - r^2(\dot{\theta}^2 + \sin^2\theta\dot{\phi}^2)\right] \tag{4.6}$$

and the Euler Lagrange equations then give the dynamics as:

$$\frac{d}{d\lambda}\left(\kappa c\dot{t}\right) = 0 \tag{4.7}$$

$$\frac{d}{d\lambda}\left(\frac{\dot{r}}{\kappa}\right) - \frac{GM}{r^2}\dot{t}^2 - \frac{1}{\kappa^2}\frac{GM}{c^2 r^2}\dot{r}^2 + 2r\left(\dot{\theta}^2 + \sin^2\theta\dot{\phi}^2\right) = 0 \tag{4.8}$$

$$\frac{d}{d\lambda}\left(r^2\dot{\theta}\right) - \sin\theta\cos\theta\dot{\phi}^2 = 0 \tag{4.9}$$

$$\frac{d}{d\lambda}\left(r^2\sin^2\theta\dot{\phi}\right) = 0 \tag{4.10}$$

where $\kappa = \left(1 - \frac{2GM}{c^2 r}\right)$.

Now, the symmetry of the metric suggests that we can restrict motion to the equatorial plane for which $\theta = \pi/2$ and $\dot{\theta} = 0$. Equation 4.10 then gives us the familiar specific angular momentum conservation around the z-axis (our assumption that the orbit is planar implies the other two components of specific angular momentum must also be conserved):

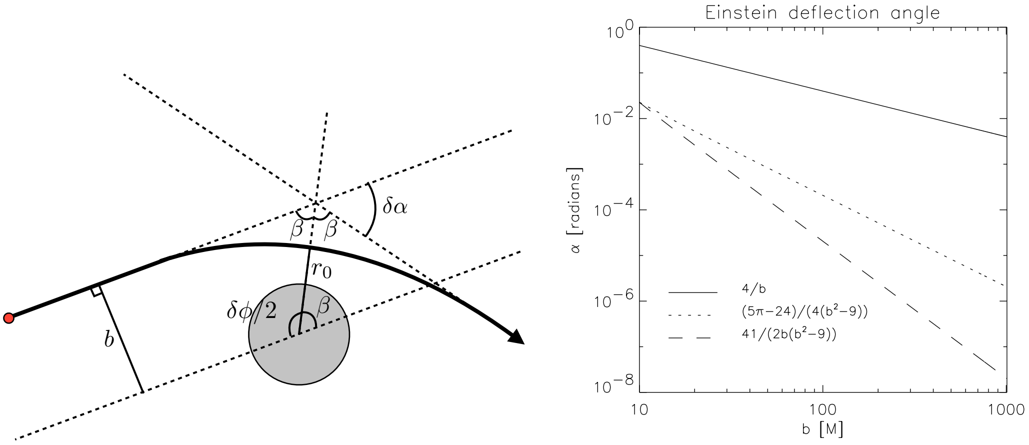


Figure 4.2: **Left:** A schematic diagram of light deflecting around a massive object. The photon path is shown by the solid black line; the undeflected path is marked by the dashed line. **Right:** The accuracy of the first order deflection angle approximation. The standard solution is shown by the solid line (see equation 4.22); higher order correction terms are shown by the dotted/dashed lines.

$$r^2\dot{\phi} = \text{const.} = J \tag{4.11}$$

Similarly, equation 4.7 gives us conservation of specific energy:

$$\kappa c^2 \dot{t} = \text{const.} = E \tag{4.12}$$

For gravitational lensing, we are interested in the 'orbits' of photons. These move along *null* geodesics for which $ds^2 = 0$. This follows from the special relativistic time dilation formula (equation 3.2). As $v \to c$, time becomes infinitely dilated and thus for photons, we must have that the proper time $d\tau = 0$ and therefore the proper length $ds = 0$. Using our above planar geometry, the metric then gives us:

$$ds^2 = c^2 d\tau^2 = 0 = \kappa c^2 dt^2 - \frac{dr^2}{\kappa} - r^2 d\phi^2 \tag{4.13}$$

Dividing through by $d\lambda^2$ and substituting for the specific energy E and angular momentum J then gives:

$$0 = \frac{E^2}{c^2} - \dot{r}^2 - \frac{J^2}{r^2}\kappa \tag{4.14}$$

Now, let us rewrite κ in terms of the *Schwarzschild radius* r_s:

$$\kappa = \left(1 - \frac{r_s}{r}\right) \tag{4.15}$$

which gives the following orbit equation for the photons:

$$\dot{r}^2 = \frac{E^2}{c^2} - \frac{J^2}{r^2}\left(1 - \frac{r_s}{r}\right) \tag{4.16}$$

And now, finally, we can calculate the deflection angle of a photon moving on a hyperbolic orbit past a point mass[1]. The geometry is shown in Figure 4.2. We will first calculate the angle $\delta\phi$ from an integral over angle along the orbit, for which we need:

[1] Really we are calculating the angle for a photon moving through a Scchwarzshild metric. However, if the photon is sufficiently far away then the Scchwarzschild solution approaches that of a point mass and hence we refer to it as such.

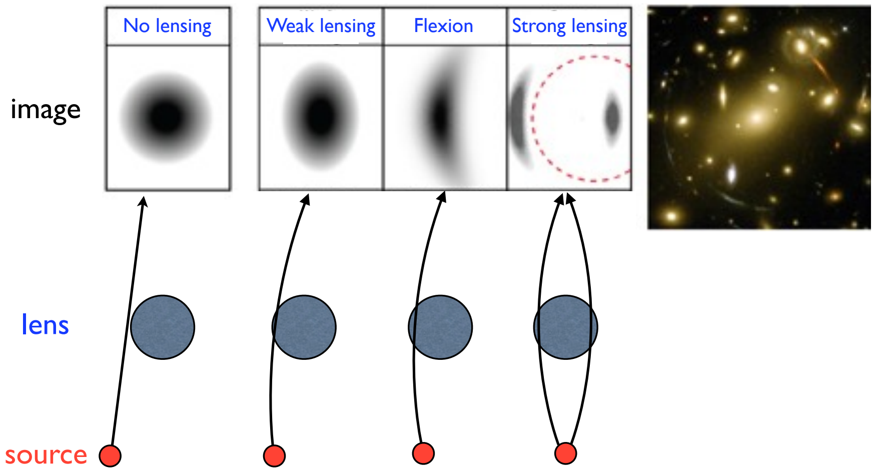


Figure 4.3: A schematic diagram of gravitational lensing. As a source approaches a lens in projection its image is first distorted (weak lensing), bent (flexion) and finally split into multiple images (strong lensing; lensing images taken from plots by Adam Amara). An example of real lensing in the galaxy cluster Abel 1669 is shown on the right. The bright lensing arcs – which are due to strong lensing of distant background galaxies – are clearly visible (credit: NASA, ESA and Johan Richard; Caltech, USA).

$$\begin{aligned}\frac{d\phi}{dr} &= \frac{d\phi}{d\tau}\left(\frac{dr}{d\tau}\right)^{-1} \\ &= \frac{J}{r^2}\left[\frac{E^2}{c^2} - \frac{J^2}{r^2}\left(1 - \frac{r_s}{r}\right)\right]^{-\frac{1}{2}} \qquad (4.17)\end{aligned}$$

where we have set our affine parameter to be the proper time ($\lambda = \tau$). The distance at the turning point r_0 follows from $\dot{r} = 0$:

$$0 = \frac{E^2}{c^2} - \frac{J^2}{r_0^2}\left(1 - \frac{r_s}{r_0}\right) \qquad (4.18)$$

which rearranging, gives a cubic equation for r_0:

$$r_0^3 - b^2 r_0 + b^2 r_s = 0 \qquad (4.19)$$

where $b^2 = \frac{c^2 J^2}{E^2}$, and the largest root must be the physical one. For distant encounters $r_0 \gg r_s$, and we have that $r_0 \simeq b$. In fact, for no mass at all $r_s \to 0$ and we have that $r_0 = b$ *exactly*. This is why b is referred to as the *impact parameter*.

From Figure 4.2, we can define the full deflection angle from the following angle relations:

$$2\beta + \delta\alpha = \pi \qquad ; \qquad \delta\phi/2 + \beta = \pi \qquad (4.20)$$

which gives $\delta\alpha = \delta\phi - \pi$ and thus:

$$\begin{aligned}\delta\alpha &= 2\int_{r_0}^{\infty} \frac{d\phi}{dr} dr - \pi \\ &= 2\int_{r_0}^{\infty} \left[b^{-2} r^4 - r^2\left(1 - \frac{r_s}{r}\right)\right]^{-\frac{1}{2}} dr - \pi \qquad (4.21)\end{aligned}$$

where the factor 2 comes about because we integrate only from r_0 to infinity which gives half of the full deflection. (Note that this is necessary to avoid a coordinate infinity when $r = r_0$.)

The solution of the above integral derives the familiar deflection or *bending angle* formula, and highlights an important point: it is only *approximately* true:

$$\delta\alpha \simeq \frac{2r_s}{b} = \frac{4GM}{c^2 b} \tag{4.22}$$

The approximation is, however, very very good as shown in Figure 4.2, right panel (taken from Mutka and Mähönen 2002).

4.3 The gravitational lens equation

Armed with the bending angle formula, we can now understand *gravitational lensing*, which is due to light bending by massive objects. Figure 4.3 illustrates the idea. If a source is a long way (in projection) from a massive object – from here on the 'lens' – then its light will be very slightly bent. To us it then appears as if the light came from a slightly different location. Resolved sources like galaxies will then appear distorted as the light coming from their centre will be bent differently from the light originating from their edge. We call this *weak lensing*. Moving the source closer to the lens in projection gives stronger distortions called *flexion*. In the limit where the source is aligned perfectly behind a spherical Scchwarzschild lens, due to the symmetry of the problem, the source will be split into a perfect *ring* of images. This is called *strong lensing* and is distinct from weak lensing and flexion by the presence of multiple images for a single source. The beautiful lensing arcs in the galaxy cluster Abel 1669, visible in the right panel of Figure 4.3, owe to strong lensing.

Let us now look at this is a little more detail. The geometry is given in Figure **??**. The angles β and θ are the angle on the sky to the source and image respectively, and $\delta\alpha$ is the bending angle as previously (β is *not* the same as in Figure 4.2). Now, from Figure **??** it is clear that (assuming small angles):

$$\theta D_S = \beta D_S + \delta\alpha D_{LS} \tag{4.23}$$

which is called the *gravitational lens equation*. Now, Figure **??** makes the problem look planar, but in fact lensing is two-dimensional on the sky. Thus, in general, we should replace all of the scalar angles by *vector* angles. For a point mass lens, the lens equation then becomes:

$$\boldsymbol{\theta} = \boldsymbol{\beta} + \frac{D_{LS}}{D_S D_L}\frac{4GM}{c^2}\frac{\boldsymbol{\theta}}{|\boldsymbol{\theta}|^2} \tag{4.24}$$

where we have used the fact that the impact parameter $b \simeq \theta D_L$, or in two dimensions: $\mathbf{b} = \frac{\boldsymbol{\theta}}{|\boldsymbol{\theta}|^2} D_L$. We can then write:

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \frac{\theta_E^2}{|\boldsymbol{\theta}|^2}\boldsymbol{\theta} \tag{4.25}$$

which defines the *Einstein radius*:

$$\theta_E^2 = \frac{D_{LS}}{D_S D_L}\frac{4GM}{c^2} \tag{4.26}$$

The meaning of the Einstein radius is then clear: it is the image position on the sky for an on-axis source $\boldsymbol{\beta} = \mathbf{0}$.

By symmetry, for a circularly symmetric lens, we can write $\boldsymbol{\theta} = (\theta, 0)$, $\boldsymbol{\beta} = (\beta, 0)$ without loss of generality (not so for non-symmetric lenses, of course). Thus in general for a point mass lens we must solve the quadratic equation:

$$\theta^2 - \beta\theta - \theta_E^2 = 0 \tag{4.27}$$

and thus, for a point mass lens there will be at most two images at $\theta = \theta_\pm$. In fact, this is true for any circularly symmetric lens as can be understood by replacing $M \to M(\theta)$ (true from Birkoff's theorem). When the source is perfectly on axis ($\beta = 0$), these two images join to form an Einstein ring (as we already noted previously from intuition).

4.4 Magnification and distortion

We have shown that massive bodies bend light and that this leads to a gravitational lensing effect. However, light is not just bent. If we consider infinitesimally close light rays, we will see that the light is also *magnified* and *distorted*. We can think of equation 4.25 as mapping the image positions $\boldsymbol{\theta}$ to the source position $\boldsymbol{\beta}$. A general map is determined (à la GR) by the matrix of second derivatives:

$$\beta_i = \frac{\partial \beta_i}{\partial \theta_j}\theta_j \tag{4.28}$$

where $i = 0, 1$ and as usual, repeated indices are summed over. The eigenvectors and eigenvalues of this matrix then define the *magnification* and distortion – called the *shear* of the source. The components of the matrix are derived by taking derivates of the lens equation (4.25). Writing as above, without loss of generality, $\boldsymbol{\theta} = (\theta, 0)$, $\boldsymbol{\beta} = (\beta, 0)$, we derive the transformation matrix (for a Schwarzschild lens) as:

$$\frac{\partial \beta_i}{\partial \theta_j} \equiv \begin{pmatrix} 1 + \frac{\theta_E^2}{\theta_\pm^2} & 0 \\ 0 & 1 - \frac{\theta_E^2}{\theta_\pm^2} \end{pmatrix} \tag{4.29}$$

where $\theta_\pm = \frac{1}{2}\left[\beta \pm \sqrt{\beta^2 + 4\theta_E^2}\right]$ defines the positions of the two images.

The above is a diagonal matrix. Its eigenvectors and eigenvalues define the transformation (from image to source) and, since the matrix is already diagonal, can be simply read off. What is more interesting however, is the *magnification tensor* that is the *inverse* of the above matrix. It defines how the source is transformed into the image:

$$\mathbf{M}_\pm \equiv \begin{pmatrix} (1 + \frac{\theta_E^2}{\theta_\pm^2})^{-1} & 0 \\ 0 & (1 - \frac{\theta_E^2}{\theta_\pm^2})^{-1} \end{pmatrix} \tag{4.30}$$

The scalar *magnification* is then defined as the determinant of this matrix (also called the *Jacobian*). Being diagonal, this is simply given by the trace:

$$M_\pm = \left| 1 - \left(\frac{\theta_E}{\theta_\pm}\right)^4 \right|^{-1} \tag{4.31}$$

Note that the two eigenvalues of $\mathbf{M}_\pm$ are different showing that images are both magnified and *distorted*. Furthermore, apart from on-axis sources (for which $\theta = \pm\theta_E$), the two images are magnified differently. One image (θ_-) is *de-magnified*; the other (θ_+) is magnified. For on-axis sources, the magnification *diverges*! This is because the image is split into a perfect Einstein ring.

4.5 Lensing and dark matter: what lensing really measures

We are ultimately interested in using lensing as a dark matter probe, and this will be covered in detail in lecture 11. However, already we can see something important from the lens equation. It is difficult to measure $\boldsymbol{\beta}$ – the position of the source on the sky since by definition we see only the images. The exception to this is $\boldsymbol{\beta} = \mathbf{0}$ for which we see the full Einstein ring. This suggests that on axis sources maximise the available information (for circularly symmetric lenses). Thus, a single source splitting into a ring really tells us only one thing: the Einstein radius θ_E which (assuming we know the distances to the source and lens) tells us the *enclosed lensing mass within* θ_E. It is then immediately clear that if we want to measure the *distribution* of mass within a lens, we will require more information than that from a single source alone. We return to this in lecture 11.

For now, however, we are already able to place a crude constraint on the mass of a galaxy cluster within θ_E from its gravitational lensing arcs – if we assume the cluster is circularly symmetric. Let us take the case of Abel 1703, shown in Figure 4.5. Assuming the outermost arc is a perfect Einstein ring (which it is not), that the cluster is spherical (which it is not), and using our current cosmological model to determine the distances (of which more in later lectures), we can use equation 4.26 to determine the mass inside the lensing arc marked 10-11 on Figure 4.5:

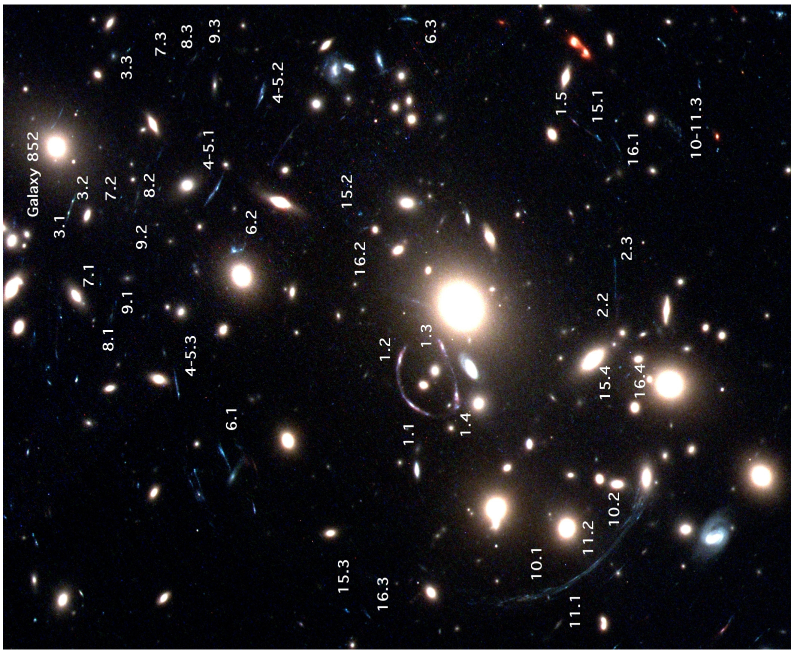


Figure 4.5: The lensing galaxy cluster Abel 1703 (credit: Limousin et al. 2008). The lensing arcs are marked by the numbers. The image is 77 × 107 arcsec. The distance to the cluster is $D_L \sim 868\,\mathrm{Mpc}$. The giant arc marked 10-11 is at a distance from the centre of ~ 35 arcsec ($\sim 150\,\mathrm{kpc}$), and a cosmic distance of $D_S \sim 1615\,\mathrm{Mpc}$.

$$M(<\theta_E) \sim \frac{\theta_E^2 D_S D_L c^2}{4GD_{LS}} \tag{4.32}$$

which gives, putting in numbers from the Figure caption, $M(150\,\mathrm{kpc}) \sim 2.8 \times 10^{14}\,\mathrm{M}_\odot$. This is not too far off results from more sophisticated analyses. Limousin *et al.* 2008 find, for example, $M(210\,\mathrm{kpc}) = 2.4 \times 10^{14}\mathrm{M}_\odot$. As for the Coma cluster, this is far larger than the amount of visible mass in this cluster: we have confirmed the existence of dark matter, but this time using a completely different technique.

Note that, in principle we could use the magnification (equation 4.31) combined with the observed position of the images to calculate θ_E, even for off-axis sources. The problem there is two-fold. Firstly, we do not (in general) know the intrinsic luminosity of the source unless we are lucky enough to catch a *standard candle* lensing, like e.g. Type Ia supernovae (c.f. the distance ladder discussion in §1). We can avoid this problem, however, by considering the *ratio* of the magnification of the two images:

$$\frac{M_+}{M_-} = \frac{\left|1 - \left(\frac{\theta_E}{\theta_-}\right)^4\right|}{\left|1 - \left(\frac{\theta_E}{\theta_+}\right)^4\right|} \tag{4.33}$$

where we now solve for θ_E using the above equation and the observed $\theta_\pm$ of the two images. Here, however, we encounter a second problem: *anomalous flux ratios*. Many lensing systems have unexpected magnification ratios between the images because the lensing potential is not smooth. A small local overdensity – even a star nearby to an image – can cause this to happen making the magnification ratios difficult to work with.

We will return to lensing as a dark matter probe in lecture 11. There we will relax the above crude assumptions and explore further how to calculate the *distribution* of mass – in particular dark matter – within the lens.

Lecture 5

What dark matter is not

In this lecture, we explore several simple solutions to the puzzle of missing matter in the Universe. We start by suggesting that the 'dark matter' is just faint stars, or difficult to detect gas. Using direct observations, we show that such 'dark baryons' are simply not there (more indirect cosmological probes of the baryon content of the Universe will be presented in §7). Next, we suggest that dark matter could comprise compact objects like black holes, or small planets/asteroids. We show that these would produce an observable microlensing signal in our galaxy that is also not seen.

5.1 Dark matter as faint stars

The best place to place constraints on dark matter as faint stars is our own Galaxy. There we can actually count individual stars and – due to their close proximity – see the very faintest stars that can possibly exist. This was first acheived using the Hubble Deep Field image which was created by staring at a single patch of the sky non stop for 10 days (Figure 5.1; Flynn *et al.* 1996).

Using the HDF, Flynn *et al.* 1996 found very few faint stars: too few by far to account for the observed dark matter. Let us look a little at their calculation to be sure there are no important caveats. From the Milky Way rotation curve, we can estimate the local dark matter density near the Solar neighbourhood. We will revisit this problem later on the course where we will present a much more detailed and rigorous analysis. For now, let's assume that the Milky Way halo density profile is a spherical power law:

$$\rho_{\rm dm} = \rho_0 \left(\frac{r}{r_s}\right)^{-\alpha} \tag{5.1}$$

where we have normalised the distribution so that the local dark matter density at the Solar neighbourhood $r = r_s \sim 8\,$kpc is given by ρ_0.

In this case, we may write the rotation curve as:

$$v_c^2 = \frac{GM_{\rm dm}(r_s)}{r_s} + v_{c,b}^2(r_s) \tag{5.2}$$

where $v_{c,b}^2(r_s) \sim (150\,{\rm km/s})^2$ is the baryonic contribution to the rotation curve at r_s (Klypin *et al.* 2002). The enclosed dark matter mass $M_{\rm dm}(r_s)$ is then given by:

$$\begin{aligned} M_{\rm dm}(r_s) &= 4\pi \int_0^{r_s} \rho_0 \left(\frac{r}{r_s}\right)^{-\alpha} r^2 dr \\ &= \frac{4\pi\rho_0 r_s^3}{3-\alpha} \end{aligned} \tag{5.3}$$

Thus, we derive:

$$\rho_0 = \frac{\left(v_c^2 - v_{c,b}^2\right)(3-\alpha)}{4\pi G r_s^2} \tag{5.4}$$

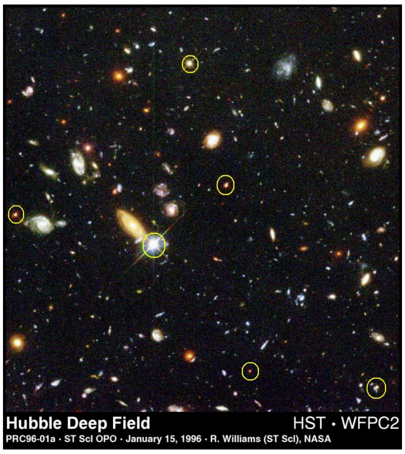

Hubble Deep Field
HST • WFPC2
PRC96-01a • ST ScI OPO • January 15, 1996 • R. Williams (ST ScI), NASA

Figure 5.1: The Hubble Deep Field: hunting for faint stars in the Milky Way. Many thousands of galaxies were found in this image, but only very few stars (marked by the yellow circles). The number of stars found is too small to account for the missing 'dark matter' in the Galaxy.

and using $v_c \sim 220\,\mathrm{km/s}$, $v_{c,b} \sim 150\,\mathrm{km/s}$ and $r_s \sim 8\,\mathrm{kpc}$, we arrive at:

$$\rho_0 = 0.0075(3-\alpha)\mathrm{M}_\odot\,\mathrm{pc}^{-3} \tag{5.5}$$

which is remarkably close to the canonical value obtained over many years in the literature (but not so close to that obtained from a purely local measure as we shall discover later on). The flatness of the observed rotation curve of the Milky Way tells us that $\alpha < 2$ (any larger, and it would fall too quickly with radius). Thus we derive a minimum density:

$$\rho_{0,\mathrm{min}} = 0.0075\mathrm{M}_\odot\,\mathrm{pc}^{-3} \tag{5.6}$$

Now, suppose that this missing mass comprised many faint undetected stars of mass m_* and luminosity L_*. If I stare for a long time along one sight line, then I can sum up all the stars along this line that I might detect. Remember, I can actually resolve each star, so what matters is my magnitude limit, not anything else. Recall from lecture 1 the definition of an apparent magnitude:

$$m_I - K_I = -2.5\log_{10}\left[\frac{L_*}{L_\odot}\frac{(10\,\mathrm{pc})^2}{d^2}\right] \tag{5.7}$$

where m_I is the apparent magnitude in waveband I, and $K_I = 4.08$ is a calibration constant for that band. Defining m_I now as a *limiting* magnitude and rearranging, we can turn this into a limiting distance:

$$\left(\frac{d_{\mathrm{max}}}{1\,\mathrm{pc}}\right) = 10\left[10^{\frac{m_I-K_I}{2.5}}\left(\frac{L_*}{L_\odot}\right)\right]^{\frac{1}{2}} \tag{5.8}$$

and we can then simply integrate over all stars of mass m_* out to d_{max}. Let us assume that the dark matter density is $\sim$ constant over this range and given by $\rho_{0,\mathrm{min}}$. Thus, the minimum expected number of detections is given by:

$$\begin{aligned} N_{\mathrm{exp}} &= \Omega\int_{d_{\mathrm{min}}}^{d_{\mathrm{max}}}\frac{\rho_0}{m_*}r^2dr \\ &= \frac{\Omega}{3}\frac{\rho_{0,\mathrm{min}}}{m_*}\left(d_{\mathrm{max}}^3 - d_{\mathrm{min}}^3\right) \end{aligned} \tag{5.9}$$

where Ω is the solid angle of sky subtended by the observation and d_{min} is set similarly to d_{max} by the *maximum* magnitude considered (we are interested only in the very faintest stars here). The

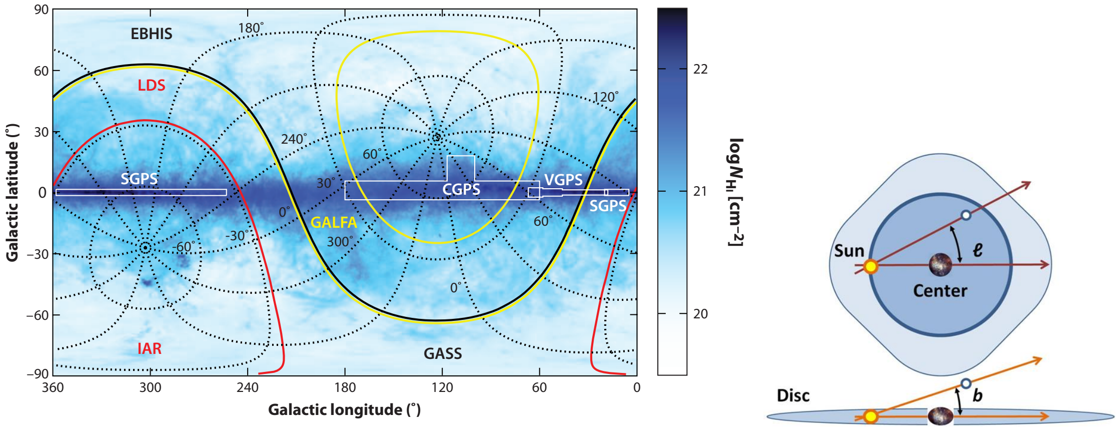


Figure 5.2: **Left:** The observed HI gas distribution in the Milky Way from a number of Galactic surveys as a function of Galactic latitude l and longitude b in degrees. **Right:** A schematic diagram explaining the l, b Sun-centred Galactic coordinate system (credit: Brews ohare).

Hubble Deep Field observations have $\Omega = 4.4\,\text{arcmin}^2$, while Flynn *et al.* 1996 find $N_{\rm det} < 3$ at 95% confidence for an I-band magnitude window of $24.63 < m_I < 26.3$ with colours $V - I > 1.8$. The faintest stars known – red dwarfs – are less than 1000 times fainter the Sun, and just 8% of the mass (Richer *et al.* 2006). Using this as a lower bound on visible objects in the halo, we can use equation 5.9 to estimate the expected number of detections of such stars if they comprise all of the missing matter in the Milky Way's halo. This gives $N_{\rm exp} = 30$ which rules out red dwarfs as making up all of the missing mass in our Galaxy. But we can go further. What if the halo comprised even fainter white dwarf stars? These have a maximum mass of $m_* = 1.4M_\odot$ (the Chandrasekhar limit; Phillips 1999). Their luminosity continuously falls with time but with a known rate. Given the age of the Universe to cool, this sets a minimum luminosity for white dwarfs; the faintest ever detected has a V-band magnitude of $M_V = 17.4$ – some 100,000 times fainter than the Sun (Richer *et al.* 2006). This corresponds to an I-band magnitude of $M_I \sim 14.9$ (Flynn *et al.* 1996). With these numbers, we would expect (assuming all of the missing matter comprises faint white dwarfs) $N_{\rm exp} \sim 0.3$ which is not ruled out by the observations. Thus, we have proven that the missing matter cannot be low mass stars, but white dwarfs remain a possibility. We will return to these in §5.3.

5.2 Dark matter as gas

Most of the gas in the Universe is hydrogen and so our observational constraints will focus on hydrogen gas in its various forms: H_2, HI, Hα etc. (c.f. §1). As above, let's start with our own Galaxy before moving to extragalactic systems.

5.2.1 The Milky Way

Neutral hydrogen Easiest to detect is neutral hydrogen – HI – because of its forbidden 21cm emission line (§1). It has temperatures in the range $\sim 100 - 3000\,$K. The total HI mass in the disc of the Milky Way out to the Solar neighbourhood is $\sim 10^{10}$M$_\odot$, with only a small fraction of this seen out of the plane (Marasco and Fraternali 2011; Binney and Merrifield 1998). This may be compared with the minimum missing mass required to fit the rotation curve, using equation 5.3, with $\alpha = 2$: $M_{\rm dm}(r_s) = 5 \times 10^{10}M_\odot$, thus HI is not sufficient to explain the discrepancy. (See Figure 5.2; Kalberla and Kerp 2009.)

Ionised gas Ionised hydrogen is straightforwards to detect through its Balmer-α lines visible at optical wavelengths (see §1), denoted Hα. The ionised gas fraction in the Milky Way is about half of the total gas mass in the disc, but not nearly enough to explain the missing matter (Ferrière 2001;

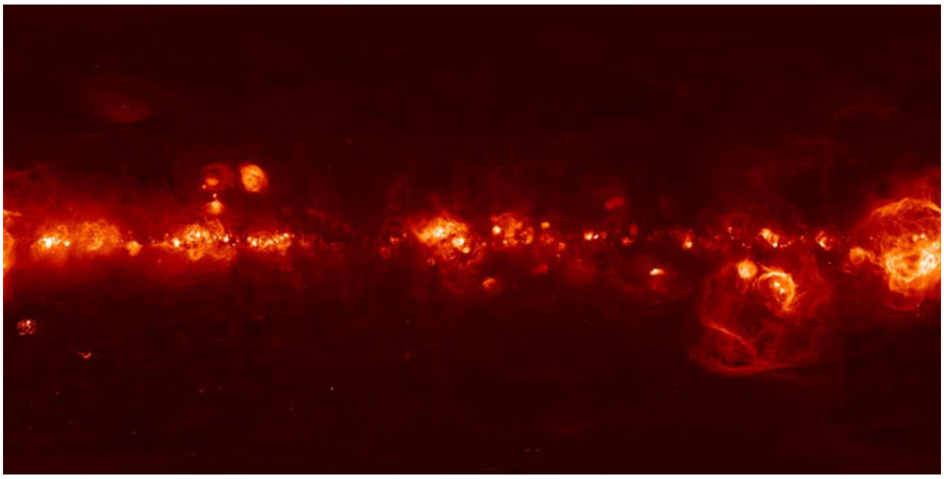


Figure 5.3: The observed Hα gas distribution in the Milky Way from a number of Galactic surveys.

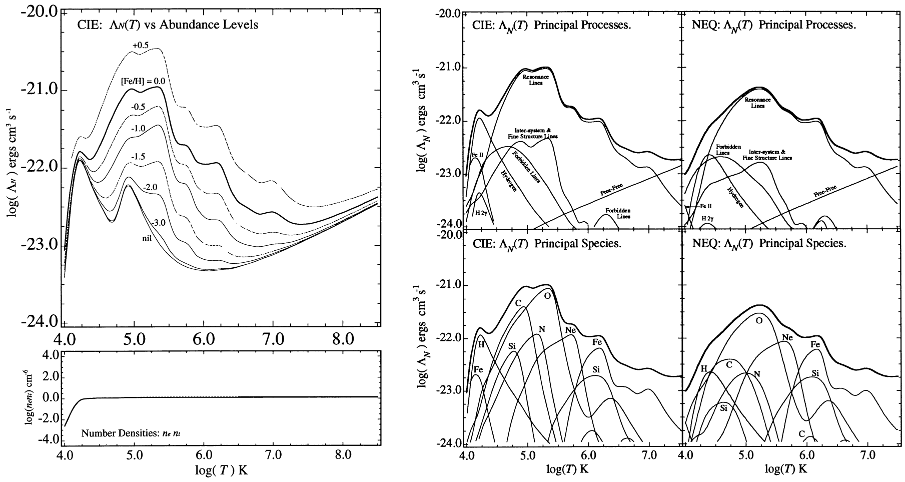


Figure 5.4: The cooling curve of a hydrogen plasma as a function of metallicity (left) and broken down by cooling mechanism (right). The cooling rate Λ is defined normalised to 1 atom/cc of gas as shown in the bottom panel.

and see Figure 5.3 from Finkbeiner 2003).

Hot gas Very hot gas $\gtrsim 10^7\mathrm{M}_\odot$ is relatively easy to detect because it emits via thermal Bremstrahlung radiation in the X-rays. We can estimate the equilibrium temperature of gas in the Milky Way using the kinetic theory of gases:

$$\frac{3}{2}k_B T \sim \frac{1}{2}m_p\sigma^2 \tag{5.10}$$

where k_B is the Boltzmann constant, T is the gas temperature, $\sigma \sim v_c/\sqrt{2}$ is the typical velocity of gas in the Milky Way halo[1], and m_p is the mass of a proton. Using $v_c \sim 220\,\mathrm{km/s}$, this gives $T \sim 10^6\,\mathrm{K}$. Unfortunately, this is a temperature for Hydrogen that is difficult to detect. This is shown by the hydrogen plasma *cooling curve* (with 'metal'[2] impurities) shown in Figure 5.4, taken from Sutherland and Dopita 1993. The left panel shows the cooling rate, normalised to a density of

[1] This is exactly correct for an isothermal sphere.

[2] Recall that we astrophysicists call every element heavier than Hydrogen a 'metal'; §1.

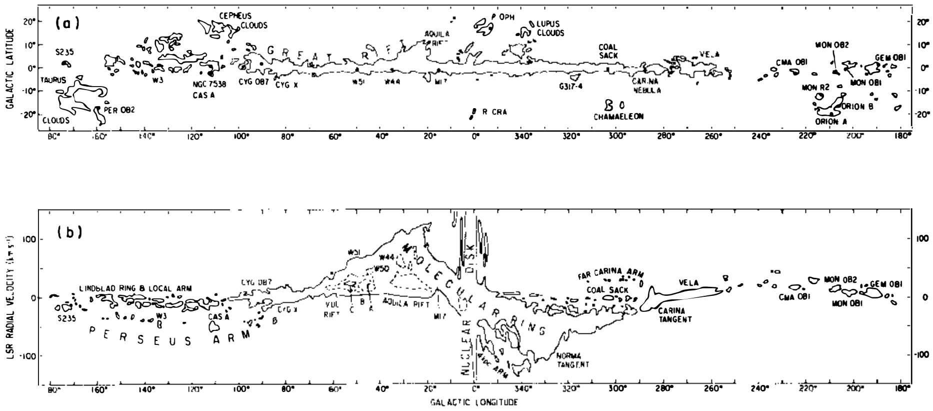


Figure 5.5: The observed distribution of CO in the Milky Way – a good tracer of the underlying H_2 gas.

1 atom/cc, as a function of metallicity. The right panel shows a breakdown of the cooling rate by cooling mechanism and corresponding elements. It also shows results assuming collisional ionisation equilibrium (CIE) – the usual approximation, and how this changes away from equilibrium (i.e. when the cooling is sufficiently fast that the rate of collisions cannot reach equilibrium). Notice that around $10^6 < T < 10^7$ K, there is a dip in the cooling rate as we move away from resonant cooling lines and towards free-free cooling (thermal Brehmstrahlung). While this warm-hot gas is difficult to detect, we can infer its presence in absorption along Pulsar sight lines (e.g. Ferrière 2001). This gives $n_H \sim 5 \times 10^{-4}$atoms/cc out to ~ 50 kpc which implies a total mass of $M(50\,\text{kpc}) \sim 10^{10}\text{M}_\odot$ (Anderson and Bregman 2010). Thus, such hot gas is unlikely to comprise the missing matter.

Cold gas Perhaps hardest to constrain is the cold molecular hydrogen in the Milky Way: H_2. Its presence can be inferred from other gaseous tracers like carbon monoxide (CO), but it is difficult to detect directly either in emission or absorption because it is visible in absorption only in the ultra violet (Combes 1991). By contrast, CO is much easier to detect because it has a low energy roto-vibrational transition that is excited by H_2 molecules (and therefore traces the cold gas), and is visible in the radio (Combes 1991). From these CO observations, the total mass in H_2 in the Milky Way is small: just $\sim 2.5{\times}10^{10}\text{M}_\odot$ (Combes 1991; and see Figure 5.5). However, the difficulty of directly probing H_2 led some authors to even claim that such cold gas can comprise all of the missing matter (Pfenniger *et al.* 1994). They proposed that the H_2 would be in the form of small dense clumps that are then hard to detect because the probability of catching one in absorption along a sight line to a star or bright distant galaxy is small. However, cosmic rays impacting such clumps will cause them to shine in γ-rays, making them detectable. Kalberla *et al.* 1999 use this to place constraints on the total mass in such clumps finding that they could still be a significant gas mass component in the Galaxy, but cannot explain all of the observed missing mass.

Even adding up all of the gaseous components of the Milky Way, we fall short of the $\sim 5 \times 10^{10}\text{M}_\odot$ required to match the rotation curve at the Solar neighbourhood (recall that some of the above estimates are the *total* mass in the Milky Way, not the mass at $R_\odot$). The situation gets worse as we move to even larger radii. Thus, the missing mass in our Galaxy cannot comprise undetected gas.

5.2.2 Galaxy clusters

Our view of our own Galaxy is unprecedented and it is not possible to do better in extragalactic systems. The *observed* distribution of stars and gas in galaxies is certainly not able to explain the observed dynamical masses (e.g. Read and Trentham 2005). However, the mass in dark baryons is assumed to be small based on the Milky Way observations, above. We cannot at present hope to hunt for faint stars even in very nearby galaxies like Andromeda. By contrast, however, *galaxy clusters*

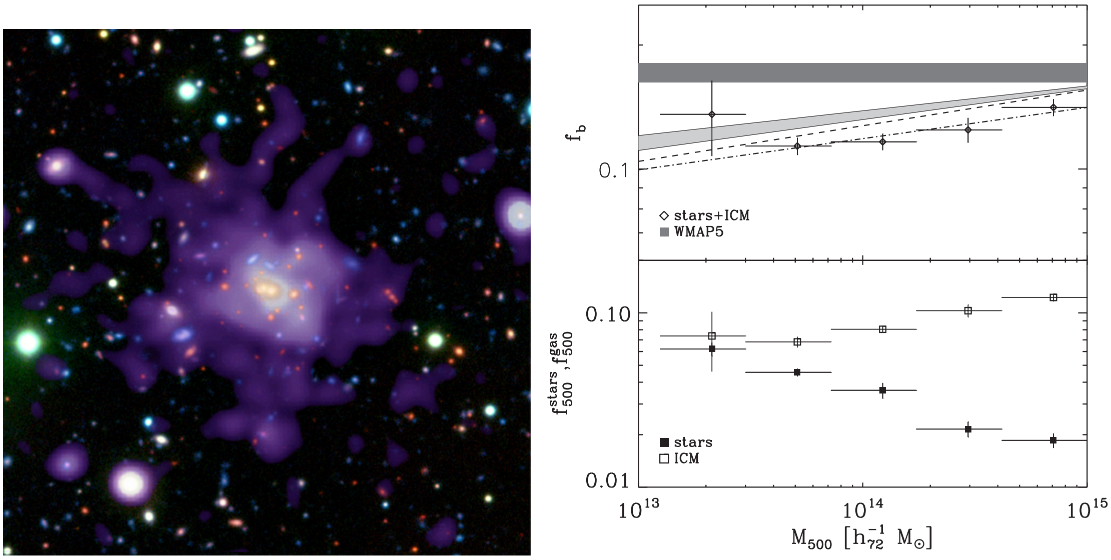


Figure 5.6: **Left:** Hot X-ray emitting gas (purple) in a galaxy cluster (credit: X-ray image NASA/CXC/ESO/P. Rosati et al.; optical image: ESO/VLT/P. Rosati et al.). **Right:** The observed baryon fraction in clusters as a function of cluster mass. The cosmological value is marked in grey; this will be discussed in later lectures.

are an excellent place to constrain dark baryons. As we saw in 2, there is plenty of missing mass in clusters too (this was where it was first discovered, after all), but dark baryons are much easier to spot. Using equation 5.10 with $\sigma \sim 1000\,\mathrm{km/s}$, we have that $T \sim 10^7\,\mathrm{K}$ and thus clusters, unlike galaxies, will shine brightly in the X-ray. Indeed, most of the baryonic mass in clusters is in X-ray emitting gas. Figure 5.6 shows an example of an X-ray emitting galaxy cluster (left), and a plot of the observed baryon fraction in clusters as a function of cluster mass (right; Giodini *et al.* 2009). The stellar mass is determined from adding the light in visible galaxies (an evolution of Zwicky's original analysis). The gas fractions come directly from the observed X-ray gas surface densities, and the masses come from a dynamical analysis that assumes that the gas is in hydrostatic equilibrium. We will study this is more detail later, but roughly one can think of this as determining the cluster mass through equation 5.10 from the observed temperature of the X-ray gas.

5.3 Dark matter as compact objects: 'MACHOs'

We have demonstrated in 5.1 that dark matter cannot comprise faint stars. However, lowering the luminosity further, it is perfectly possible to have dark matter comprise massive compact objects that are just a few times fainter than the faintest stars. These are called *Massive Compact Halo Objects*, or MACHOs and could be stellar mass black holes, planets, asteroids, or something more exotic like primordial black holes (e.g. Hawking 1971). Apart from perhaps primordial black holes, there can be difficulties in understanding how such objects got there, but until we have empirically ruled out such a possibility then, formally, it remains. Luckily, we can directly test this idea using an effect called *microlensing* that we describe next.

5.3.1 Microlensing

Microlensing is simply *unresolved* strong gravitational lensing. Recall from lecture 4, the Einstein radius:

$$\theta_E^2 = \frac{D_{LS}}{D_S D_L}\frac{4GM}{c^2} \tag{5.11}$$

Figure 5.7: **Left:** The Large and Small Magellanic Clouds (LMC/SMC) – two dwarf galaxy companions of the Milky Way just $\sim 50\,\mathrm{kpc}$ away (credit: ESO/S. Brunier). **Right:** The Magellanic Clouds, visible with the naked eye from the Southern hemisphere.

Now, we are interesting in seeing a lensing signal from MACHOs within our Galaxy. First, we need some sources to be lensed. Ideally, we would like a large number of stars in one patch of the sky that are resolved, and for which we know the distance. This suggests looking towards a nearby pair of dwarf galaxies: the Large and Small Magellanic Clouds (LMC/SMC), shown in Figure 5.7. These are just $\sim 50\,\mathrm{kpc}$ away and bright enough that you can see then with the naked eye from the Southern hemisphere. Second, we need to know the distance to the lenses. From equation 5.11, we see that if $D_L = D_S$ then the Einstein radius shrinks to zero. The limit $D_L \to 0$ is also bad for lensing since the Einstein radius expands to infinity. Thus there is a sweet spot for lenses somewhere between the source and observer. Let us consider halfway: $D_L = 0.5 D_S$. Thus the Einstein radius becomes:

$$\theta_E^2 = \frac{4GM}{D_S c^2} \tag{5.12}$$

Putting $M \sim \mathrm{M}_\odot$ for Solar mass lenses and $D_S \sim 50\,\mathrm{kpc}$ for the Magellanic clouds, this gives $\theta_E = 4 \times 10^{-4}$ arcsec. This is far smaller than the point spread function of current instruments and so the lensing will be *unresolved*. This seems like a disaster: how are we to know that a source is being lensed? The solution lies in the *time domain*. Recall that the typical velocity of objects moving in the halo of our Galaxy is $v_{\rm typ} \sim v_c \sim 200\,\mathrm{km/s}$. So the time it takes for such objects to move through an angle θ_E is: $t = \theta_E D_L / v_{\rm typ} \sim 90$ days. This is measurably short! When a lens moves in front of a background star, although the images are not resolved we can still see the magnification effect. The star will brighten and then dim as the lens passes in front of it. This is called *microlensing* (because it would require micro-arcsec resolution to see the resolved lensing effect). We can determine the form of this brightening and dimming – the *lightcurve* – from equation 4.31:

$$M_\pm = \left|1 - \left(\frac{\theta_E}{\theta_\pm}\right)^4\right|^{-1} \tag{5.13}$$

Since the two images are now unresolved, we will just see the sum of their brightening:

$$M = M_+ + M_- = \left|1 - \left(\frac{\theta_E}{\theta_+}\right)^4\right|^{-1} + \left|1 - \left(\frac{\theta_E}{\theta_-}\right)^4\right|^{-1} \tag{5.14}$$

where the image positions are given by: $\theta_\pm = \frac{1}{2}\left[\beta \pm \sqrt{\beta^2 + 4\theta_E^2}\right]$ and β is the position of the source on the sky. Notice that for $\beta \to \infty$, $\theta_+ \to \beta$, $\theta_- \to 0$ and thus $M \to 1$ as it should. The brightening is maximised for on-axis sources ($\beta = 0$). The lightcurve equation can be further simplified by substituting for $\theta_\pm$ to give:

$$M = \left|\frac{u^2 + 2}{u\sqrt{u^2 + 4}}\right| \tag{5.15}$$

where $u = \beta/\theta_E$.

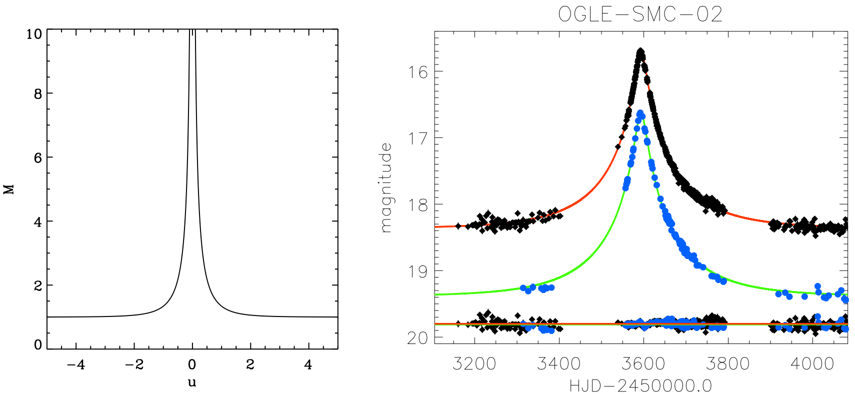


Figure 5.8: **Left:** The characteristic lightcurve of a microlensing event (obtained from equation 5.15). **Right:** An example of a real 'poster-child' microlensing event observed towards the SMC: OGLE-SMC-02. The black and blue points show data for the I- and V-bands, respectively. Notice that this event is achromatic as expected for microlensing. The solid curves show fits using the lightcurve equation (equation 5.15). The residuals of the model fitting are shown at the bottom.

Figure 5.8 shows a plot of a lightcurve determined using equation 5.15 (left), and a real observation of a microlensing lightcurve: OGLE-SMC-02 (right; plot taken from Wyrzykowski *et al.* 2011). This particular microlensing event is fascinating because the data are so good that Dong *et al.* 2007 could not fit a single point mass lens to the data, finding instead that a *binary lens* is a better fit. There is something even more remarkable, though about this event. If the lens cannot be actually seen (which is the case here), then we have no hope of determining the distance D_L. This then leads to a degeneracy between the distance and the mass of the lens as can be seen by simple inspection of equation 5.11. However, in a visionary paper Refsdal 1966 suggested that this degeneracy could be broken if the lensing event is viewed from two different locations separated by a large baseline. This is now possible by combining measurements from the ground and in space. Dong *et al.* 2007 did exactly this to measure what is called the *microlensing parallax* for OGLE-SMC-02. This allowed them to break the degeneracy and determine that this lens lies in the halo of the Milky Way, most likely with a mass $\sim 10\mathrm{M}_\odot$. Thus it must be a *binary black hole* in the halo of the Milky Way: a truly dark Massive Compact Halo Object!

The above demonstrates that we can hunt for MACHOs by staring for a long time at stars in the LMC and SMC and waiting for them to show a characteristic brightening and dimming due to microlensing, an idea first proposed by Paczynski 1986. We can separate the microlensing signal from the intrinsic luminosity variations of stars because the microlensing signal should have a characteristic shape as shown in Figure 5.8, should be *achromatic*, and should occur only once. Thus, the main source of background will actually be real microlensing events but from other *stars* (or stellar remnants), rather than from dark matter. We can weed these out too, however, by hunting for the actual lensing star that should also be visible in the data. Even if this proves difficult (because the lensing star is too faint, for example), the amount of microlensing needed to explain all of the missing matter is much larger than expected from the visible light alone. So a simple excess of mircolensing events is enough to determine whether MACHOs are indeed the dark matter.

Three main collaborations have been performing the above experiment: MACHO (Alcock *et al.* 1993), EROS (Aubourg *et al.* 1993), and OGLE (which looked initially towards the Galactic bulge Udalski *et al.* 1992, and only later towards the LMC/SMC; e.g. Wyrzykowski *et al.* 2011). MACHO and EROS both reported their first microlensing events in 1993 (Alcock *et al.* 1993; Aubourg *et al.* 1993). The latest results from all three experiments are summarised in Figure 5.9 (taken from Wyrzykowski *et al.* 2011). An earlier claim from the MACHO collaboration that the halo could comprise a significant fraction of $\sim$ solar mass MACHOs is now excluded at 95% confidence by both the EROS and OGLE experiments. These experiments – in particular EROS-1 and EROS-2 – prove that dark matter cannot comprise massive compact objects in the range $10^{-7} < M/\mathrm{M}_\odot < 20$ (Tisserand *et al.* 2007).

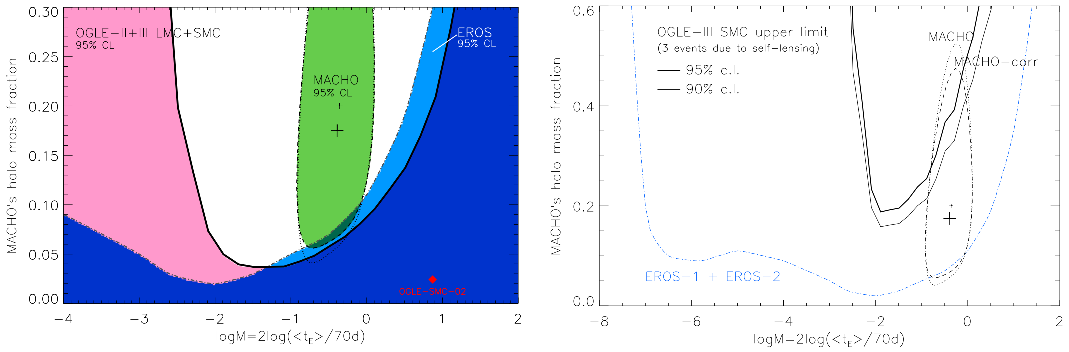


Figure 5.9: The latest constraints on MACHO dark matter from the OGLE, EROS and MACHO experiments. **Left:** Constraints from the combined LMC+SMC OGLE data with constraints from MACHO and EROS overlaid. Only the blue region is still allowed by the combined data. Marked in red is the implied fraction of halo mass comprised of objects like OGLE-SMC-02 (shown in Figure 5.8). An earlier claim from the MACHO collaboration that the halo could comprise a significant fraction of $\sim$ solar mass MACHOs (shown in green) is now excluded at 95% confidence by both the EROS and OGLE experiments. **Right:** The full exclusion region from EROS-1 and EROS-2 (everything above the blue dashed line is excluded). Here the OGLE constraints appear weaker because only the SMC data have been used.

Lecture 6

Dark matter as alternative gravity

In this lecture, we consider whether modifications to our theory of gravity could explain dark matter. We show that a definitive answer remains elusive due to the challenge of making concrete predictions. However, current data appears to disfavour wide classes of models as a complete explanation of the dark matter. Combined with results from previous lectures, we are left with one remaining possibility (apart from our lack of imagination): that dark matter is some new as yet undetected type of matter.

6.1 Introduction

We have demonstrated that dark matter does not comprise faint stars or gas, or even massive compact non-emitting bodies – at least over the mass range $10^{-7} < M/\mathrm{M}_\odot < 20$. This suggests more exotic explanations. One possibility is that we simply have gravity wrong. Recall from §3 that while general relativistic dynamics is quite secure, the Einstein field equations are only the simplest thing we could write down – not the only possibility. Could it be that modifications to the field equations could masquerade as missing mass on galaxy and cluster scales in the Universe?

6.2 Lagrangians and action principles for field theories

We have already encountered the Lagrangian L for describing the dynamics of systems. For example, the Lagrangian for classical mechanics is given by $L = T - V$ where T is the kinetic, and V the potential energy of the system. We can also write down a Lagrangian to describe a *field* (really a Lagrangian *density*) $\mathcal{L}$ which is a function of the fields $\phi_i(x, y, z, t)$ and their derivatives. These fields can be, for example, the temperature at different points in a room, the velocity of a fluid, or the gravitational potential Φ. This Lagrangian density appears inside the action:

$$S = \int \mathcal{L} d^n\mathbf{x} \tag{6.1}$$

where n is the number of dimensions (3 for classical gravity; 4 for GR) and the action is now a (hyper)-*volume* integral which is why $\mathcal{L}$ is called a Lagrangian *density.* The field equations then follow from extremising the action $\delta S = 0$, which yields the Euler-Lagrange equations:

$$\partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi_i)}\right) - \frac{\partial \mathcal{L}}{\partial \phi_i} = 0 \tag{6.2}$$

where now our “coordinates” are the fields ϕ_i themselves.

The above is a very useful abstraction because we can now construct field equations through the Lagrangian $\mathcal{L}$ that obey symmetries *by construction.* Let us consider a concrete example: the Lagrangian density for Newtonian gravity:

$$\mathcal{L}_N = -\frac{|\nabla\Phi|^2}{8\pi G} - \rho\Phi \tag{6.3}$$

Application of the Euler-Lagrange equations (6.2) with $\phi_i = \Phi$ then gives:

$$\nabla \cdot \nabla \Phi = 4\pi G \rho \tag{6.4}$$

which is the familiar Poisson equation.

Similarly, we can write down a Lagrangian density for general relativity (first derived by Hilbert):

$$\mathcal{L}_{GR} = \left[\frac{c^4}{16\pi G}(R - 2\Lambda) + \mathcal{L}_M\right]\sqrt{-g} \tag{6.5}$$

where R is the Ricci scalar, Λ is the cosmological constant, $\mathcal{L}_M$ is the Lagrangian density describing the matter field, and $g = \det(g_{\mu\nu})$ is the determinant of the metric. The Einstein field equations then follow by application of the Euler-Lagrange equations with $\phi_i = g_{\mu\nu}$[1].

As we noted in §3, the Einstein field equations are just one possibility. Once we allow more complex theories, there are an infinity of possibilities that can only be constrained by empirical data from the real Universe (at least for now). As an example, we consider one such extension of the field equations here: TeVeS (Tensor, Vector, Scalar) gravity (Bekenstein 2004). However, there are many other possible forms of modified gravity (e.g. Moffat 2005; Moffat 2006). We will write down the full Lagrangian for TeVeS, but will then consider mainly the non-relativistic *weak field limit*. The full theory is important for cosmological probes that we discuss in later lectures, it also plays a role in lensing since, although the bending angle formula requires only the weak-field limit, the cosmological theory is required to turn observed doppler shifts (redshift) into distance. However, the non-relativistic weak field limit is sufficient for galaxy and galaxy cluster dynamics where dark matter was first discovered (see §2). In this limit, we will show that we can generalise to a broad class of theories that may then be simultaneously tested.

6.3 Tensor Vector Scalar gravity (TeVeS) and MOND

TeVeS is a relativistic version of an older non-relativistic modified gravity theory: MOdified Newtonian Dynamics (MOND; Milgrom 1983; Bekenstein and Milgrom 1984). The relativistic extension is required to study gravitational lensing and cosmology in modified gravity, but rotation curves and the dynamics of galaxies can be understood in the much simpler non-relativistic limit. We will mostly consider the latter limit, but discuss here briefly the full theory.

6.3.1 The relativistic theory

The basic idea with TeVeS is to introduce *two metrics* – hence it is called a *bi-metric* theory. The first metric, $g_{\mu\nu}$ behaves as in standard GR. It is responsible for raising and lowering indices, for example: $A^\mu = g^{\mu\nu}A_\nu$. It is also the metric that appears in the geodesic equation (see §3):

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{\alpha\beta}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = 0 \tag{6.6}$$

where

$$\Gamma^\alpha_{\lambda\mu} = \frac{1}{2}g^{\alpha\nu}\left(\frac{\partial g_{\mu\nu}}{\partial x^\lambda} + \frac{\partial g_{\lambda\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\lambda}}{\partial x^\nu}\right) \tag{6.7}$$

is the *Christoffel* symbol that is a function of first derivatives of the metric $g_{\mu\nu}$.

So far, we are identical to standard GR. Now, we introduce a second metric $\tilde{g}_{\mu\nu}$ that appears in the field equations:

$$\tilde{R}^{\mu\nu} - \frac{1}{2}\tilde{g}^{\mu\nu}\tilde{R} = \frac{8\pi G}{c^4}T^{\mu\nu} \tag{6.8}$$

where we have dropped the cosmological constant Λ (though it is straightforward to include it if required), and the tilde symbol reminds us that the Riemann tensor is now a function of second derivatives of $\tilde{g}_{\mu\nu}$ – our new metric.

To understand what the above achieves for us, it is worth considering the Newtonian weak field limit. Recall from §3, that this occurs when:

[1] For a derivation of this, see `http://preposterousuniverse.com/grnotes/`.

1. Objects move slowly, i.e. $\frac{1}{c}\frac{dx^\mu}{d\tau} \ll \frac{dt}{d\tau}$.
2. Gravity is weak such that spacetime is very close to Minkowski. In this case, we may write the metric as Minkowski plus some small perturbation, $h_{\mu\nu}$:

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \tag{6.9}$$

3. The metric is static and not a function of time.

In TeVeS, the geodesic equation behaves just as in standard GR, so must share the same Newtonian weak field approximation:

$$\frac{d^2\mathbf{x}}{dt^2} = -\frac{c^2}{2}\nabla h_{00} = -\nabla\Phi \tag{6.10}$$

The field equations must also behave similarly, but involve our new metric $\tilde{g}_{\mu\nu}$. The left hand side becomes:

$$\tilde{R}_{00} - \frac{1}{2}\tilde{R}\tilde{g}_{00} = \frac{2\nabla^2\tilde{\Phi}}{c^2} \tag{6.11}$$

while the right hand side gives (as for standard GR):

$$\frac{8\pi G T_{00}}{c^4} = \frac{8\pi G\rho}{c^2} \tag{6.12}$$

Thus we derive a Poisson equation in a different static potential $\tilde{\Phi}$:

$$\nabla^2\tilde{\Phi} = 4\pi G\rho \tag{6.13}$$

We must now relate this new potential to the one that gives rise to the forces. In other words, we must specify a relation between $\nabla\Phi$ and $\nabla\tilde{\Phi}$. This means specifying some relationship between our two metrics $g_{\mu\nu}$ and $\tilde{g}_{\mu\nu}$. In general, we can expect that they are interrelated by some function f:

$$f\nabla\Phi = \nabla\tilde{\Phi} \tag{6.14}$$

which gives a modified Poisson equation:

$$\nabla\cdot(f\nabla\Phi) = 4\pi G\rho \tag{6.15}$$

This is the weak-field gravitational field equation for TeVeS also called MOdified Newtonian Dynamics (MOND).

The above is fine for the weak-field limit, but for a full general relativistic theory, we must specify how $g_{\mu\nu}$ and $\tilde{g}_{\mu\nu}$ interrelate (and thus how to determine f). There is quite some freedom in how to do this. In TeVeS, the modified metric $\tilde{g}_{\mu\nu}$ is related to the usual GR metric $g_{\mu\nu}$ via scalar and tensor fields (hence the name *Te*nsor *Ve*ctor *S*calar):

$$\tilde{g}_{\mu\nu} = e^{-2\phi}g_{\mu\nu} - 2U_\mu U_\nu \sinh(2\phi) \tag{6.16}$$

where ϕ and U_μ are a new scalar and vector field, respectively. The vector field satisfies the relation:

$$U^\mu U_\mu = -1 \tag{6.17}$$

These two new fields also evolve according to their own Lagrangians:

$$\mathcal{L}_U = \frac{c^4}{32\pi G}\sqrt{-g}\left[KA_{\mu\nu}A^{\mu\nu} - 2\lambda(U_\mu U^\mu + 1)\right] \tag{6.18}$$

where $A_{\mu\nu} = \nabla_\mu U_\nu - \nabla_\nu U_\mu$, λ is a Lagrange multiplier that ensures that U^μ satisfies equation 6.17[2], and:

[2] A derivation of λ is given in Bekenstein 2004. Recall that Lagrange multipliers allow us to apply constraints when trying to find extrema. Here we are finding the extremal path subject to the Lagrangian L_{TS} and the constraint given by equation 6.17.

$$\mathcal{L}_\phi = \frac{c^4 F_0}{8\pi G l^2}\sqrt{-g}F(l^2\sigma) \tag{6.19}$$

where F_0 and K are constants, $\sigma = (\tilde{g}^{\mu\nu} - U^\mu U^\nu)\nabla_\mu\phi\nabla_\nu\phi$, l is some length scale, and F is a free function designed to interpolate between the Newtonian and MOND weak-field regimes.

Thus, the full TeVeS Lagrangian density is given by:

$$\mathcal{L}_{TS} = \mathcal{L}_{TSG} + \mathcal{L}_U + \mathcal{L}_\phi \tag{6.20}$$

where $\mathcal{L}_U$ and $\mathcal{L}_\phi$ are as above, and $\mathcal{L}_{TSG}$ is given by the standard GR Lagrangian density, but with $g_{\mu\nu} \to \tilde{g}_{\mu\nu}$. In the absence of matter terms or cosmological constant, this is given by:

$$\mathcal{L}_{TSG} = \frac{c^4}{16\pi G}\sqrt{-\tilde{g}}\tilde{R} \tag{6.21}$$

TeVeS approaches standard general relativity when the two metrics become equal $\tilde{g}_{\mu\nu} \to g_{\mu\nu}$. In this case, from equation 6.16, we have that $\phi \to$ const. $= 0$, which occurs for $K = F_0 \to 0$.

We have already given an outline of the weak field limit in TeVeS, but not derived how the function f is expressed in terms of the new fields ϕ and U^μ. However, the full derivation is a bit involved for us here and does not really gain us much since in the end, the function f is really just chosen to match data for nearby disc galaxy rotation curves (as we shall see shortly). For a derivation of the weak-field limit, gravitational lensing and cosmology in TeVeS, we refer the interested reader to Bekenstein 2004.

Phew! the above looks rather complicated. TeVeS's complex structure is, however, *designed* to: (i) approach a weak field modified gravity that can explain the dark matter; (ii) not violate known GR Solar system tests; and (iii) maintain manifest covariance and causality. These combined constraints make it difficult to construct a workable theory, but it is possible.

6.3.2 The weak field limit

In the weak field, TeVeS becomes much more manageable and we will study the theory mainly in this limit. In this case, the Lagrangian density becomes:

$$\mathcal{L}_{\mathrm{MOND}} = -\frac{a_0^2}{8\pi G}F(\frac{|\nabla\Phi|^2}{a_0^2}) - \rho\Phi \tag{6.22}$$

where a_0 is an acceleration parameter and F is a free function that interpolates between the Newtonian and MOND regimes. Applying the Euler-Lagrange equations, we recover the MOND field equations:

$$-\nabla\cdot\left(\frac{a_0}{8\pi G}\frac{dF(x^2)}{dx^2}2\mathbf{x}\right) + \rho = 0 \tag{6.23}$$

$\Rightarrow$

$$\nabla\cdot(f\nabla\Phi) = 4\pi G\rho \tag{6.24}$$

where $x = |\nabla\Phi|/a_0$ and $f(x) = \frac{dF(x^2)}{dx^2}$.

Thus, MOND gravity is described by a modified *non-linear* Poisson equation. Typically, f is assumed to take the form:

$$f(x) = x(1+x^2)^{-1/2} \tag{6.25}$$

which – as we shall see shortly – is *designed* to match rotation curves for nearby spiral and disc galaxies without the need for dark matter.

Note that this modified Poisson equation is difficult to solve analytically. It is straightforward to show that substituting $\Phi \to \Phi_1 + \Phi_2$ does not give $\rho \to \rho_1 + \rho_2$. This means that solutions cannot be superposed as in normal Newtonian mechanics. Indeed much of our Newtonian intuition must be discarded as you will discover on the problem sheet. Every mass configuration will have its own unique potential which should be determined by (numerically) inverting equation 6.24.

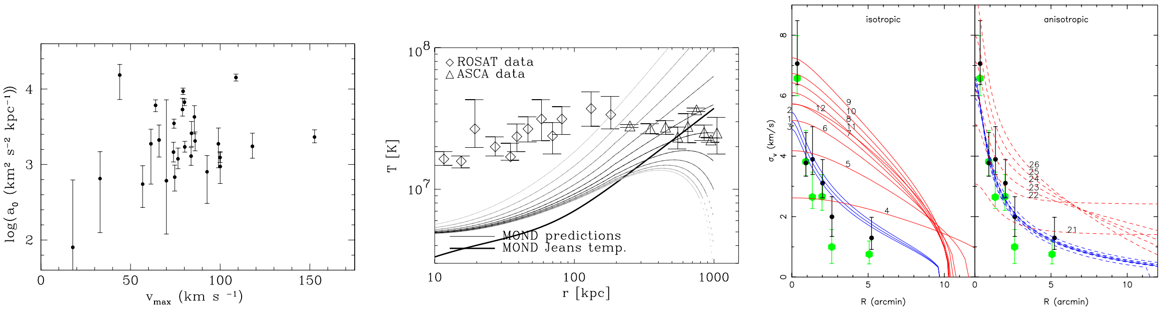


Figure 6.1: **Left:** Fits of the MOND acceleration parameter a_0 to dwarf and low surface brightness (LSB) galaxies. The typical value found is $a_0 \sim 10^{-10}$m s^{-2}. **Middle:** derived MOND temperature profiles for X-ray emitting gas in the Virgo cluster (lines) as compared to data (points). **Right:** The measured and predicted velocity dispersion profile for the globular cluster NGC2419 using Newtonian gravity (blue lines) and MOND (red lines). At a distance of $d \sim 100$ kpc, 5 arcmin corresponds to ~ 140 pc. The black and green data points show different velocity outlier cuts. The left panel shows results for isotropic models (where the dispersion is the same in all directions). The right panel shows results for a class of anisotropic models where orbits become increasingly radially anisotropic towards the edge of the cluster.

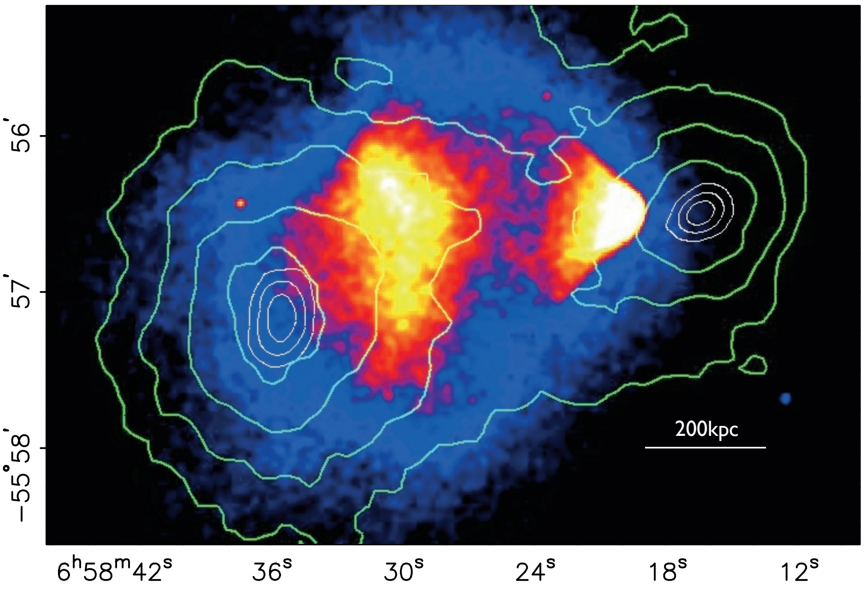


Figure 6.2: The 'bullet' cluster: a merging pair of galaxy clusters. Shown in green are isodensity contours derived from a *weak gravitational lensing* mass map of the cluster. The two peaks are centred on the observed distribution of galaxies in the two clusters (not shown). However, most of the baryonic mass is actually in the X-ray emitting gas, shown in the white through blue shaded contours. This hydrogen plasma, being collisional, has been stripped away from the galaxy cluster centres by the collision. The image is strong evidence that dark matter moves like the galaxies and not like their gas: dark matter is a collisionless fluid.

6.3.3 Rotation curves in MOND

The MOND theory is *designed* to match observed data from rotation curves (Milgrom 1983). Although hard to solve in general, in spherical symmetry, the MOND-Poisson equation can be solved analytically. To prove this, consider the substitution $f\nabla\Phi = \nabla\Phi_N + \nabla \times \mathbf{h}$, where $\mathbf{h}$ is some vector field and Φ_N is a *Newtonian* potential that satisfies the usual Poisson equation. Let us substitute this into the MOND-Poisson equation:

$$\nabla \cdot (f\nabla\Phi) = \nabla \cdot \nabla\Phi_N + \nabla \cdot \nabla \times \mathbf{h} \tag{6.26}$$

In spherical symmetry, $\mathbf{h} \propto \hat{\mathbf{r}}$ and therefore $\nabla \cdot \nabla \times \mathbf{h} = 0$. Thus:

$$f\nabla\Phi = \nabla\Phi_N \tag{6.27}$$

The above means that in spherical symmetry the MOND force must point in the *same direction* as the Newtonian force. Thus, we may write:

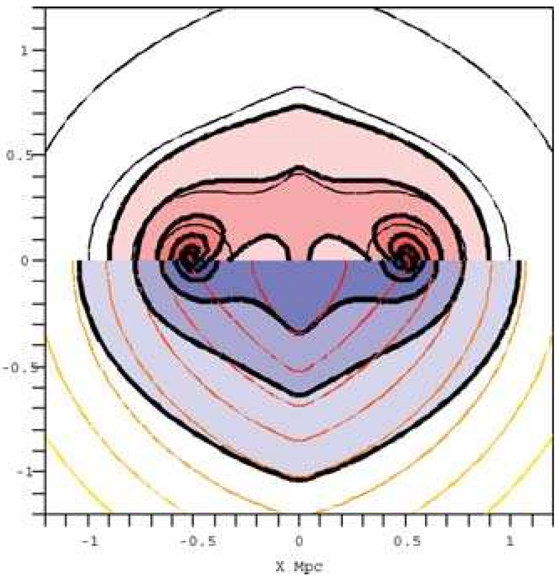


Figure 6.3: The visible matter in TeVeS/MOND can be peaked where the gravitational potential is not. This plot shows a toy model system that has a smooth ∼ double peaked potential shown by the thin red contours (bottom half), but a *triple peaked* baryonic distribution (derived from the MOND-Poisson equation assuming the classical limit; thick black lines and red shaded regions, top half). The thick black lines and blue contours in the bottom half show the lensing map for this system that is also double peaked. The thin black lines in the top half show the Newtonian baryonic distribution for this potential (derived from the normal Poisson equation) that is double peaked.

$$\mathbf{g} = \mathbf{g}_N \frac{g}{g_N} \tag{6.28}$$

where $\mathbf{g} = \nabla\Phi$, $g = |\mathbf{g}|$, and similar.

Taking the modulus of both sides of equation 6.27 and substituting for $f(g/a_0)$, we obtain the following quadratic equation:

$$g^4 - g^2 g_N^2 - g_N^2 a_0^2 = 0 \tag{6.29}$$

which solving (taking the physical positive root) gives:

$$\mathbf{g} = \mathbf{g}_N \frac{\left[1 + \sqrt{1 + 4a_0^2/g_N^2}\right]^{1/2}}{\sqrt{2}} \tag{6.30}$$

Thus, in spherical symmetry, MOND becomes straightforward to solve. We may solve the usual Poisson equation to obtain $\mathbf{g_N}$ from some mass distribution ρ. Equation 6.30 then tells us how this force is *modified* in MOND.

Let us consider some interesting limits from equation 6.30. First, for large accelerations, $g_N \gg a_0$, we have that $\mathbf{g} \to \mathbf{g}_N$ and we return to normal Newtonian mechanics. MOND effects appear only at *low acceleration*. Secondly, for $g_N \ll a_0$, we have that $\mathbf{g} \to \mathbf{g}_N\sqrt{a_0/g_N}$. Let us consider what effect this has on rotation curves in galaxies. For spherical symmetry, the gravitational force is balanced by the centripetal force:

$$m\frac{v_c^2}{r} = mg \tag{6.31}$$

At small radii, the rotation curve will agree with the Newtonian case because the accelerations are large. At large radii, where the accelerations drop to zero, we have $g \simeq m g_N\sqrt{a_0/g_N}$ and substituting $g_N = GM(r)/r^2$, we derive:

$$v_c^2 = \sqrt{GM(r)a_0} \tag{6.32}$$

Thus, at large radii, where the mass becomes constant, the rotation curve will become constant and flat. Indeed, MOND is specifically *designed* to achieve this.

In the deep-MOND regime, MOND has just one free parameter in this weak field limit: a_0. Figure 6.1 (left) shows some recent fits of a_0 to dwarf and low surface brightness (LSB) galaxy rotation curves (Swaters *et al.* 2010). The typical value found is $a_0 \sim 10^{-10}$m s^{-2}. However, already the results are not encouraging because there is a wide scatter in a_0 and possibly even a correlation with $v_{\rm max}$ as was reported in earlier work by Lake 1989 (note that this Figure has a *logarithmic* scale for the y-axis). The sympathetic MONDite would argue, however, that systematics in the observations and modelling of rotation curves could plausibly explain this scatter. Things get worse, however, when we move away from the disc galaxies which MOND was designed to fit. On galaxy cluster scales Aguirre *et al.* 2001 show that the observed temperature profiles of X-ray gas disagree with expectations from MOND (see Figure 6.1, middle). Sanders 2003 propose adding 'dark matter' to solve this problem, but then if we must have dark matter *and* alternative gravity, the motivation for modifying gravity appears somewhat diminished. Angus *et al.* 2007 argue that this dark matter could be familiar massive neutrinos. But this idea has now been ruled out Natarajan and Zhao 2008. On small scales, MOND fares poorly also. Globular clusters (GCs) are massive star clusters that orbit within the Milky Way. These are not thought to contain dark matter, but some are still in the deep MOND regime with low accelerations. Thus, they make for an interesting test of MOND because MOND demands that we see 'dark matter' like effects in the outskirts of these clusters. Ibata *et al.* 2011 have recently used this idea for the GC NGC2419 that lies ~ 100 kpc from the Milky Way. This cluster is particularly special because it is massive ($M_{cl} \sim 10^6\,$M$_\odot$), distant ($d \sim 100$ kpc) and unusually large ($r_h \sim 20$ pc) for a GC. At $d \sim 100$ kpc from the Milky Way, the acceleration due to our Galaxy is $a_{\rm gal} \sim GM_{\rm gal}/d^2$, which for $M_{\rm gal} \sim 10^{12}\,M_\odot$ gives $a_{\rm gal} \sim 10^{-11}$m s^{-2} – an order of magnitude smaller than the MOND scale a_0. Thus, we may safely ignore the Galactic contribution to the potential (that would otherwise complicate the analysis). Furthermore, for stars at the half light radius within the cluster (r_h), the acceleration scale is $a_{1/2} \sim GM_{cl}/r_h^2 \sim 3 \times 10^{-10}$m s^{-2}. Thus stars at the outskirts of the cluster will be in the deep-MOND regime. Figure 6.1 (right) shows some mass models for NGC2419 assuming Newtonian gravity (blue) and MONDian gravity (red). Again, MOND fares poorly here struggling to reproduce the observed kinematics. By contrast Newtonian gravity gives an excellent fit.

Unfortunately, MOND appears to only succeed in the disc galaxies for which it was designed to succeed. The mark of a good theory is that it performs well beyond the regime in which it was proposed. MOND does not appear to pass this test. This does not mean, however, that *all* alternative gravity theories are ruled out. But we must now look to alternative alternative gravity theories.

6.4 A generalised weak-field alternative gravity theory

Given the enormous possibility for creating new gravity theories, ruling out one at a time seems like a never-ending task. More satisfying would be to be able to rule out broad classes of theory simultaneously. Or better still, to rule out all such theories once and for all. There is a neat route to achieving this in the weak field non-relativistic limit.

MOND in the weak field is simply a modified Poisson equation. In fact, this must be true for *any scalar gravity theory*, since the force must be the gradient of some potential Φ. In general, then, we may write:

$$\mathbf{O} \cdot \nabla\Phi = \rho \tag{6.33}$$

where $\mathbf{O}$ is some *operator*. In Newtonian mechanics, we have $\mathbf{O} = \frac{\nabla}{4\pi G}$, in MOND $\mathbf{O} = \frac{\nabla f}{4\pi G}$. To give another example, Moffat 2006 propose a modified gravity theory (MOG) where Newton's gravitational constant G becomes a function of space and time. In this case, in the weak field, we have $\mathbf{O} = \frac{\nabla}{4\pi G(\mathbf{r},t)}$.

Common to all such modified gravity theories is that the mapping between Φ and ρ – however complex – must be *symmetry preserving*. A spherical distribution of mass *must* have a spherical potential (indeed, we saw this already for MOND). Similarly, a flattened mass distribution like our own Milky Way disc Galaxy must then have a flattened gravitational potential. In relatively recent work, we tried to use such arguments to rule out MOND (Read and Moore 2005), and you will explore this further on the problem sheet. In fact, such arguments are completely general and test any weak-field modified gravity that purports to explain all of the missing matter phenomenon. The slight snag, as you will see, is that non-linear Poisson equations often behave in entirely counter-intuitive ways ...

6.5 Lensing and cosmology constraints on alternative gravity

So far we have discussed only classical dynamics constraints on alternative gravity. There is a good reason for this. Moving away from the non-relativistic weak field limit, we must specify a full general relativistic alternative gravity theory like TeVeS, and we must be able to make detailed calculations of distances and bending angles in this theory. Such a calculation has been done for gravitational lensing in TeVeS, assuming a simple lens geometry (Zhao *et al.* 2006). They find that many lenses can be reasonably fit by the theory, but there are some significant outliers that hint at problems with the theory similar to those we have already discussed above.

Perhaps the most famous non-classical test of alternative gravity models is the 'bullet' cluster – a merging pair of galaxy clusters (Clowe *et al.* 2006). The key result is shown in Figure 6.2. Shown in green are isodensity contours derived from a *weak gravitational lensing* mass map of the cluster. The two peaks are centred on the observed distribution of galaxies in the two clusters (not shown). However, most of the baryonic mass is actually in the X-ray emitting gas, shown in the white through blue shaded contours. This hydrogen plasma, being collisional, has been stripped away from the galaxy cluster centres by the collision. If dark matter is really just alternative gravity, we might expect the gravitational field to be peaked where the visible matter is peaked – i.e. on top of the observed X-ray gas peaks. However, the potential appears to peak instead where the galaxies are. This suggests that whatever dark matter is, it moves like the galaxies do as a *collisionless fluid*[3] and not as the gas does which is a collisional fluid. Most astronomers take this as strong evidence that dark matter cannot be explained by alternative gravity theories. Indeed, as we shall see in later lectures, assuming that dark matter is indeed to a good approximation a collisionless fluid, we can calculate its expected distribution in the Universe. This matches very well observations over a very large range of scales from the cosmic microwave background radiation to galaxy clustering in the nearby Universe.

The problem with the 'bullet', however, is that it relies on the weak lensing mass map which is derived using standard general relativity. There remains a niggling doubt, then, that in some alternative gravity theory things might look different. In particular, our neat symmetry arguments in §6.4 no longer apply since the distribution of galaxies and X-ray gas in the bullet is clearly very complex. Worse still, this makes mass modelling in alternative gravity theories very difficult. The simple lens models presented in Zhao *et al.* 2006 for TeVeS, for example, will no longer be adequate. Furthermore, our work then becomes seemingly never-ending since each new alternative gravity theory will have to be tested in detail against the bullet and other similar systems.

The complexity of modelling the bullet in new gravity theories has meant that to date only idealised toy models have been attempted. Angus *et al.* 2006 demonstrate that in MOND it is possible to have the peak in the baryonic mass offset from the peak in the gravitational potential (see Figure 6.3). The simplest way to demonstrate this is to specify a smooth potential Φ and then derive the baryonic distribution associated with this using the MONDian Poisson equation (6.24). This is straightforward since ρ follows from Φ by simple differentiation. It is going the other way round and modelling the relativistic theory correctly (required to actually fit the bullet cluster) that is difficult. Consider the following MOND toy potential:

$$\Phi_{tc}(x,y,z) = [k_1 + (1-k_1-k_2)H(x)]\,\Phi(r_1) + [k_2 + (1-k_1-k_2)H(-x)]\,\Phi(r_2) \tag{6.34}$$

where $H(x)$ is the Heaviside step function:

$$H(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases} \tag{6.35}$$

and $\Phi(r_1)$, $\Phi(r_2)$ are spherical potentials centred on r_1 and r_2, respectively. The Heaviside functions create a thin disc at $z = 0$; the other potentials peak at r_1 and r_2.

Figure 6.3 shows the density distribution derived from the above potential assuming $k_1 = k_2 = 0.2$ in MOND (i.e. using equation 6.24). The potential for this choice of parameters is smooth and $\sim$ double peaked (shown by the thin red contours; bottom half), but the baryonic distribution is *triple peaked* (thick black lines and red shaded regions; top half). The thick black lines and blue contours in the bottom half show the lensing map for this system that is also double peaked. The thin black lines

[3] Recall the definition of a collisionless fluid was given in §1.

in the top half show the Newtonian baryonic distribution for this potential (derived from the normal Poisson equation) that is double peaked.

Although MOND/TeVeS has many problems on all scales from GCs to clusters (see Figure 6.1), the above argument due to Angus *et al.* 2006 is a proof of concept that seeing offsets between light and gravity does not necessarily rule out alternative gravity theories. Similarly counter-intuitive results were found by Brownstein and Moffat 2007 for the MOG theory. There, the gravitational constant G is a function of space and time. Allowing the 'MOG centre' where $G = G_N$ (the Newtonian value) to vary, they could also reproduce something like the bullet cluster results.

6.6 Some final musings

Alternative gravity as an explanation for dark matter is embattled, but hard to definitively rule out. The main alternate theory for the past two decades has been Milgrom/Bekenstein's MOND/TeVeS. This is now disfavoured by kinematic observations of Globular Clusters, spiral galaxies and galaxy clusters. But other theories are arriving to fill the void, like Moffat's MOG theory. Seemingly clean tests for alternative gravity like the 'bullet' cluster are often less constraining than they first appear because non-linearities in alternative gravity theories can lead to counter-intuitive results (like the peak of the baryonic matter distribution being off-set from the peak of the gravitational potential). The bullet cluster is, however, *beautifully* explained by a collisionless fluid dark matter. This not only fits its instantaneous gravitational potential, but helps us understand how it came to look the way it does. A collisionless fluid moves like the galaxies and so the current observed lensing map is expected as a result of a cluster-cluster gas rich merger. Furthermore, as a result of ever-improving data, alternative gravity theories now require more and more degrees of freedom to match the observations. MOND has just one parameter in the weak field which makes it possible to rule out the theory. MOG appears more successful at fitting the observations, but has many more degrees of freedom in how G varies. In principle, there is nothing to stop us from simply building ever more complex alternative gravity theories with more and more free parameters until all observations are explained. Such a theory should be viewed with caution, however, because of its lack of predictive power. As we will see in the following lectures, assuming that dark matter is a collisionless fluid, we can make detailed *predictions* for its distribution in the Universe that really do match observations remarkably well.

Ultimately, it would be good to have a clean test that rules out alternative gravity once and for all. Such a test is possible in the weak field based purely on symmetry arguments, but a super-clean observational system in which to present such a test has not yet been found. Happily, there is a key test of alternative gravity that we have not yet discussed: the Cosmic Microwave Background radiation (Skordis *et al.* 2006; Dodelson 2011). As we shall see, this is probably the most challenging test for alternative gravity theories found to date, but it requires us to first understand the standard cosmological model. We discuss this, next.

Lecture 7

Cosmological probes of dark matter I: The homogeneous Universe

In this lecture, we lay the groundwork for studying cosmological probes of dark matter: the cosmic microwave background radiation, big bang nucleosynthesis and large scale structure. We argue that the Universe on large scales is isotropic and homogeneous, and we present and study the Friedmann equations that describe a such a Universe.

7.1 The homogeneous Universe

Building a cosmological model means finding a solution to Einstein's field equations that gives a good description of the distribution of matter and energy in the Universe both now and backwards in time. Fitting this model to a broad range of observables gives us information about the *composition* of the Universe, and hence about dark matter. If we knew nothing about the observed distribution, we might start with the simplest assumption (rather like we did when guessing the form of Einstein's field equations in §3): a homogeneous and isotropic Universe. In fact, this assumption agrees very well with the data as we shall see. For now, we can satisfy ourselves that the Universe is certainly quite close to homogeneous today on large scales, as can be seen from galaxy surveys like the Sloan Digital Sky Survey (SDSS; Figure 7.1; Yadav *et al.* 2005).

A more theoretical argument for homogeneity comes from the observed local expansion. Lemaitre (1927) and later Hubble (1929) found that nearby galaxies are all receding from us with a velocity proportional to distance (Hubble 1929; Nussbaumer and Bieri 2011; and Figure 7.2). The *Copernican Principle* states that there is nothing special about our place in the Universe. We may then reasonably

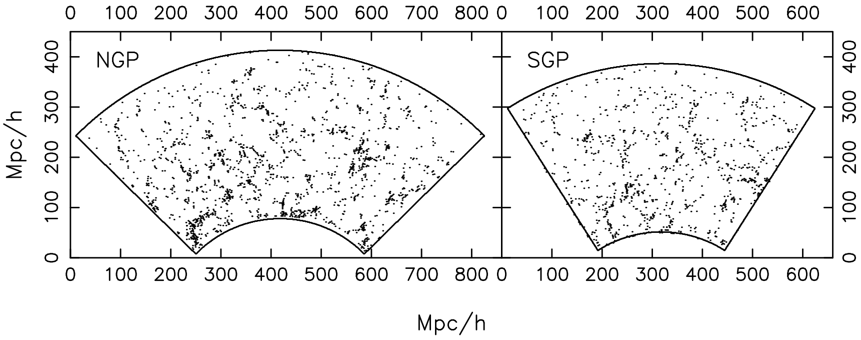


Figure 7.1: The Universe today on large scales is observed to be very homogeneous. This image shows the distribution of galaxies in the SDSS galaxy survey towards the North Galactic Pole (left) and the South Galactic Pole (right). The distribution becomes statistically homogeneous on scales larger than $\sim 70\,\mathrm{Mpc}$.

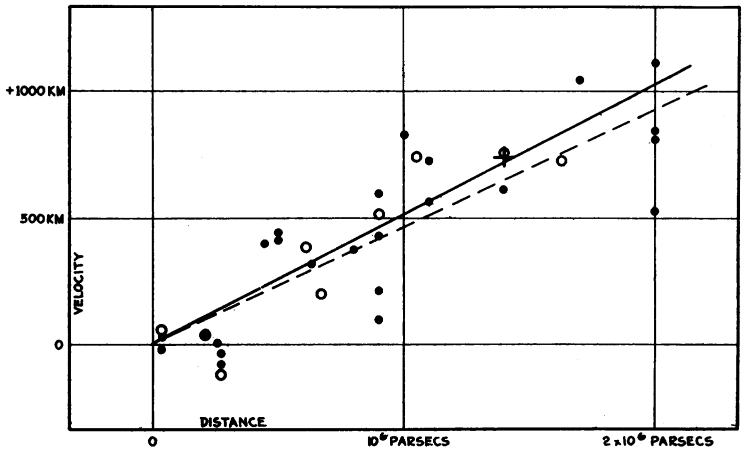


FIGURE 1

Velocity-Distance Relation among Extra-Galactic Nebulae.

Radial velocities, corrected for solar motion, are plotted against distances estimated from involved stars and mean luminosities of nebulae in a cluster. The black discs and full line represent the solution for solar motion using the nebulae individually; the circles and broken line represent the solution combining the nebulae into groups; the cross represents the mean velocity corresponding to the mean distance of 22 nebulae whose distances could not be estimated individually.

Figure 7.2: Hubble's original data showing that the Universe is expanding with velocity proportional to distance.

assume that *all* observers must see the Universe expanding away from them, which implies that the Universe is *isotropic*. But, if the Universe is isotropic then it *must* be homogeneous. The proof is geometrical and given in Figure 7.3 (argument taken from Peacock 1999). The converse is, however, not true. A homogeneous Universe can be anisotropic (can you think of an example?).

Note that we could have realised much sooner than either Hubble or Friedmann the Universe is expanding due to a paradox commonly attributed to the German amateur astronomer Olbers in 1823, but in fact dating back much earlier than him to Thomas Digges in the late 1500's (Harrison 1989). The paradox is as follows: if the Universe were static and infinite then we would see a star along ever single light line in the night sky. Thus the photons arriving from each of these stars would light up the night sky until it were as bright as the Sun. Edgar Allen Poe put it very well in a poem, *Eureka* (1848) in which he wrote:

> *"Were the succession of stars endless, then the background of the sky would present us a uniform luminosity, like that displayed by the Galaxy – since there could be absolutely no point, in all that background, at which would not exist a star. The only mode, therefore, in which, under such a state of affairs, we could comprehend the voids which our telescopes find in innumerable directions, would be by supposing the distance of the invisible background so immense that no ray from it has yet been able to reach us at all."*

The fact that the sky at night is dark is indeed a compelling mystery! It can be understood, however, if the Universe is expanding. In this case, the light from distant stars are *redshifted* (more so with increasing distance). Infinitely distant stars will be redshifted infinitely until they can no longer be seen.

7.2 The FLRW metric

In §3, we wrote down the metric that describes an isotropic and homogeneous Universe, the FLRW metric:

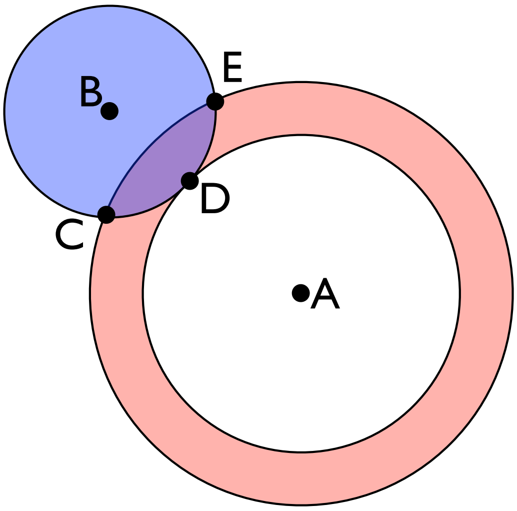


Figure 7.3: A geometric proof that an isotropic Universe is homogeneous. The converse is not necessarily true. Isotropy about point B tells us that the density at C, D and E is the same. By expanding spheres of different radii around point A, we see that the overlapping purple shaded area must be homogeneous. For large enough shells, we may extend this argument to the whole Universe.

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t)\left(\frac{dr^2}{1-kr^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2)\right) \tag{7.1}$$

where $R(t)$ is called the *scale factor*; $k = [-1, 0, 1]$ is a parameter that measures the fundamental *curvature* of the spacetime; and r is a time independent *co-moving* coordinate. We can see that k describes curvature by considering $k = 0$. In this case, the FLRW metric looks very similar to Minkowski space, just with an expansion factor $R(t)$. Thus, $k = 0$ is often called 'flat space', even though it still has some spacetime curvature. $k = \pm 1$ are called *closed* and *open* Universes respectively, which will describe in more detail shortly.

Since distant galaxies are observed to all be moving away from us with ever greater speeds, we will mostly be dealing with pure radial motion in the FLRW metric. For this reason, it is useful to transform to a different coordinate system that eliminates the $1 - kr^2$ term in the denominator of the radial part. Consider the function:

$$S_k(r') = \begin{cases} \sin r' & k = +1 \\ \sinh r' & k = -1 \\ r' & k = 0 \end{cases} \tag{7.2}$$

Now, we have that (taking $k = 1$ as an example):

$$\frac{dr^2}{1-kr^2} \to \frac{dS_k^2}{1-kS_k^2} = \frac{\cos^2 r' dr'^2}{1-\sin^2 r'} = dr'^2 \tag{7.3}$$

and similarly for $k = -1$ and $k = 0$. Thus, the metric becomes:

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t)\left(dr'^2 + S_k(r')^2 d\psi^2\right) \tag{7.4}$$

where $d\psi^2 = (d\theta^2 + \sin^2\theta d\phi^2)$. We will use the above metric with the notation $r = r'$ from here on.

Since the Universe is observed to be homogeneous, and the Copernican Principle suggests that it must be isotropic, the FLRW metric, and perturbations around it, form the basis of our current cosmological model. Straight away this makes an important *prediction*. The scale factor $R(t)$ acts to cause either an expansion or contraction of the length scales in the metric as a function of time: the FLRW metric describes either expanding or collapsing Universes. Initially this worried Einstein who introduced the cosmological constant to try to counter-act the expansion term. But this static solution is unstable and Einstein later called it his greatest blunder (c.f §3). Now we can think of this instead as a beautiful prediction of general relativity and Einstein's field equations: an isotropic, homogeneous Universe must either expand or contract.

7.3 Dynamics of the expansion

In this section, we study *dynamics* in the FLRW metric. This means substituting the metric into Einstein's field equations to obtain the equations of motion (c.f. §3 and §4). We leave this as an exercise for the reader and quote the result: Friedmann's equation:

$$\dot{R}^2 - \frac{8\pi G}{3}\rho R^2 = -kc^2 \tag{7.5}$$

where k is the curvature as previously, c is the speed of light in a vacuum (as previously), and ρ is the density of matter and energy in the Universe. Note that this density contains *all* contributions including radiation and potentially the vacuum itself.

Note that we can 'derive' the above using a semi-Newtonian analogy where we demand simply conservation of energy:

$$\frac{1}{2}(\dot{R})^2 - \frac{GM}{R} = \text{const.} \tag{7.6}$$

which, substituting $M = \frac{4}{3}\pi R^3 \rho$ (valid because the FLRW Universe is homogeneous), gives Friedmann's equation (7.5).

The above may be a useful way to remember the equation, but the 'derivation' is dodgy in many ways and we favour here instead proper substitution of the FLRW metric into the field equations to derive the dynamics. It does give one useful insight, however. We can think of flat Universes ($k = 0$) as being just bound; open Universes ($k = -1$) are unbound and will expand forever; and closed Universes ($k = +1$) are over-bound and will eventually collapse.

The Friedmann equation tells us that there is an intimate link between the geometry of the Universe and its density. A flat Universe ($k = 0$) will follow by construction if the density has a critical value (the *critical density*):

$$\rho_c = \frac{3H^2}{8\pi G} \tag{7.7}$$

where $H = \dot{R}/R$ is called the *Hubble parameter*. It is useful then to define densities relative to ρ_c which defines the *density parameter*: $\Omega = \rho/\rho_c$. Note that H and therefore ρ_c *change with time*. Thus, Ω will also be time dependent. Its value at the present epoch is often denoted Ω_0. But it is so commonly used that the $_0$ is often dropped. To avoid confusion, we will refer to Ω at earlier times as $\Omega(t)$, explicitly expressing the time dependence.

As we stated previously, the density of the Universe will comprise matter, radiation, and vacuum contributions each of which have different equations of state. Thus, we may divide up $\Omega(t)$ into these separate contributions that all evolve with time:

$$\Omega(t) = \Omega_\Lambda(t) + \Omega_m(t) + \Omega_r(t) \tag{7.8}$$

However, we can expect each of these contributions to evolve *differently* with time according to their respective equations of state. The matter term should evolve as $\rho \propto R^{-3}$, the radiation term[1] as $\rho \propto R^{-4}$ and the vacuum term as $\rho \propto$ const.[2]. Thus, we may write:

$$\frac{8\pi G\rho}{3} = H^2\Omega(t) = H_0^2\left(\Omega_\Lambda + \Omega_m a^{-3} + \Omega_r a^{-4}\right) \tag{7.9}$$

where $a(t)$ is the dimensionless scale factor: $a(t) = R(t)/R_0$; $H_0 = \dot{R}_0/R_0$; R_0 defines the scale at the present epoch; and Ω_m is a constant defined also at the present epoch (and similarly for the other contributions to Ω).

Thus, we may now rewrite the Friedmann equation as:

[1] We can treat the radiation as a relativistic gas.

[2] This follows because the vacuum is described by Einstein's cosmological constant (c.f. §3).

$$H^2 = H_0^2 \left[\Omega_\Lambda + \Omega_m a^{-3} + \Omega_r a^{-4}\right] - c^2 k R^{-2} \tag{7.10}$$

and using $H_0^2(1-\Omega_0) = -c^2 k R_0^{-2}$, we derive:

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = H_0^2 \left[\Omega_\Lambda + \Omega_m a^{-3} + \Omega_r a^{-4} - (\Omega_0 - 1)a^{-2}\right] \tag{7.11}$$

where $\Omega_0 = \Omega_\Lambda + \Omega_m + \Omega_r$. Note that the Hubble constant at the present epoch (H_0) is often written in dimensionless form as:

$$h = \frac{H_0}{100\mathrm{km\,s^{-1}Mpc^{-1}}} \tag{7.12}$$

Equation 7.11 is the key equation in cosmology because it tells us how the scale factor $a(t)$ will evolve with time depending on the *composition* of the Universe. The first thing this gives us is the *age* of the Universe. Notice from the definition of the scale factor that $a = 1$ at the present time, while $a = 0$ at the 'beginning' when the Universe had zero size (presumably things break down at this point, but we should be OK until very close to that moment). Thus, we can simply use equation 7.11 to calculate the age of the Universe for a given set of cosmological parameters:

$$t_{\mathrm{univ}} = \int_0^1 \frac{dt}{da} da = \int_0^1 \frac{1}{\dot{a}} da \tag{7.13}$$

Secondly, notice that the age and the expansion rate depend on the composition of the Universe. Thus, by measuring the expansion rate backwards in time, and fitting the above formula we can actually measure the equation of state of the Universe. This is how cosmology will become a probe of dark matter. The Friedmann equation must in general be solved numerically, but there are a few analytic limits that are worth considering. We discuss these, next.

7.3.1 Interesting limits

Let us consider first some limiting cases of equation 7.11. First notice that if $\dot{a} = 0$ then we have a *turning point*: the Universe will stop expanding, turn around and re-collapse. This can only occur if:

$$\Omega_\Lambda + \Omega_m a^{-3} + \Omega_r a^{-4} = (\Omega_0 - 1)a^{-2} \tag{7.14}$$

Since a and all of the Ω's are positive, this is only possible if $\Omega_0 - 1$ is *positive*. Thus the Universe can re-collapse if $\Omega_0 > 1$ (i.e. $k = +1$). This is called a *closed Universe*. If $\Omega_0 < 1$ ($k = -1$) then we have an *open Universe* that will expand forever.

Another important limit occurs at *early times* ($a \to 0$). In these first moments, only the radiation term is important. Later, the matter term dominates, then the curvature. Finally at late times $a > 1$, the vacuum energy term (if Ω_Λ is sufficiently large) will dominate over all of the other terms.

7.3.2 A matter dominated Universe

We can now solve the Friedmann equation under different assumptions about the equation of state to derive the evolution of the scale factor $a(t)$. Let us consider first a matter dominated flat Universe with $\Omega_r = \Omega_\Lambda = k = 0$. In this case, equation 7.11 becomes:

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2 \Omega_m a^{-3} \tag{7.15}$$

which rearranging gives:

$$\int da a^{\frac{1}{2}} = \int H_0 \sqrt{\Omega_m} dt \tag{7.16}$$

and thus, $a \propto t^{2/3}$.

For non-flat Universes, things are a little more tricky but still analytic. Now we have:

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2\left[\Omega_m a^{-3} - (\Omega_0 - 1)a^{-2}\right] \tag{7.17}$$

At early times, the curvature is not important and we have $a \propto t^{2/3}$ as above. At late times it depends on whether the Universe is open or closed.

For a closed Universe, we may write $\kappa = \Omega_0 - 1 > 0$ and the solution is then the *cycloid solution*:

$$a = a_*(1 - \cos\alpha) \quad ; \quad t = t_*(\alpha - \sin\alpha) \tag{7.18}$$

which we can verify is a solution to equation 7.17 by substitution:

$$\dot{a} = \frac{da}{d\alpha}\left(\frac{dt}{d\alpha}\right)^{-1} = \frac{a_* \sin\alpha}{t_*(1 - \cos\alpha)} \tag{7.19}$$

Taking the square, using the trigonometric identity $\sin^2\alpha = (1 - \cos^2\alpha) = (1 - \cos\alpha)(1 + \cos\alpha)$, and substituting for $\cos\alpha = 1 - a/a_*$ gives:

$$\dot{a}^2 = \frac{a_*^2}{t_*^2}\left[2\frac{a_*}{a} - 1\right] \tag{7.20}$$

which is the matter dominated Friedmann equation if $2a_*^3/t_*^2 = H_0^2\Omega_m$ and $a_*^2/t_*^2 = H_0^2\kappa$. This latter is why the cycloid solution is only valid for $\kappa > 0$. Thus: $a_* = \Omega_m/(2\kappa)$; $t_* = \Omega_m/(2H_0\kappa^{3/2})$.

The above is called the cycloid solution because it describes the motion of a point on the surface of a circle as it rolls along. It is plotted in Figure 7.4 (black line).

For an open Universe, we may similarly write down a parametric solution. Now $\kappa < 0$ and we have:

$$a = a_*(\cosh\alpha - 1) \quad ; \quad t = t_*(\sinh\alpha - \alpha) \tag{7.21}$$

and substituting similarly to the above, we derive $a_* = \Omega_m/|\kappa|$ and $t_* = \Omega_m/(H_0|\kappa|^{3/2})$. This solution is plotted also in Figure 7.4 (blue line).

7.3.3 The eternal static Universe

We have already argued against a static Universe – at least an infinite one – based on observations that it is currently expanding and that the night sky is dark. But there is a good theoretical reason to reject such a Universe also. We may construct one only by having the vacuum energy exactly balance the expansion such that $\dot{a} = 0$ always. Thus occurs if:

$$\Omega_\Lambda = \Omega_m a^{-3} + \Omega_r a^{-4} - (\Omega_0 - 1)a^{-2} \tag{7.22}$$

which implies a *positive* cosmological constant with a funky equation of state (not the same as that derived in §3 assuming that it describes the vacuum), carefully tuned to balance the contributions from matter and radiation as a function of time. Any slight error in the cancellation and the Universe will either expand or contract. Such fine tuning is theoretically undesirable.

7.3.4 A radiation dominated Universe

This is the situation at very early times. Equation 7.11 becomes:

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2(\Omega_r a^{-4} - (\Omega_0 - 1)a^{-2}) \tag{7.23}$$

and since we consider early times, we may neglect the curvature term:

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2\Omega_r a^{-4} \tag{7.24}$$

which may be straightforwardly solved to give $a = \left(2H_0\sqrt{\Omega_r}t\right)^{1/2}$.

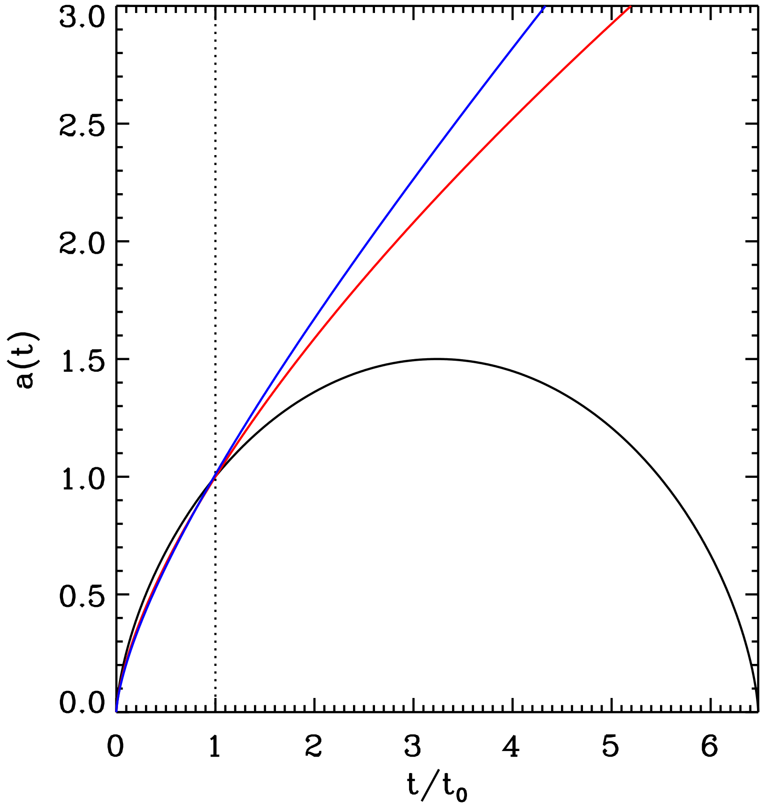


Figure 7.4: Solutions of the Friedmann equation (7.11) for a matter dominated Universe. The red line marks a flat Universe that scales as a power law with $a \propto t^{2/3}$. The black line marks the cycloid solution for a closed Universe. This is also the curve that describes the motion of a point on the surface of a circle as it rolls along – hence the name. The blue line marks the solution for an open Universe that will expand forever. The vertical dotted line marks the time t_0 today.

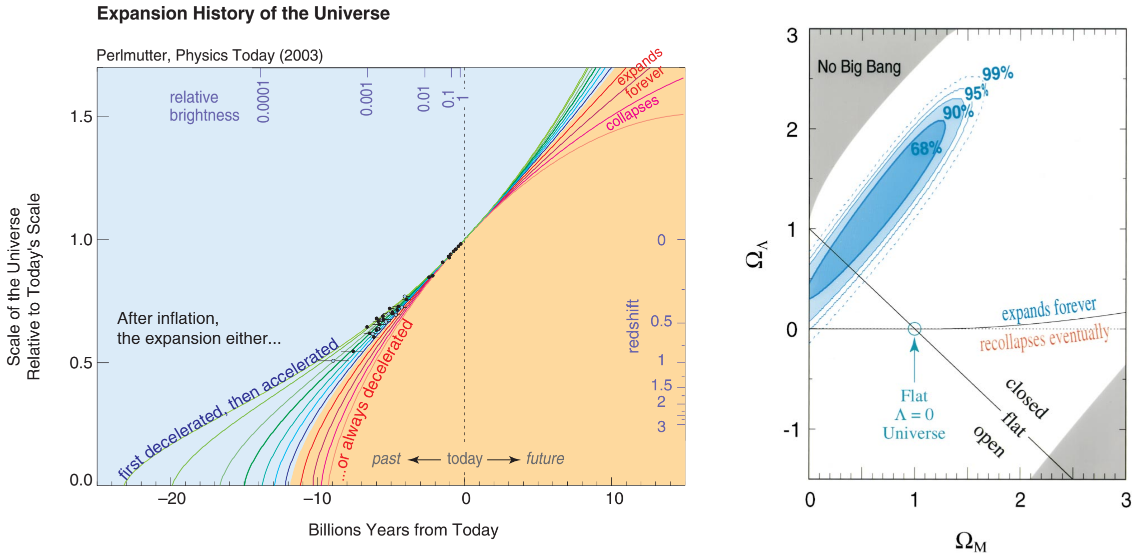


Figure 7.5: Solutions of the Friedmann equation for general Universes. The left panel shows the expansion factor a as a function of time for different cosmological models, as marked. The black data points show data from Type Ia supernova standard candles (more in this in the next lecture). The right panel shows which models best fit these data. The x-axis shows the matter density Ω_m, the y-axis the vacuum energy contribution Ω_Λ. Open, flat and closed Universes are marked. There is a small region where curvature and matter can beat the cosmological constant, but mostly the cosmological constant wins and causes the Universe to expand forever. The current data favour a model for our Universe with $\Omega_\Lambda \sim 0.7$ and $\Omega_m \sim 0.3$ suggesting that we will expand forever.

7.3.5 A vacuum dominated Universe

This situation must occur at late times (if there is a vacuum component). Neglecting curvature, equation 7.11 becomes:

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2 \Omega_\Lambda \tag{7.25}$$

and we derive an exponential expansion: $a \propto \exp(H_0\sqrt{\Omega_\Lambda}t)$. Thus the future for our Universe which does indeed appear to have significant Ω_Λ appears rather bleak. The timescales are sufficiently long, however, that we need not start worrying just yet.

A summary of solutions to the Friedmann equation is given in Figure 7.5 (Perlmutter *et al.* 1999). This encapsulates the limits we have just derived and everything inbetween. Our own Universe (as we shall see later) appears to be flat $\Omega = 1; k = 0$ and dominated by vacuum energy: $\Omega_m \sim 0.3$; $\Omega_\Lambda \sim 0.7$. Thus most of the energy density of our Universe is dark: dark matter and dark energy – both of which remain mysterious. Exciting times for a theoretical physicist!

7.4 Making observations in cosmology

To derive observables from the FLRW metric, we must first work out how to determine distances. Let's start with some useful theoretical distances before moving to some more observable ones.

1. *Proper distance.* Suppose we place an observer on a distant receding galaxy that co-moves with the expansion. We will call this observer a *fundamental* or *co-moving* observer. Since she moves with the expansion, the proper separation between her and us is just $D_{\text{prop}} = R(t)r$. This defines the proper distance.

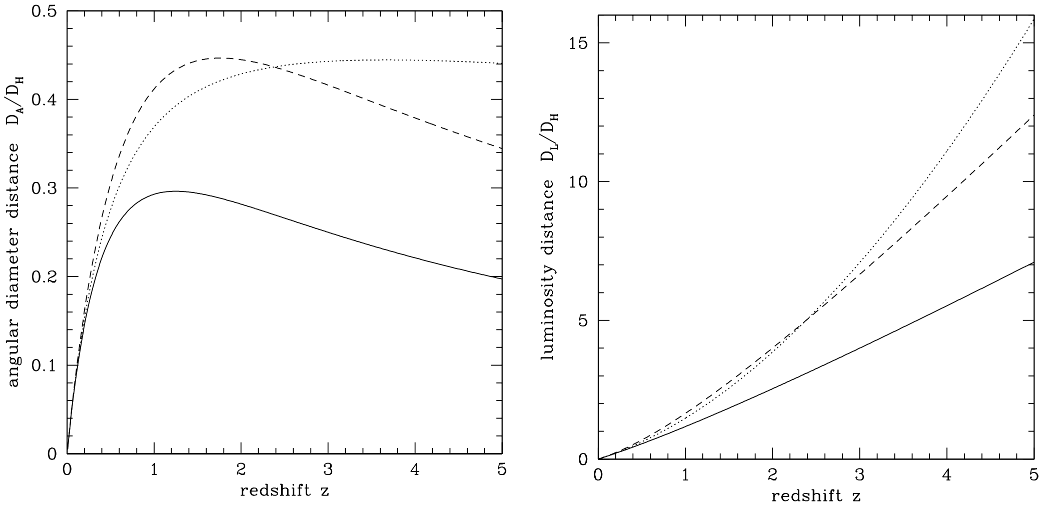


Figure 7.6: Two distance measures for the FLRW metric: the angular diameter distance (left) and the luminosity distance (right). The solid line shows a model with $(\Omega_m, \Omega_\Lambda) = (1, 0)$, the dotted line is for $(\Omega_m, \Omega_\Lambda) = (0.05, 0)$, and the dashed is for: $(\Omega_m, \Omega_\Lambda) = (0.2, 0.8)$. Distances are plotted relative to the 'Hubble' distance: $D_H = c/H_0$.

2. *Co-moving distance.* This is simply the proper distance divided by the scale factor: $D_{\text{comov}} = D_{\text{prop}}/R = r$. The co-moving distance for co-moving observers does not change with time.

The proper distance allows us to connect the FLRW metric expansion to Hubble's law. Writing the velocity of the expansion as:

$$\dot{D}_{\text{prop}} = v = \dot{R}r = \frac{\dot{R}}{R} D_{\text{prop}} \tag{7.26}$$

we derive Hubble's law with $H = \dot{R}/R$ (recall that r is not a function of time).

Now, both of the above distances are well defined but hard to measure. In practice all we really see is the *redshift* z of photons from distant receding galaxies. For nearby galaxies, we can interpret this as a Doppler shift due to some recession velocity:

$$\frac{\nu_e}{\nu_o} = 1 + z \simeq 1 + \frac{v}{c} \tag{7.27}$$

where $\nu_{e,o}$ is the frequency of the emitted and observed photon, respectively. However, this is not the case for more distant galaxies. In general, the redshift is determined by the line integral along a photon null geodesic in the FLRW metric. For pure radial motion, and using $d\tau = 0$, the line integral is:

$$r = \int_{t_e}^{t_o} \frac{cdt}{R(t)} \tag{7.28}$$

where $t_{e,o}$ are the time of emission and observation, respectively. Notice that r is a *co-moving* distance, and is therefore time invariant. Thus, we may write:

$$cdt = Rdr \Rightarrow \tag{7.29}$$

$$\frac{dt_e}{dt_o} = \frac{d\nu_o}{d\nu_e} = \frac{R(t_e)}{R(t_o)} = \frac{1}{1+z} \tag{7.30}$$

This is *not* the Doppler shift formula! It is emphatically not the same as equation 7.27. Only for very nearby galaxies can we reliably interpret this cosmological redshift as a recession velocity. However, we *can* still use the redshift to derive distances. Equation 7.30 nicely relates the scale factor to the redshift:

$$a = \frac{R}{R_0} = (1+z)^{-1} \tag{7.31}$$

since $z = 0$ corresponds to $R_0 = R(t_o)$.

And thus, we may now use equation 7.11 to determine distances from redshifts for a given equation of state for the Universe. Writing:

$$Rdr = cdt = cdR/\dot{R} = cdR/(RH) \tag{7.32}$$

and using $R = R_0/(1+z)$, we have from equation 7.11:

$$R_0 dr = \frac{c}{H(z)}dz = \frac{c}{H_0}\left[(1-\Omega_0)(1+z)^2 + \Omega_\Lambda + \Omega_m(1+z)^3 + \Omega_r(1+z)^4\right]^{-1/2} dz \tag{7.33}$$

which now relates observed redshift z to proper distance measured today $R_0 dr$.

Other useful distances measures are:

1. *The angular diameter distance*: $D_A = (1+z)^{-1}R_0 S_k(r)$. This is particularly useful for gravitational lensing because it relates the proper transverse size of an object r_p to a measured angle on the sky $d\psi$:

   $$d\psi D_A = r_p \tag{7.34}$$

   The angular diameter distance follows from considering pure transverse motion for photons on null geodesics in the FLRW metric:

   $$\begin{aligned} ds^2 &= c^2dt^2 - R^2S_k(r)^2d\psi^2 \\ &= c^2dt^2 - \frac{R_0^2}{(1+z)^2}S_k(r)^2d\psi^2 \\ &= 0 \end{aligned} \tag{7.35}$$

   $\Rightarrow$

   $$cdt = \frac{R_0}{1+z}S_k d\psi \tag{7.36}$$

   which comparing with equation 7.34 gives:

   $$D_A = \frac{R_0 S_k}{1+z} \tag{7.37}$$

2. *The Luminosity distance*: $D_L = (1+z)R_0S_k(r)$. This is how surface brightness falls off with 'distance' in the FLRW metric.

It is clear that we must exercise care when thinking about observations on cosmological scales. Our usual intuition that 'flux falls off with distance squared', or 'size is distance times angle' requires careful thought about the definition of 'distance' in each case. Similar care is required when thinking about velocities on cosmological scales. Remember, we measure the shift of spectral lines, not velocities!

A plot of the angular diameter distance and luminosity distance as function of redshift for various cosmologies is given in Figure 7.6 (taken from Hogg 1999).

7.5 The Big Bang

So far, we have been winding the clock forwards to see what our Universe will do next. But the same Friedmann equations also allow us to wind the clock backwards to see what happened in the past. If we are expanding into the future, then we must be shrinking into the past. This is what leads us to believe that our Universe started in a hot 'Big-Bang': the limit where the scale factor $a \rightarrow 0$. We may now use equation 7.11 to integrate from the ‘beginning’ (the Big Bang; $a = 0$) to the present time ($a = 1$) to give us the age of the Universe (see problem sheet).

The name ‘Big Bang’ has a curious history. It is attributed to Fred Hoyle who originally meant it has a disparaging remark: how could the Universe have such a beginning? He favoured instead a steady state model for the Universe, where the Universe still expands, but matter is continually created everywhere from the vacuum. In Hoyle’s Universe, there is no beginning nor end. The trouble is that no evidence for continually created matter has ever been found. The Big Bang has not only become our premier cosmological model, but it has adopted its name from an honourable competitor.

Lecture 8

Cosmological probes of dark matter II: The inhomogeneous Universe

In this lecture, we build on results from the previous lecture to study the onset of inhomogeneities in the Universe.

8.1 The inhomogeneous Universe

As we discussed in the last lecture, the Universe on large scales is observed to be very close to homogeneous. On smaller scales in the nearby Universe, however ($\lesssim 70\,\mathrm{Mpc}$), the Universe becomes very inhomogeneous. We start to see local fluctuations in the density field due to the presence of *structure*: galaxies like our own Milky Way. This suggests that after the Big Bang, the Universe was not perfectly homogeneous. Tiny fluctuations – perhaps seeded by quantum effects – provided just enough inhomogeneity to grow into the galaxies and local structure we see today. In this lecture, we study the growth of such fluctuations using linear perturbation theory.

8.2 Two types of perturbation

Before discussing how perturbations will evolve and grow, we should discuss briefly how they started out. As we mentioned previously, most of the Universe is dominated by radiation and matter at early times so we focus on perturbations to these two coupled fluids. Let us define the density perturbation relative to the homogeneous expanding background ρ_0 as:

$$\delta = \frac{\rho - \rho_0}{\rho_0} = \frac{\delta\rho}{\rho_0} \tag{8.1}$$

This can be a matter perturbation δ_m, or a radiation one δ_r. These behave differently since the matter perturbation is applied to a non-relativistic fluid, while the radiation one applies to a relativistic fluid. There are two main types of perturbation: adiabatic, and isocurvature.

- **Adiabatic perturbations:** If we squeeze or stretch our matter/radiation fluid very slowly (slow as compared to any other timescale of interest), then we will induce adiabatic perturbations. Recall that such slow adiabatic perturbations conserve the *entropy density* of the fluid. In this case $\delta_r = \frac{4}{3}\delta_m$ and the matter and radiation fields deform *together*.
- **Isocurvature perturbations:** This is the *orthogonal* perturbation to the above. Here we perturb the entropy density, but not the energy density. Now we have that $\rho_r\delta_r = -\rho_m\delta_m$ and the matter and radiation fields deform in opposite directions to produce constant spatial curvature (hence the name isocurvature). At early times $\rho_r \gg \rho_m$ and thus isocurvature perturbations imply that $\delta_m \gg \delta_r$ and all of the initial perturbation appears only in the matter field and not the radiation field.

Whether perturbations are adiabatic, isocurvature or some linear combination of the above makes a difference as we shall see later on (Figure 8.1).

8.3 Linear perturbation theory

We have presented the two main types of initial perturbation. We now study how these will grow. A full treatment of the growth of structure in an expanding FLRW spacetime involves linear (and potentially higher order) perturbations of the GR field equations, which is a bit involved for us here. Instead, we will use a useful approximation. Since any inhomogeneity must be *local*, and local space must be Minkowski, we can approximate the local dynamics as being pure Newtonian weak field embedded in an expanding FLRW spacetime. Thus, locally, the Universe must approach a classical fluid. This can be a rather complex fluid (e.g. a mix of non-relativistic matter, relativistic radiation, and interaction terms between the two; see later), but for now let us imagine that it is a viscous-free fluid that obeys the Euler equations:

$$\frac{D\rho}{Dt} = -\rho\nabla\cdot\mathbf{v} \qquad ; \qquad \text{Continuity} \tag{8.2}$$

$$\frac{D\mathbf{v}}{Dt} = -\frac{\nabla p}{\rho} - \nabla\Phi \qquad ; \qquad \text{Momentum} \tag{8.3}$$

$$\nabla^2\Phi = 4\pi G\rho \qquad ; \qquad \text{Poisson} \tag{8.4}$$

where ρ is the density of the fluid; p is the pressure; Φ is the gravitational potential; $\mathbf{v}$ is the velocity; and $D/Dt = \partial/\partial t + \mathbf{v}\cdot\nabla$ is the convective derivative (i.e. the local time derivative taken by someone moving with the flow). The above equations are closed by the equation of state that links pressure to density. Since this is different for matter/radiation/etc. we will not specify it just yet.

Now, consider a small perturbation:

$$\rho = \rho_0 + \delta\rho \qquad ; \qquad \mathbf{v} = \mathbf{v}_0 + \delta\mathbf{v} \qquad ; \qquad \text{and similar} \tag{8.5}$$

where the unperturbed solution ρ_0 etc. refers to the *expanding FLRW spacetime*. In other words, a observer moving at $\mathbf{v}_0$ would be moving with the Hubble flow: $\mathbf{v}_0 = H\mathbf{x}_0$. Plugging this perturbation into the continuity equation, we obtain:

$$\left[\frac{\partial}{\partial t} + (\mathbf{v}_0 + \delta\mathbf{v})\cdot\nabla\right](\rho_0 + \delta\rho) = -(\rho_0 + \delta\rho)\nabla\cdot(\mathbf{v}_0 + \delta\mathbf{v}) \tag{8.6}$$

Subtracting the zeroth order equation $\left[\frac{\partial}{\partial t} + \mathbf{v}_0\cdot\nabla\right]\rho_0 = -\rho_0\nabla\cdot\mathbf{v}_0$, using the fact that ρ_0 is homogeneous ($\nabla\rho_0 = 0$) and retaining terms at linear order, we obtain:

$$\frac{d\delta\rho}{dt} \simeq -\rho_0\nabla\cdot(\delta\mathbf{v}) - \delta\rho\nabla\cdot\mathbf{v}_0 + O(\delta^2) \tag{8.7}$$

where $d/dt = \left[\frac{\partial}{\partial t} + \mathbf{v}_0\cdot\nabla\right]$ is now a time derivative with respect to an observer co-moving with the *unperturbed* expansion. Similarly, the momentum and Poisson equations become:

$$\frac{d\delta\mathbf{v}}{dt} \simeq -\frac{\nabla\delta p}{\rho_0} - \nabla\delta\Phi - (\delta\mathbf{v}\cdot\nabla)\mathbf{v}_0 + O(\delta^2) \tag{8.8}$$

$$\nabla^2\delta\Phi \simeq 4\pi G\delta\rho + O(\delta^2) \tag{8.9}$$

Now, recall that $\mathbf{v}_0 = H\mathbf{x}_0$. Thus, we may simplify the tricky term $(\delta\mathbf{v}\cdot\nabla)\mathbf{v}_0$. Writing it out explicitly, we have:

$$H\left[\delta v_x\frac{\partial}{\partial x} + \delta v_y\frac{\partial}{\partial y} + \delta v_z\frac{\partial}{\partial z}\right]\mathbf{x}_0 = H\delta\mathbf{v} \tag{8.10}$$

Finally, recalling our definition of a fractional density perturbation (equation 8.1):

$$\delta = \frac{\delta\rho}{\rho_0} \tag{8.11}$$

we may transform the continuity equation:

$$\begin{aligned}\frac{d}{dt}\left(\frac{\delta\rho}{\rho_0}\right) &= \frac{1}{\rho_0}\frac{d}{dt}(\delta\rho) - \frac{\delta\rho}{\rho_0^2}\frac{d}{dt}\rho_0 \\ &= -\nabla\cdot\delta\mathbf{v} - \frac{\delta\rho}{\rho_0}\nabla\cdot\mathbf{v}_0 + \frac{\delta\rho}{\rho_0}\nabla\cdot\mathbf{v}_0 \qquad (8.12)\end{aligned}$$

⇒

$$\frac{d\delta}{dt} = -\nabla\cdot\delta\mathbf{v} \tag{8.13}$$

and we see that the above *linearised* system of equations do not depend on the expansion velocity $\mathbf{v}_0$.

The above equations take on an even simpler form if we transform to co-moving coordinates:

$$\mathbf{x}(t) = a(t)\mathbf{r}(t) \tag{8.14}$$

where $a(t)$ is derived from the Friedmann equation (equation 7.11; §7), and the co-moving coordinate $\mathbf{r}$ is now a *function of time*, reflecting the departure from homogeneity caused by structure formation.

Taking the time derivative (using 'dot' $\equiv d/dt$), we have:

$$\begin{aligned}\dot{\mathbf{x}} &= \dot{a}\mathbf{r} + a\dot{\mathbf{r}} \\ &= \mathbf{v}_0 + \delta\mathbf{v} \qquad (8.15)\end{aligned}$$

and the term on the left is the familiar Hubble expansion; the term on the right is called the *peculiar velocity*, which is the local velocity with respect to the Hubble flow.

Finally, using $\nabla_x = \frac{1}{a}\nabla_r$, our linearised equations of motion become:

$$\frac{d}{dt}(a\dot{\mathbf{r}}) = -\frac{1}{a}\frac{\nabla_r\delta p}{\rho_0} - \frac{1}{a}\nabla_r\delta\Phi - \frac{\dot{a}}{a}a\dot{\mathbf{r}} \tag{8.16}$$

⇒

$$\ddot{\mathbf{r}} + 2\frac{\dot{a}}{a}\dot{\mathbf{r}} = -\frac{1}{a^2}\nabla_r\delta\Phi - \frac{1}{a^2\rho_0}\nabla_r\delta p \tag{8.17}$$

with:

$$\dot{\delta} = -\nabla_r\cdot\dot{\mathbf{r}} \tag{8.18}$$

and:

$$\frac{\nabla_r^2\delta\Phi}{a^2} = 4\pi G\rho_0\delta \tag{8.19}$$

where we recall that δ describes the over/under-density of a region of the Universe $\delta = (\rho - \rho_0)/\rho_0$.

It is worth taking a moment to think about what the above equations mean. We have assumed that the Universe can be described locally by a classical fluid. We have then perturbed this fluid, assuming that the unperturbed solution simply follows the FLRW expanding homogeneous Universe. Finally, we transformed to co-moving coordinates that allowed us to factor out the expansion. This last point is key. We can now use the evolution equation we derived in the last lecture for $a(t)$: the Friedmann equation. Equations 8.17, 8.18 and 8.19 then describe the local inhomogeneous motion *relative to this expanding Universe.*

In the absence of a perturbation, $\delta p = \delta\Phi = \delta = 0$ and thus $\dot{\mathbf{r}} = \ddot{\mathbf{r}} = 0$ and we recover the unperturbed Friedmann solution with $\mathbf{r} = \text{const.}$.

8.3.1 Non-relativistic fluids

Up to now, we have not specified an equation of state for our fluid. Specifying one allows us to reduce to just one evolution equation for the density perturbation δ. First, let us multiply equation 8.17 by the ∇_r operator:

$$\nabla_r \cdot \ddot{\mathbf{r}} + 2\frac{\dot{a}}{a}\nabla_r \cdot \dot{\mathbf{r}} = -\frac{1}{a^2}\nabla_r^2 \delta\Phi - \frac{1}{a^2\rho_0}\nabla_r^2 \delta p \tag{8.20}$$

Now we can substitute for $\nabla_r \cdot \dot{\mathbf{r}}$ and $\nabla_r \cdot \ddot{\mathbf{r}}$ using equation 8.18:

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = \delta 4\pi G\rho_0 + \frac{1}{a^2\rho_0}\nabla_r^2 \delta p \tag{8.21}$$

and finally, we must specify something about the equation of state to substitute for δp. We can write rather generally: $c_s^2 = \frac{\partial p}{\partial \rho}$, where c_s is the sound speed for the fluid. Thus:

$$\nabla \delta p = c_s^2 \nabla \delta \rho \tag{8.22}$$

and, recalling that $\delta = \delta\rho/\rho_0$, our equation for δ becomes:

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = \delta 4\pi G\rho_0 + \frac{c_s^2}{a^2}\nabla_r^2 \delta \tag{8.23}$$

where we have used the fact that the unperturbed Universe is homogeneous ($\nabla\rho_0 = 0$).

Before we attempt a full solution, let us consider some simple cases. Perhaps the simplest is a plane wave perturbation of the form:

$$\delta(\mathbf{r}, t) = A(t)\exp(-i\mathbf{k}_r \cdot \mathbf{r}) \tag{8.24}$$

where $\mathbf{k}_r = a\mathbf{k}$ is a co-moving wavevector. Physically, this represents a single Fourier mode whose wavelength stretches with the Universe. Thus, the right hand side of equation 8.23 further simplifies to:

$$\frac{c_s^2}{a^2}\nabla_r^2 \delta = -c_s^2 k^2 \delta \tag{8.25}$$

where k is the amplitude of the proper wavevector $\mathbf{k}$. We now gain an important insight. Equation 8.23 for plane waves represents a competition between the expanding Universe and local gravitational collapse. On small enough scales, $\dot{a}$ becomes dynamically unimportant, and we have: $\ddot{\delta} = \delta(4\pi G\rho_0 - c_s^2 k^2)$, which is straightforwardly solved for the time dependence of $\delta(\mathbf{r}, t)$:

$$A(t) = \exp(\pm t/\tau) \tag{8.26}$$

with $\tau = \left(4\pi G\rho_0 - c_s^2 k^2\right)^{-1/2}$.

Thus, there is a critical wavelength[1]:

$$\lambda_J = c_s\sqrt{\frac{\pi}{\rho_0 G}} \tag{8.27}$$

where we move from oscillating stationary solutions, to either exponentially growing or shrinking solutions. This is called the *Jeans length*.

The above is, of course, made more complicated by the $\dot{a}$ factor that appears in the full equation. But once the Hubble expansion becomes negligible then runaway growth of structure is inevitable.

A full solution to equation 8.23 is usually expressed in terms of the linear *growth factor* that separates out the spatial and temporal evolution of the density perturbation $\delta(\mathbf{x}, t)$:

$$\delta(\mathbf{r}, t) = A(\mathbf{r})D_1(t) + B(\mathbf{r})D_2(t) \tag{8.28}$$

where D_1 is the growing and D_2 the decaying mode. The growth factor can also be written as a function of redshift instead of time using the redshift dependence of the scale factor $a(z)$ (equation 7.31).

[1] Recall that the wavenumber is defined such that the wavelength $\lambda = \frac{2\pi}{k}$.

We will not present solutions for the growth factor for different cosmologies here, though in many cases these are analytic or semi-analytic (requiring only the solution of an integral). For further details see Peebles 1980.

8.3.2 Relativistic fluids

So far, we a have considered only non-relativistic fluids. Relativistic fluids – like radiation – have similar but slightly different fluid equations:

$$\frac{D}{Dt}(\rho + p/c^2) = \frac{\partial}{\partial t}(p/c^2) - (\rho + p/c^2)\nabla \cdot \mathbf{v} \tag{8.29}$$

$$\frac{D\mathbf{v}}{Dt} = -\nabla\Phi \tag{8.30}$$

$$\nabla^2\Phi = 4\pi G(\rho + 3p/c^2) \tag{8.31}$$

which are the continuity, momentum and Poisson equations for a pure special relativistic fluid assuming Newtonian gravity. These are similar to previously, but now we must remember to flux momentum and mass in the continuity equation, and to include the momentum terms as a source of gravity in the Poisson equation. We have also assumed here that the pressure gradients are negligible as compared to gravity.

The above is not entirely satisfactory. We are assuming a special relativistic fluid and Newtonian gravity. In the very early Universe this must fail and we ought to do a proper linear perturbation of the full GR equations. But we can gain some further insight from the above 'toy' equations. For pure radiation, the equation of state is given by $p = \rho c^2/3$. Performing an analysis similarly to the above for the non-relativistic fluid, we derive [exercise]:

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = \frac{32\pi}{3}G\rho_0\delta + \frac{c_s^2}{a^2}\nabla_r^2\delta \tag{8.32}$$

and thus in a pure radiation dominated Universe, the equations are similar to the non-relativistic case, but the driving term is ∼ 8 times higher.

8.3.3 Beyond simple fluids

In reality, the very early Universe is a highly complex 'fluid' – in fact, not really a fluid at all. It has a collisionless component (dark matter); a non-relativistic fluid component (baryons); a relativistic fluid component (photons); and vacuum. Each of these fluids interact via gravity and/or other forces and an analytic treatment is not promising. Instead, we must resort to a numerical approach.

A full treatment requires first properly perturbing the Einstein field equations with the FLRW metric as input. Already this is problematic. While the field equations are coordinate independent, once we perturb about the solution we often are forced to pick a particular coordinate system or 'gauge'. Linear perturbations typically break coordinate invariance and so we must be careful that the results do not depend on the initial coordinate choice. Best of all is to perform a perturbation analysis that is gauge invariant (e.g. Bardeen 1980). We will not go into the details here.

With a proper perturbation analysis, we must then solve the full system of coupled 'fluid' equations for the early Universe. This means solving the general relativistic form of the Boltzmann equation with the correct interaction terms included. It is a highly involved problem (a full course in its own right). We will give some of the results from such codes when we discuss the CMB, but we will not give any great detail. The interested reader is referred to Seljak and Zaldarriaga 1996 and references therein. It is, of course, remarkable that the results of such a calculation can be trusted. However, three independent codes have been recently shown to converge on the same results at 0.1% precision for most parameters of interest (Seljak *et al.* 2003).

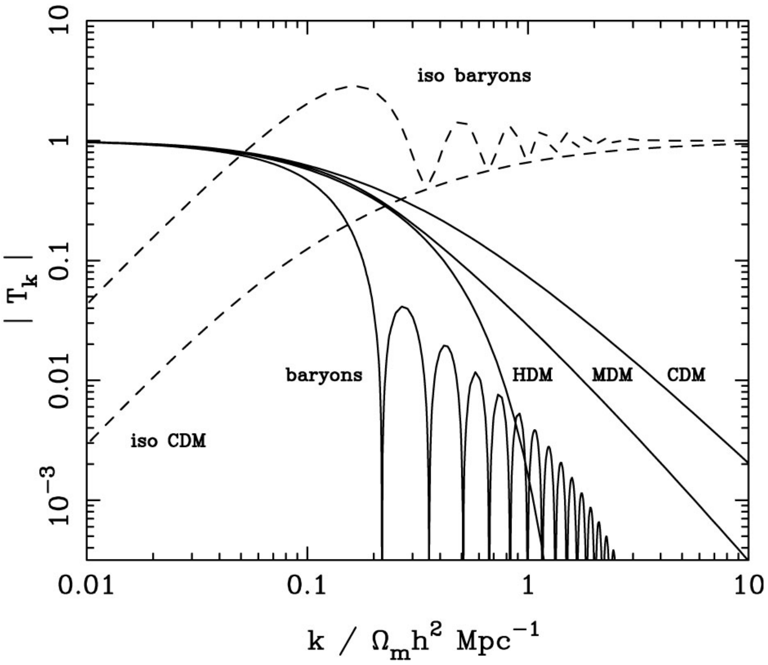


Figure 8.1: The numerically calculated transfer function for Universes with different composition and different initial conditions. All cases assume a flat Universe $\Omega_0 = 1$ with a Hubble constant $H_0 = 50$ km s^{-1} /Mpc and adiabatic perturbations initially, unless otherwise stated. The models shown are for baryons; Hot Dark Matter (HDM); Cold Dark Matter (CDM); Mixed Dark Matter (MDM); and two isocurvature models.

8.3.4 The transfer function

The transfer function encapsulates the results of numerical integrations of the growth of density perturbations, δ for complex non-fluids in the early Universe:

$$T_k \equiv \frac{\delta_k(z=0)}{\delta_k(z)D_1(z)} \tag{8.33}$$

where $D_1(z)$ is the linear growth factor between redshift z and the present (equation 8.28). The transfer function necessarily depends on the composition of these fluids; some examples are given in Figure 8.1 (taken from Peacock 1999).

Notice that both the initial conditions and the composition of the Universe greatly affect the transfer function. Thus, the distribution of density peaks in the Universe today and its time evolution encodes information about the initial conditions and the composition. Notice also the striking oscillations in the baryonic transfer function that are not present in any of the dark matter models. There is also a strong damping tail to high wavenumber k. The former is a result of acoustic waves travelling through the baryonic fluid that is collisional and therefore has a pressure term. These are not present for the dark matter fluids that are assumed to be pressureless (though, of course, one could test this using such models!). The latter is a result of a process called 'Silk damping' (Silk 1968). This occurs because baryonic matter interacts with the photon fluid in the early Universe. The photon mean free path acts as a damping scale that suppresses the growth of structure. Again, this is not the case for the dark matter models. In fact, this is the basis for the *need* for dark matter to explain cosmological observations. Without any dark matter, structure would be suppressed on scales smaller than ~ 1 Mpc.

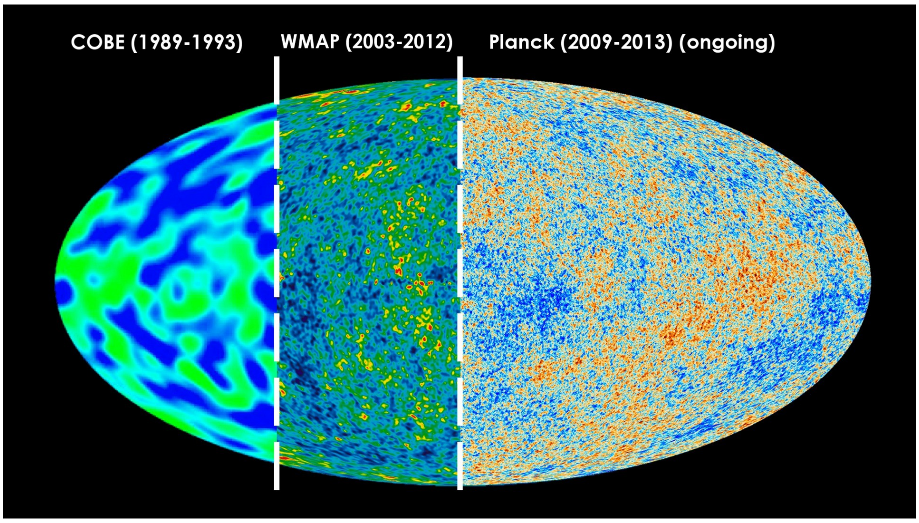


Figure 8.2: The cosmic microwave background radiation as seen by the COBE, WMAP and Planck satellites.

8.4 The cosmic microwave background radiation

The cosmic microwave background radiation (CMB) is a beautiful *prediction* of our cosmological model. It was predicted to exist by Dicke *et al.* 1965 in the very same year that Penzias & Wilson – two scientists at Bell labs who were working on something completely different – stumbled across the CMB radiation (Penzias and Wilson 1965). As we have discussed previously, the early Universe is dominated by the radiation field (c.f. equation 7.11; §7). At very early times, this is in thermal equilibrium with the matter in the Universe and forms a plasma which is largely opaque to light. A critical moment occurs when this plasma cools to the point where photons can efficiently escape: *recombination*. It is this moment which gave birth to the cosmic microwave background radiation, or CMB – the afterglow of the Big Bang – that we see everywhere around us in the sky today. A modern picture of the observed CMB sky is shown in Figure 8.2.

The truly fascinating thing about the CMB is that it is not perfectly smooth. Looking at the COBE picture of the CMB, we see tiny temperature fluctuations at the level of 10^{-5} (Smoot *et al.* 1992). These encode information about the inhomogeneities present in the Universe at very early times and, therefore, about the *composition* of the Universe. Furthermore, since the fluctuations in temperature are so small, we are still in the *linear* regime and can use the machinery of linear perturbation theory derived above to understand the CMB (Hu and Dodelson 2002).

8.4.1 The CMB power spectrum

The CMB is very close to being a black body spectrum, with tiny fluctuations around this. For this reason, the fluctuations are typically characterised by a function $\Theta(\theta, \phi) = \Delta T/T$, where (θ, ϕ) is the direction on the sky. We may write this function as some spherical harmonic decomposition:

$$\Theta(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{m=l} \Theta_{lm} Y_{lm}(\theta, \phi) \tag{8.34}$$

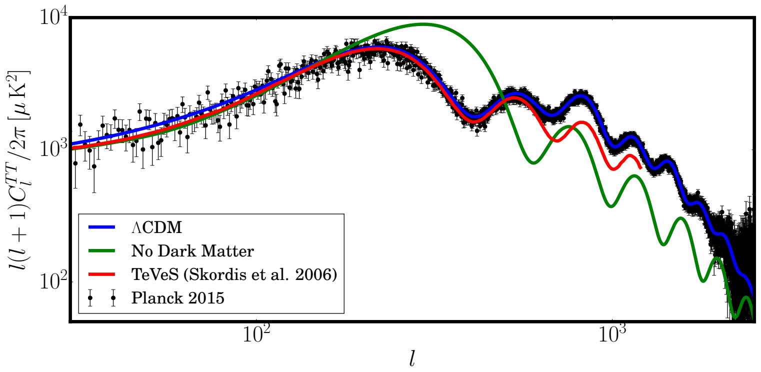


Figure 8.3: The CMB angular power spectrum measured by the Planck satellite. The blue line shows our current best fitting standard cosmological model; the green shows what happens if all the matter in this model is 'baryonic', all other things being equal (i.e. without any dark matter); the red line shows the best-fitting TeVeS alternative gravity model from Skordis et al. (2006). See text for further details.

where $Y_{lm}(\theta, \phi)$ are the spherical harmonics (c.f. Appendix F), and the coefficients Θ_{lm} are given in the usual way by:

$$\Theta_{lm} = \int_0^{\pi} \sin\theta d\theta \int_0^{2\pi} d\phi Y_{lm}^*(\theta,\phi)\Theta(\theta,\phi) \tag{8.35}$$

If the fluctuations are pure Gaussian, then they have no preferred direction. Thus, we can 'integrate out' the dependence on m, and express all of the information purely by the power spectrum (the intensity of each spectral component):

$$\begin{aligned} C_l &= \frac{\sum_{m=-l}^{m=l}\sum_{m'=-l'}^{m'=l'}\Theta_{lm}\Theta_{l'm'}^*}{2l+1} \\ &= \frac{\sum_m |\Theta_{ml}|^2}{2l+1} \end{aligned} \tag{8.36}$$

where $2l+1$ is the number of terms in the sum and so is a normalisation.

Typically, what is plotted for the CMB angular power spectrum is:

$$\Delta_T^2 \equiv \frac{l(l+1)}{2\pi} C_l T^2 \tag{8.37}$$

where T is the temperature. Because typically $l \gg 1$, Δ_T^2 approximately encodes the power per natural logarithmic interval in l.

A recent plot of the Δ_T^2 for the CMB determined from the Planck experiment is given in Figure 8.3 (data taken from Planck Collaboration *et al.* 2013). Our current best-fitting cosmological model is marked in blue. On large scales, $l \lesssim 10$, the errors become dominated by 'cosmic variance'. This simply means that there can only be $2l+1$ measurements for each C_l (c.f. equation 8.36). Thus, low l moments will be fundamentally more poorly sampled. Also shown are a model with no dark matter (green) and a model with no dark matter and the TeVeS alternative gravity model (§6; red). I discuss these in §8.4.3.

8.4.2 The standard cosmological model: ΛCDM

As mentioned previously (§8.3.3), using linear perturbation of the FLRW metric we can numerically solve the coupled Boltzmann fluid equations to predict the distribution of density fluctuations in the early Universe. Calculating this at the time of last scattering (recombination), allows us to predict the CMB angular power spectrum for a given choice of cosmological parameters. Figure 8.4 shows the results of such a calculation for different choices of key cosmological parameters: $\Omega_m, \Omega_\Lambda, \Omega_b$ and Ω_0 (see §7 for definitions; Figure taken from Hu and Dodelson 2002). Notice that there are degeneracies

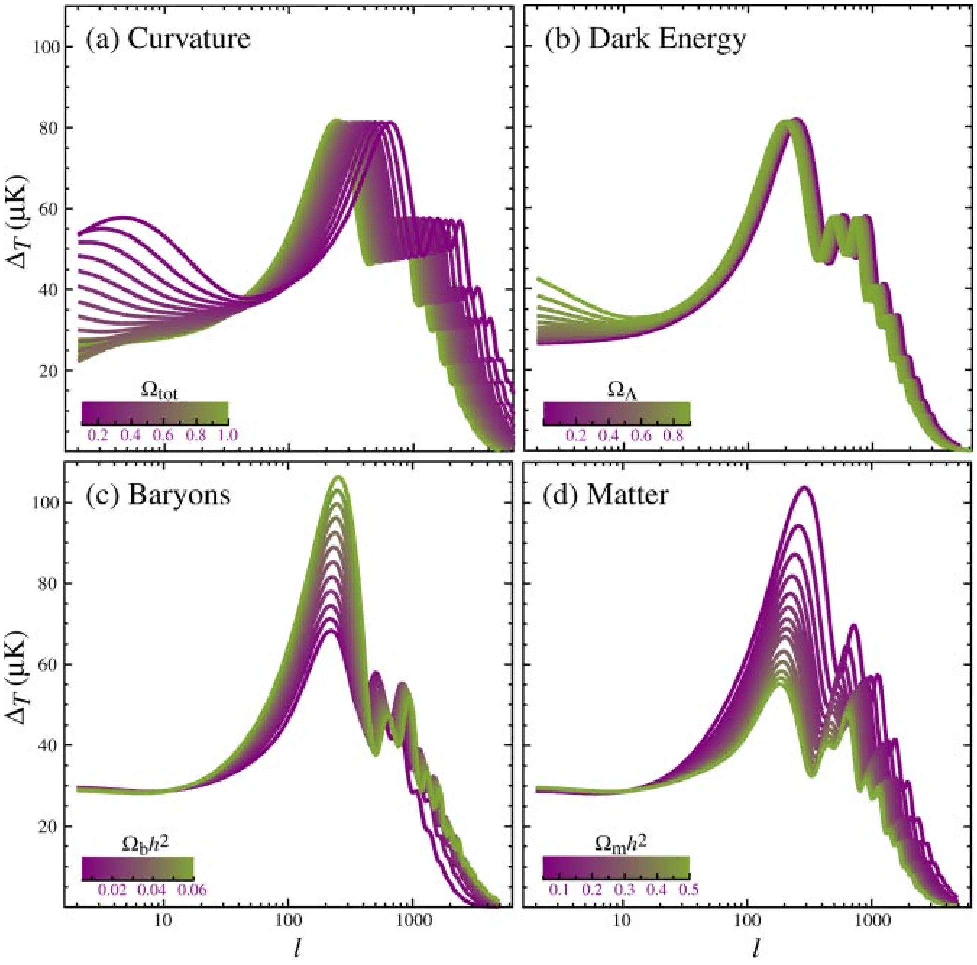


Figure 8.4: The sensitivity of the CMB to changes in the cosmological parameters. Here $\Omega_{\rm tot}$ is what we have previously called Ω_0. All parameters are varied around a fiducial model with: $\Omega_0 = 1$, $\Omega_\Lambda = 0.65$, $\Omega_b h^2 = 0.02$ and $\Omega_m h^2 = 0.147$.

between the effects of each parameter. For this reason, we cannot expect the CMB alone to fully derive the cosmological model – we will have to combine it with other probes. We briefly explain the trends in the data as we vary the cosmological parameters, next.

- The curvature ($k \propto \Omega_0 - 1$) shifts the position of the first and subsequent peaks. Smaller Ω_0 (more negative curvature) pushes the peaks to larger l. This effect is a result of the difference between coordinate and angular diameter distance. In a Universe with positive curvature, two points separated by a given angle on the sky are really further apart than we would expect in Euclidean space; the converse is true for negative curvature. Thus, all other things being equal, a spatially open Universe will push the power to smaller spatial scales (larger l). This is what is seen in Figure 8.4. The observed position of the first peak in the CMB demands a near-perfectly flat Universe with $\Omega_0 = 1$.

- Dark energy (Ω_Λ) also shifts the position of the peaks, though the effect is much smaller than the curvature. Larger Ω_Λ (for a flat Universe) shifts the peak to smaller l (larger scales). This can be understood because dark energy acts to delay structure formation. Since the smallest scales form earliest (they have the shortest dynamical times), this necessarily shifts power to larger scales at recombination. The effect is small, however, since dark energy is not a dominant contributor to the Friedmann equations at these early times (c.f. §7), and is not directly contributing to the growth of structure either.

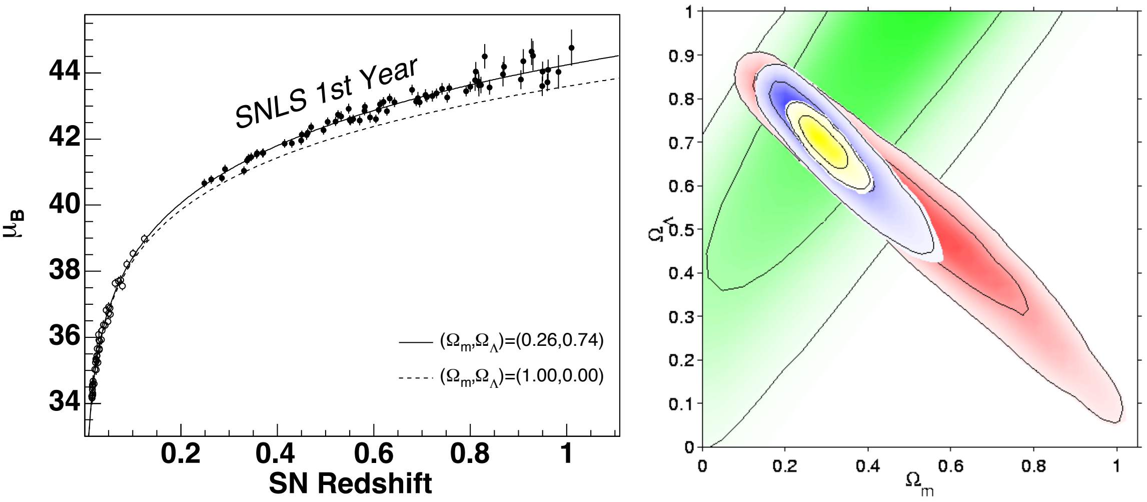


Figure 8.5: **Left:** The SuperNovae Legacy Survey data (SNLS; Astier et al. 2006). The SNLS data measure the luminosity distance (μ_B) as a function of redshift z and, therefore, the scale factor $a(z)$ (see §7). Two cosmological models are over-plotted. **Right:** In combination with the CMB (red contours) and the CMB with a measurement of the Hubble constant h (blue contours), the supernovae data (green contours) break the orthogonal degeneracy between Ω_m and Ω_Λ in each data set (yellow contours; and see also Figure 8.6).

- The total matter content (Ω_m) mainly raises the amplitude of the peaks. This can be understood since more non-relativistic matter means more growth of structure on all scales. There is a slight shift, however, also to smaller scales (larger l) because smaller structures grow faster due to the smaller dynamical times.

- The baryon content (Ω_b) mainly affects the amplitude of the *first peak* and not the subsequent peaks. This is how the CMB data can differentiate between baryonic matter (that interacts with the relativistic radiation fluid) and collisionless non-relativistic matter like dark matter (that does not). This photon-baryon interactions damps the growth of structure on small scales at these early times suppressing peaks higher than the first one. The effect does not occur for dark matter since it does not couple to the radiation fluid.

The position and amplitude of the first peak in the CMB data is enough on its own to tell us that the Universe is almost perfectly flat and that it *must* contain some non-relativistic matter that does not interact with photons: dark matter[2] (see Figures 8.3 and 8.6). In combination with other independent cosmological probes, we can actually measure the cosmological parameters to an impressive accuracy (see Figure 8.6 taken from Spergel *et al.* 2007; and see e.g. Lewis and Bridle 2002; Seljak *et al.* 2006b; Sullivan *et al.* 2011). Two key additional constraints are considered here: data from Type Ia supernovae standard candles (Figure 8.5), and data from low redshift galaxy surveys (which we will discuss in more detail in §9). The Type Ia supernovae data are particularly interesting. Type Ia supernovae have a light curve decay that depends on their luminosity in a characteristic way. Calibrating the luminosity-decay rate relation using low redshift supernovae, Type Ia supernovae then act as excellent standard candles out to high redshift. They tell us the luminosity distance as a function of redshift z, and therefore are a direct probe of the scale factor $a(z)$ (c.f. §7). Some recent data from the SuperNovae Legacy Survey (SNLS) is given in Figure 8.5.

Our current best fitting cosmological model suggests that – remarkably – most of the Universe is dark: $\Omega_0 = 1$, $\Omega_\Lambda = 0.694{\pm}0.007$, $\Omega_b h^2 = 0.02227{\pm}0.0002$, $\Omega_m h^2 = 0.306{\pm}0.007$ and $h = 0.679{\pm}0.006$ with very small errors (Planck Collaboration *et al.* 2013). The success of the standard cosmological

[2] Recall we are working here under the assumption that Einstein's field equations and GR describe gravity, and that we do not live in a special place in the Universe (the Copernican principle). See §6 and §7 for a discussion of the validity of these key assumptions.

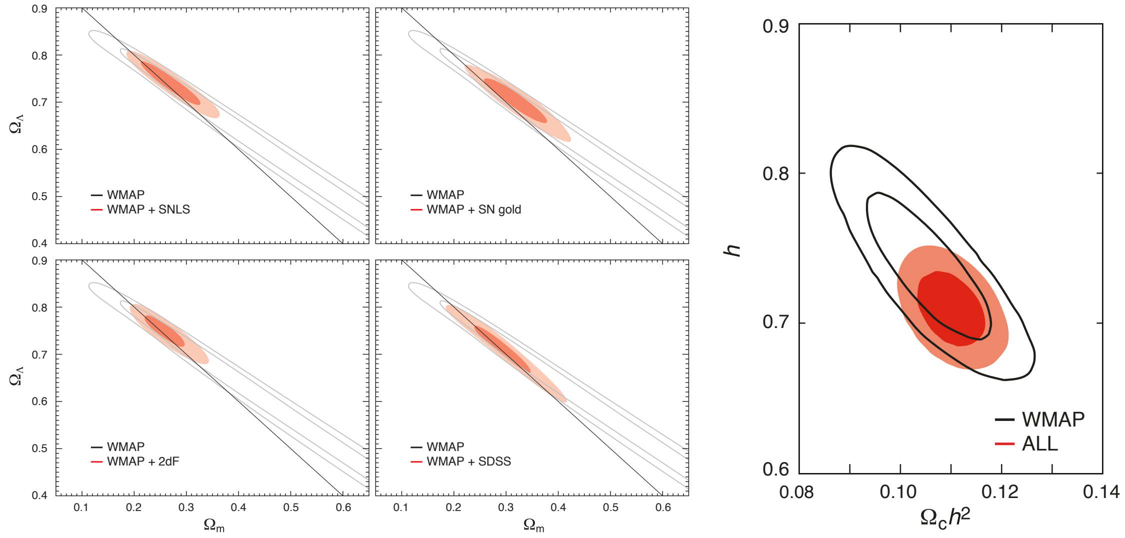


Figure 8.6: Combined constraints on our current cosmological model. **Left:** CMB constraints on dark energy (Ω_Λ), and the total matter content (Ω_m). Notice that the CMB on its own (WMAP; black contours) tells us that the Universe is close to flat $\Omega_m + \Omega_\Lambda \sim 1$. In combination with just one other probe, the degeneracy is broken and we favour the standard cosmological model values $\Omega_\Lambda \sim 0.7; \Omega_m \sim 0.3$. The other probes considered here are the SuperNovae Legacy Survey (Astier et al. 2006; Figure 8.5), the SN 'gold' survey (Riess et al. 2004), and constraints on the non-linear growth of structure from two low-redshift galaxy surveys: the Sloan Digital Sky Survey (Tegmark et al. 2004), and the 2dF survey (Cole et al. 2005). We will discuss the non-linear growth of structure in §9. **Right:** Combined constraints (using all of the previous data and constraints on h) on the dark matter content of the Universe $\Omega_c = \Omega_m - \Omega_b$. Notice that the CMB on its own (WMAP; black contours) tells us already that the Universe must contain dark matter. In combination with other probes, the amount of dark matter needed is well constrained.

model and the era of precision cosmology has led to several nobel prizes over the past few years and a recognition that what was once barely deemed science is now driving our understanding of the structure of the Universe both on incredibly large scales and on sub-atomic scales. The fraction of dark matter required to fit the CMB and other probes is $(\Omega_m - \Omega_b)/\Omega_m \sim 0.83$ (Figure 8.6). This is remarkably similar to the amount of dark matter required to explain the dynamics and lensing data in galaxies and galaxy clusters (c.f. Figure 5.6). Yet it is derived from a very different and independent analysis.

Our standard cosmological model is typically referred to as "ΛCDM", meaning that it requires a cosmological constant Λ, and a 'cold dark matter' component (a non-relativistic fluid that does not noticeably interact with the photon fluid in the early Universe).

8.4.3 The real problem with alternative gravity models

Finally, we can return to the issue of alternative gravity models as an explanation for dark matter. As mentioned at the end of lecture §6, a key evidence for non-baryonic dark matter is the cosmic microwave background radiation. We are now in a position to revisit this. The green line in Figure 8.3 shows what the ΛCDM model looks like if we put all the matter in baryons rather than dark matter, keeping all other things equal. As can be seen, this is a very poor fit to the Planck data (black data points). More importantly, the high l modes are damped as compared to the primary peak due to photon-baryon interactions, while the oscillations due to acoustic waves in the baryonic fluid are too large as compared to the data. Both of these problems go away with the addition of a dark matter fluid that does not interact with the photon fluid. The damping and oscillations are then both suppressed, leading to the ΛCDM model. This is, then, a fundamental problem with all models that seek to explain dark matter using alternative gravity (Skordis *et al.* 2006; Dodelson 2011). Even if we boost the gravitational force to speed up structure formation, if the Universe comprises only baryons, then it will be over-damped on small scales with overly strong oscillations due to acoustic waves. Indeed, this is exactly what can be seen in the red line on Figure 8.3 that shows the best-fitting TeVeS alternative gravity model from Skordis *et al.* 2006. As can be seen, this model fits the lower l peaks, but it ultimately fails at high l. (Actually, even a fit this good requires some dubious gymnastics; this model has three massive neutrinos, each with $m_\nu \sim 2\,\mathrm{eV}$.)

Lecture 9

Cosmological probes of dark matter III: The non-linear growth of structure

In this lecture, we build on results from the previous lecture to study the non-linear growth of structure in the Universe through to the present day.

9.1 The non-linear growth of structure: evolution equations

After the de-coupling of photons (that became the CMB), things became very much simpler. From this point on, radiation is negligible and only matter, curvature and vacuum energy are important. This is a great simplification because we need deal only with two types of non-relativistic fluid: one for baryons, and one for dark matter. Furthermore, these couple only through gravity (through the Poisson equation) that is linear and therefore straightforward to deal with.

First, let us derive the full non-linear equations of motion. We will still assume, as previously, that we can describe the motion as locally Newtonian within an expanding FLRW background (we shall critique this approach in §10.3). Thus, we are still in locally Newtonian weak-field linearised GR. But we will no longer linearise the fluid equations in this limit. (In this sense, what we have been considering so far is a sort of *doubly* linear approximation: linear GR *and* linearised Newtonian fluids. It is the latter assumption that we relax here.)

The derivation is easiest done in co-moving rather than Eulerian coordinates. Writing the Eulerian position $\mathbf{x}$ as $\mathbf{x} = a(t)\mathbf{r}(t)$ as previously, we may differentiate twice to give:

$$\dot{\mathbf{x}} = \dot{a}\mathbf{r} + a\dot{\mathbf{r}} \tag{9.1}$$

$$\ddot{\mathbf{x}} = \ddot{a}\mathbf{r} + 2\dot{a}\dot{\mathbf{r}} + a\ddot{\mathbf{r}} \tag{9.2}$$

where the 'dot' refers to d/dt – the co-moving temporal derivative, as previously. Rearranging in terms of the co-moving acceleration, we have:

$$\ddot{\mathbf{r}} = -2\frac{\dot{a}}{a}\dot{\mathbf{r}} - \frac{\nabla_r \Delta\Phi}{a^2} - \mathbf{g}_0 \tag{9.3}$$

where we have written the acceleration terms as the *peculiar acceleration*: $\ddot{\mathbf{x}}/a = \frac{\nabla_r \Delta\Phi}{a^2}$, and an unperturbed acceleration: $\mathbf{g}_0 = \frac{\ddot{a}}{a}\mathbf{r}$, and ignored pressure forces (for now). Assuming weak field general relativity, the Poisson equation still holds locally and we have:

$$\frac{\nabla_r^2(\Phi_0 + \Delta\Phi)}{a^2} = 4\pi G(\rho_0 + \Delta\rho) \tag{9.4}$$

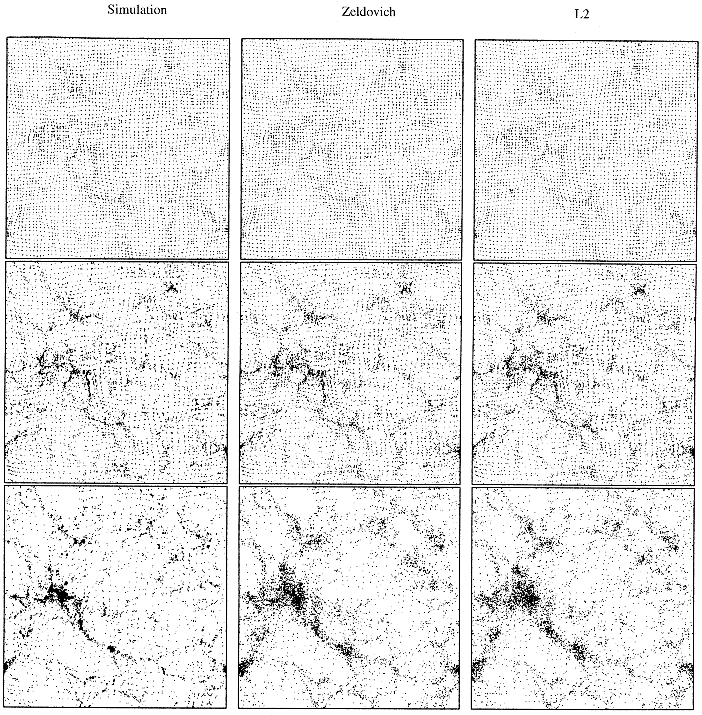


Figure 9.1: A comparison of an N-body models of the non-linear evolution of structure (left), the Zeldovich approximation (middle) and second order Lagrangian perturbation theory (right). Figure taken from Bouchet et al. 1995.

Thus, subtracting the unperturbed parts of the above equations ($\mathbf{g}_0$, Φ_0 and ρ_0)[1], we arrive at the full equations of motion:

$$\ddot{\mathbf{r}} = -2\frac{\dot{a}}{a}\dot{\mathbf{r}} - \frac{\nabla_r \Delta\Phi}{a^2} \tag{9.5}$$

$$\frac{\nabla_r^2 \Delta\Phi}{a^2} = 4\pi G \Delta\rho \tag{9.6}$$

this is actually *identical* to the linearised equation for non-relativistic fluids that we derived previously. It turns out that the linearised equation is correct for arbitrary over-density $\Delta\rho$!

Before turning to the numerical solution of equations 9.5 and 9.6, it is worth looking at two analytic approaches that give us some important insight.

9.2 The Zeldovich approximation

A first obvious thing to try in moving into the non-linear regime is to add higher order terms to our linear perturbation theory (§8). In fact, we can do better by perturbing instead about the above equations in co-moving coordinates (since these are exact). This was a key new idea due to Zel'Dovich 1970. The nature of the approximation is to assume that the co-moving coordinates can be decomposed into a time independent part $\mathbf{q}$ and a time dependent part that stretches the initial perturbation field $f(\mathbf{q})$ with time:

$$\mathbf{r} = \mathbf{q} + b(t)\mathbf{f}(\mathbf{q}) \tag{9.7}$$

[1] This relies on the infamous *Jeans swindle* after Jeans (1902). It is, of course, dodgy to "subtract away" the unperturbed gravitational field. In the unperturbed limit, we have $\dot{\mathbf{r}} = \ddot{\mathbf{r}} = 0$ and therefore $\nabla\Phi_0 = 0$. But we must have from the Poisson equation that $\nabla^2\Phi_0 = 4\pi G\rho_0$. These can only both be true if $\rho_0 = 0$! The reason the swindle works is really just that it leads to the answer one gets if doing a proper linearised analysis in GR. Such an approach is beyond the scope of this course, however, and we must accept the swindle for now.

Now, initially $t = 0$, $b(t) = 0$, $\mathbf{r} = \mathbf{q}$ and the density is homogeneous $\rho = \rho_0$. We can think of equation 9.7 as a time dependent *map* from $\mathbf{q}$ (a homogeneous Universe) to $\mathbf{r}$ (including non-linear structure growth). Thus, the density that is initially homogeneous will be mapped via the Jacobian of this transformation ($d^3\mathbf{r} = \left|\frac{\partial r_i}{\partial q_j}\right| d^3\mathbf{q}$):

$$\rho = \rho_0 \left|\frac{\partial r_i}{\partial q_j}\right|^{-1} = \rho_0 \left|\delta_{ij} + b(t)\frac{\partial f_i}{\partial q_j}\right|^{-1} \tag{9.8}$$

If we assume that the deformation matrix $\frac{\partial f_i}{\partial q_j}$ is irrotational, then it is symmetric. Diagonalising it, we may in general determine the eigenvalues and eigenvectors of this transformation $(-\alpha, -\beta, -\gamma)$, and using the fact that a matrix determinant is simply the product of the eigenvalues, we derive:

$$\rho = \rho_0 \left[(1 - b\alpha)(1 - b\beta)(1 - b\gamma)\right]^{-1} \tag{9.9}$$

For perfectly spherical collapse (see next), $\alpha = \beta = \gamma$. But in general, we will have one smaller than the other two. Thus, the Zeldovich approximation predicts that collapse will proceed first along one axis to form pancakes. As these pancakes interest, we expect to see filaments. At the intersection of filaments, nodes (see Figure 9.1). Now, let us linearise equation 9.9 assuming $b\alpha, b\beta, b\gamma \ll 1$:

$$\rho \simeq \rho_0 \left[1 + b(\alpha + \beta + \gamma)\right] \tag{9.10}$$

⇒

$$\delta = \frac{\rho - \rho_0}{\rho_0} \simeq b(\alpha + \beta + \gamma) = b\nabla \cdot f \tag{9.11}$$

Taking the divergence of the equation of motion (9.5), we can eliminate $\mathbf{f}$ to obtain an evolution equation for $b(t)$:

$$\nabla \cdot \ddot{\mathbf{r}} = -2\frac{\dot{a}}{a}\nabla \cdot \dot{\mathbf{r}} - \frac{\nabla_r^2 \Delta\Phi}{a^2} \tag{9.12}$$

⇒

$$\ddot{b}\nabla \cdot \mathbf{f} = -2\frac{\dot{a}}{a}\dot{b}\nabla \cdot \mathbf{f} - \frac{4\pi G\rho_0 b}{a^2}\nabla \cdot f \tag{9.13}$$

where we used equation 9.11 to substitute for the over-density δ. Thus:

$$\ddot{b} = -2\frac{\dot{a}}{a}\dot{b} - \frac{4\pi G\rho_0}{a^2}b \tag{9.14}$$

and we may solve for $b(t)$ and, through equation 9.11, also for $\mathbf{f}$.

The above is typically used to set up just-beyond-linear initial conditions for the full N-body simulations that we will describe in §9.5. A comparison of the Zeldovich approximation, second order Lagrangian perturbation theory and an N-body simulation (of which more in a moment) are given in Figure 9.1.

9.3 Spherical collapse

You will have looked at this approximation already on the problem sheet. The special case of spherical collapse is particularly amenable to analytic treatment. The trick is to think of a spherical perturbation as a mini-Universe where, from Birkoff's theorem, we may think of the density now as the *mean enclosed density* $\overline{\rho}$. Thus, our perturbation is described by the Friedmann equation for a matter dominated Universe:

$$\dot{r}^2 - \frac{8\pi G}{3}\overline{\rho}r^2 = -kc^2 \tag{9.15}$$

which, writing $M = 4/3\pi r^3\overline{\rho}$ becomes:

$$\dot{r}^2 - \frac{2GM}{r} = -kc^2 \tag{9.16}$$

and we can simply write down the cycloid solution for closed Universes we already derived in §7:

$$r = r_*(1 - \cos\alpha) \qquad ; \qquad t = t_*(\alpha - \sin\alpha) \tag{9.17}$$

Substituting the above into the Friedmann equation, we derive similarly to equation 7.20:

$$\dot{r}^2 = \frac{r_*^2}{t_*^2}\left[2\frac{r_*}{r} - 1\right] \tag{9.18}$$

and matching terms with the Friedmann equation, we derive:

$$r_* = (GMt_*^2)^{1/3} \tag{9.19}$$

$$t_* = r_*/(\sqrt{k}c) \tag{9.20}$$

We can now use this simple model to study the evolution into the deep non-linear regime. We will not gain the intuition that pancakes and filaments should form, as we saw from the Zeldovich approximation (§9.2), since we assume spherical symmetry. But we can imagine that this approach is reasonable at a node connecting three intersecting pancakes that is point-like. This is where we expect the highest density structures to form that will ultimately be the sites of galaxy formation.

Structure formation proceeds in three stages:

1. **Turnaround:** The sphere reaches a maximum radius when $\dot{r} = 0$, for which $r = 2r_*; \alpha = \pi; t = \pi t_*$.

2. **Collapse:** If perfectly spherical, the collapse will proceed to a singularity (see the problem sheet for a discussion of this). This occurs for $r = 0; \alpha = 2\pi; t = 2\pi t_*$.

3. **Virialisation:** In practice, slight departures from perfect symmetry will cause the system to virialise before it collapses to a point. Virialisation means that $2T + V = 0$, where T is the total kinetic, and V the total potential energy, respectively (c.f. §2). This occurs at a radius r_V where $\dot{r}_V^2 = GM/r_V$, which from the Friedmann equation gives: $GM/r_V = kc^2$ and therefore $r_V = r_*; \alpha = 3\pi/2; t = (3\pi/2 + 1)t_*$.

The density contrast at 'collapse' can be estimated either at $r = 0$ (full collapse) or $r = r_*$ (virialisation). The expansion rate of the 'background' can be estimated by extrapolating the small t growth rate of our perturbation into the distant future. Small t means $r \ll r_*$ and from equation 9.18, we have:

$$\dot{r} \simeq \sqrt{2}\frac{r_*^{3/2}}{t_* r^{1/2}} \tag{9.21}$$

which integrating gives:

$$r_b \simeq r_* \left(\frac{9}{2}\right)^{1/3} \left(\frac{t}{t_*}\right)^{2/3} \tag{9.22}$$

where now r_b is the expansion rate of the background. At virialisation, we can then estimate the local overdensity as:

$$\delta \simeq \left(\frac{r}{r_b}\right)^{-3} \tag{9.23}$$

which gives $\delta_V \simeq \frac{9}{2}(3\pi/2 + 1)^2 = 147$. Some authors argue, however, that $r_V = r_*$ is not reached until the actual collapse time, for which $t_c = 2\pi t_*$ rather than $t_V = (3\pi/2 + 1)t_*$. Using this assumption, we derive a slightly larger overdensity of $\delta_c \simeq \frac{9}{2}(2\pi)^2 = 178$ – the 'collapse' overdensity.

The above overdensities are typically used as a rule of thumb to define the 'edge' of bound structures – called dark matter *halos* – that form in cosmological N-body simulations. We define here the 'virial mass' M_V and 'virial radius' r_V as the mass and radius where the mean enclosed density is χ times the background density:

$$\overline{\rho} = \frac{M_V}{4/3\pi r_V^3} = \chi\rho_0 \tag{9.24}$$

where the exact value for χ depends on the cosmology and on whether we assume virialisation or 'collapse' to define the edge. Typically, $\chi \sim 200$ is assumed for our standard cosmological model (ΛCDM).

9.4 The statistics of halo formation

Armed with the spherical top hat collapse model, we can now study the *statistics* of the formation of bound dark matter structures called dark matter *halos*. The initial power spectrum of perturbations is given from the transfer function (the fit to full Boltzmann code calculations in the early Universe; equation 8.33) by (Eisenstein and Hu 1999):

$$\frac{k^3}{2\pi^2}\delta_H^2\left(\frac{ck}{H_0}\right)^{3+n}\frac{T^2(k,z)D_1^2(z)}{D_1^2(0)} \tag{9.25}$$

where δ_H is the amplitude of perturbations on the Horizon scale, k is the wavenumber, n is the power law index of initial perturbations ($n = 1$ is scale invariant) and $D_1(z)$ is the linear growth factor of the Universe (equation 8.28).

The Press-Schechter approach and its refinement – Excursion-set theory – assumes that each wavelength in the power spectrum collapses individually in a spherical top-hat collapse (Press and Schechter 1974; Bond *et al.* 1991). With this assumption, we can calculate the fraction of bound structures of a given mass in a given volume at a given redshift: the *mass function* of halos. The semi-analytic theory gives a remarkable fit to full N-body simulations, but we do not have space to discuss it in detail here. A good review can be found in Zentner 2007.

9.5 N-body models

The full numerical solution of equations 9.5 and 9.6 is usually done by means of *N-body* simulations (e.g. Dehnen and Read 2011). The idea is to sample the overdensity field using discrete sampling points called 'particles' (not to be confused with sub-atomic particles!). The 'particles' (we will drop the quotation marks from here on), then evolve according to the equations of motion: equations 9.5 and 9.6. First, we solve Poisson's equation for the overdensity, next the forces are estimated and the particles are integrated forward over some small interval in time.

One problem immediately presents itself: the sheer size of the Universe. We cannot hope to reasonably sample small scales if simulating the whole shebang. Typically this is circumvented by modelling instead a cubic patch of the Universe of volume L^3. Periodic boundary conditions are applied so that the simulated Universe is really a infinity of replicas of the small patch. This approximation is acceptable so long as the Universe is not collapsing on scales comparable to the box size. This necessarily sets a limit to the redshift down to which a box of a given size can be reliably evolved. The Universe at redshift $z = 0$ is still linear on scales ~ 8 Mpc today, thus the minimum box size required to evolve a simulation to the present time should be a few times larger than this in co-moving coordinates ($\gtrsim 24$ Mpc co-moving).

9.5.1 Solving Poisson's equation

Ignoring the periodic boundary conditions for a moment, the first challenge is to solve Poisson's equation for the particles. One simple possibility is to treat the particles as discrete point masses. Then the (Newtonian) force on a particle i follows from a simple sum over particles j:

$$\mathbf{F}_i = \sum_j^N \frac{Gm_im_j(\mathbf{x}_j - \mathbf{x}_i)}{|\mathbf{x}_i - \mathbf{x}_j|^3} \tag{9.26}$$

where $\mathbf{F}_{ij}$ is the force between particle pairs i and j at positions $\mathbf{x}_i$ and $\mathbf{x}_j$ and N is the total number of particles.

This runs into two computational problems. The first is that we must compute $O(N)$ sums for each particle and thus the algorithm scales as $O(N^2)$ which is very slow (i.e. if I increase the number of particles by a factor 10, the computational costs will increase 100 fold!). Secondly, recall that

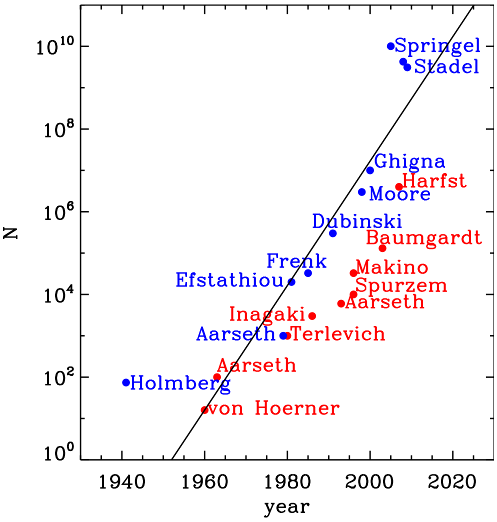


Figure 9.2: The increase in particle number in N-body simulations over the past 50 years for selected collisional (red) and collisionless (blue) N-body simulations[1]. The solid line shows the scaling $N = N_0 2^{(\text{year}-y_0)/2}$ (with $N_0 = 16$ and $y_0 = 1960$ valid for von Hoerner's calculation) expected from Moore's law if the costs scale linearly with N.

these 'particles' are merely sampling points in the density field. If two approach one another, they should *not* really behave like giant point masses. Yet equation 9.26 has that $\mathbf{F}_i$ *diverges* for $\mathbf{x}_i \to \mathbf{x}_j$. This latter problem is typically solved by introducing a *force softening* ϵ such that the force equation becomes:

$$\mathbf{F}_i = \sum_j^N \frac{G m_i m_j (\mathbf{x}_j - \mathbf{x}_i)}{(\epsilon^2 + |\mathbf{x}_i - \mathbf{x}_j|^2)^{3/2}} \tag{9.27}$$

which removes the diverging force for approaching particles.

It is clear that in the limit $N \to \infty$; $\epsilon \to 0$, the above force equation approaches the correct Newtonian dynamics. But this not not necessarily mean that we will converge for finite N, ϵ. The thorny issue of force softening and its relation to real collisionless fluid equations is beyond the scope of this present course (see Dehnen and Read 2011 for a discussion). But it is worth mentioning a useful rule of thumb for determining the magnitude of ϵ. We require that the maximum acceleration on a particle ($a_{\text{max}} \simeq Gm/\epsilon^2$; where m is the particle mass) is less than the minimum mean-field acceleration ($a_{\text{min}} \simeq GM_{\text{tot}}/R^2$; where M_{tot} and R are the total mass and scale length of the system). This gives:

$$\epsilon > \frac{R}{\sqrt{N_{\text{tot}}}} = \epsilon_{\text{min}} \tag{9.28}$$

Of course, in a cosmological context, it is tricky to decide on what the scale R should correspond to, but we can think of it as the virial radius of the most massive collapsed object in the box. In practice, a few times ϵ_{min} is a 'safe' choice. This may all seem a bit voodoo at this stage and, unfortunately, it remains so at present. There have been many papers exploring different force softening formulae (called *Kernels*; e.g. Dehnen 2001) and varying softening (e.g. Power *et al.* 2003). Luckily most simulations are not sensitive to these choices and produce converged results for quite different (but reasonable) choices.

We still have the second problem of $O(N^2)$ scaling, however. We can significantly improve on this by utilising the fact that close interactions are in any case softened. This is incredibly useful and allows the $O(N^2)$ force calculation for direct summation to be reduced to $O(N)\ln(N)$, or even $O(N)$. Such techniques have allowed modern collisionless simulations to keep pace with Moore's law (that computer power doubles every two years), which is not the case for direct (collisional) $O(N^2)$ calculation of the force (see Figure 9.2). We now describe two popular methods for achieving this vast improvement in speed.

[1] Our selection was taken from the review paper Dehnen and Read 2011.

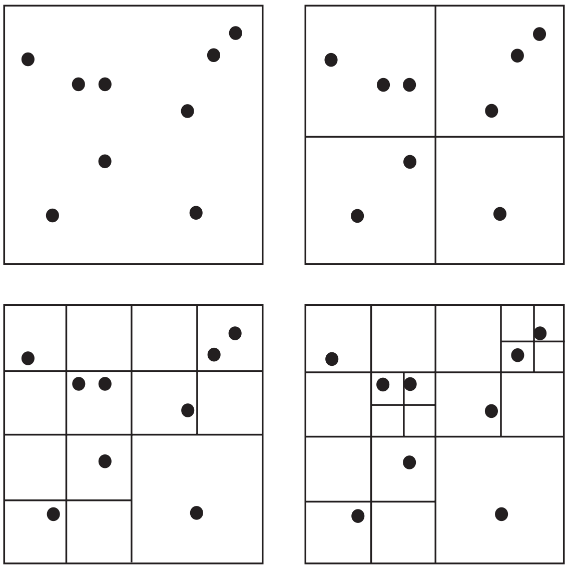


Figure 9.3: Subdivision of space in to a *grid* or *mesh*. The particles are shown as black filled circles.

9.5.1.1 Fourier techniques

In the Fourier method, we divide space up into a grid or *mesh* (see Figure 9.3). It is assumed that the the matter in each cell is concentrated at the centre. Provided the mesh is fine enough that it is smaller than the mean inter-particle separation, the fact that our system is collisionless means that we do not need to resolve below this scale. (This is ultimately the reason why collisionless systems are so much easier to deal with than, for example, gas dynamical systems where the smallest scale which ought to be resolved is the molecular mean free path.)

The key idea is then to write the gravitational potential as a *convolution*:

$$\Phi(\mathbf{x}) = \int \mathcal{G}(\mathbf{x} - \mathbf{x}')\rho(\mathbf{x}')d^3\mathbf{x}' = \mathcal{G} * \rho \tag{9.29}$$

which defines the *Green's* function for the Poisson equation:

$$\mathcal{G} = -\frac{G}{|\mathbf{x} - \mathbf{x}'|} \tag{9.30}$$

As above, it is more usual to use a *softened* Green's function that does not diverge for $\mathbf{x} = \mathbf{x}'$. For Plummer force softening, we have:

$$\mathcal{G} = -\frac{G}{\sqrt{\epsilon^2 + |\mathbf{x} - \mathbf{x}'|^2}} \tag{9.31}$$

As you may remember, the Fourier transform of a convolution is given by:

$$F.T.\,\{\Phi(\mathbf{x})\} = \tilde{\mathcal{G}}(\mathbf{k})\tilde{\rho}(\mathbf{k}) \tag{9.32}$$

where $\tilde{\mathcal{G}}(\mathbf{k}) = F.T.\,\{\mathcal{G}(\mathbf{x})\}$, and similarly for $\tilde{\rho}(\mathbf{k})$.

Now, we know $\tilde{\mathcal{G}}(\mathbf{k})$ analytically and so the only hard work which remains is in finding $\tilde{\rho}(\mathbf{k})$. This may be done very rapidly by using the method of *Fast Fourier Transforms (FFTs)*. For more information on this algorithm, the reader is referred to the excellent Press *et al.* 1992. Thanks to the Fast Fourier transform, this method scales as $O(N \ln N)$ which is a dramatic improvement on N^2. Forces may be similarly calculated by noting that:

$$\nabla_x\Phi(\mathbf{x}) = \int \nabla_x\mathcal{G}(\mathbf{x} - \mathbf{x}')\rho(\mathbf{x}')d^3\mathbf{x}' = \nabla_x\mathcal{G} * \rho \tag{9.33}$$

which simply gives us a different – but also analytic – Green's function for the force calculation.

There is some additional complication in how the particles are mapped onto the grid cells and how the forces are then mapped back on to the particles. Also, in practice, adaptive meshes are often employed rather than a fixed grid to put resolution only where it is needed. A more detailed account of this method that includes these complications is presented in Binney and Tremaine 2008.

9.5.1.2 Tree techniques

The other obvious thing to do is to solve the multipole expansion. In practice, this is often combined with *tree* techniques. The density is represented by particles, as in the direct summation technique, but now we divide up space into a *tree* structure (see Figure 9.4). At the base of the tree is the *root node*. This is then subdivided into *branches* of the tree which are themselves subdivided until we arrive at one particle per sub-division – the *leaves* of the tree. The tree can be built by dividing space in a number of different ways. A popular choice is the *oct-tree* where each *parent* cube is divided into eight equal *children* (also called the *Barnes and Hut* oct-tree after Barnes and Hut 1986). This is useful because all cells are cubic. But other, more complicated, choices can be better. Binary trees, for example, divide the cubes into two halves which leads to rectangular cells, but a more adaptive (and therefore more efficient) space division (see e.g. Stadel 2001).

Having built the tree, we calculate the potential of each tree node as:

$$\Phi_{\text{node}}(\mathbf{r}) = -G\int_{\text{node}} \mathrm{d}^3\mathbf{x}\frac{\rho(\mathbf{x})}{\sqrt{\epsilon^2+|\mathbf{r}-\mathbf{x}|^2}} \tag{9.34}$$

where $\mathbf{x}$ is the distance to the centre of mass of the node, and we have used the *softened* potential corresponding to the softened force of equation 9.27 (other choices of softening Kernel are also possible; see e.g. Dehnen 2001). For particle simulations, the density within the node is a sum over delta functions:

$$\rho_{\text{node}}(\mathbf{x}) = \sum_\alpha m_\alpha \delta(|\mathbf{x}-\mathbf{x}^\alpha|) \tag{9.35}$$

where $\mathbf{x}^\alpha$ is the distance from the centre of mass of the node to one of the particles.

Substituting equation 9.35 in 9.34 then gives:

$$\Phi_{\text{node}}(\mathbf{r}) = -G\sum_\alpha \frac{m_\alpha}{\sqrt{\epsilon^2+|\mathbf{r}-\mathbf{x}^\alpha|^2}} \tag{9.36}$$

Now, since $|\mathbf{r}| \gg |\mathbf{x}^\alpha|$, we may Taylor expand[2] the square root to give:

$$\Phi_{\text{node}}(\mathbf{r}) = -G\sum_\alpha m_\alpha\left(\frac{1}{s}+\frac{r_i x_i^\alpha}{s^3}+\frac{3}{2}\frac{r_i x_i^\alpha r_j x_j^\alpha}{s^5}+...\right) \tag{9.37}$$

where $s^2 = \epsilon^2 + |\mathbf{r}|^2$ and we use the summation convention[3].

The above is just the multipole expansion for the node in Cartesian coordinates. The first term is the monopole, the second the dipole – that must be zero because we use coordinates about the centre of mass, and the third is the quadrupole. It is useful because the dependence on $\mathbf{r}$ now falls out linearly: we may calculate these multipole sums for each node and then sum over all nodes to obtain the potential at a given point. This presents us with a trade-off between building more branches in the tree and including more multiple moments: both give increased force accuracy. Branching the tree is controlled by the *opening angle* θ, which is defined by comparing the size of the quadrupole term with the monopole term for a node:

$$\frac{1}{s^5}\sum_\alpha m_\alpha x_i^\alpha x_j^\alpha r_i r_j < \theta^2\frac{1}{s}\sum_\alpha m_\alpha \tag{9.38}$$

If the size of one size of the cubic node is $\sim l$ and $s \sim r$, then the above reduces to:

$$\frac{l}{r} < \theta \tag{9.39}$$

[2] see Appendix D.

[3] see appendix B.

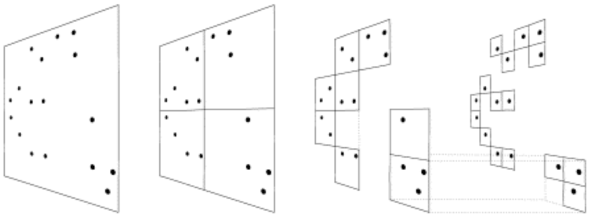


Figure 9.4: Schematic illustration of the Barnes and Hut oct-tree in two dimensions. The particles are first enclosed in a square (root node). This square is then iteratively subdivided in four squares of half the size, until exactly one particle is left in each final square (the leaves of the tree). In the resulting tree structure, each square can be a 'parent' of up to four 'children'. Note that empty squares need not be stored. For a three-dimensional simulation, the tree nodes are cubes instead of squares.

which is the branching criteria proposed by Barnes and Hut 1986. Other branching criteria can compare the size of higher order moments (e.g. hexadecapole) giving different trade-offs between having more branches on the tree versus a more accurate force calculation for each node (see e.g. Springel *et al.* 2001).

The above algorithm scales as $O(N)\ln N$. This can be simply understood by considering a binary tree for a constant density particle distribution. The space is continually subdivided until we have one particle per cell − this is $\sim$ the scale of the force softening ϵ. For constant density, each subdivision halves the number of particles. Thus, the total number of particles N can be written as n subdivisions:

$$N = 2^n \tag{9.40}$$

$\Rightarrow$

$$\ln N = n\ln(2) \tag{9.41}$$

and after $n \sim \ln N$ subdivisions, we reach the leaf nodes. Then, for each particle we 'walk the tree' to calculate the force, by summing over the nodes. This requires $\ln N$ force computations per particle giving overall $\sim O(N)\ln N$ time.

In fact, we can do even better than the above by calculating the forces between *nodes*, rather than between particles and using the symmetry of the gravitational force between node-paris. A careful ordering of the sums can reduce the order of the algorithm further to $O(N)$ (see Dehnen 2000). This improvement is just becoming really important. With state-of-the-art simulations now using $O(10^9)$ particles, we can obtain speed-ups of $\sim$ 20 by eliminating the $\ln N$ dependence. This is why, for gravity only simulations, tree techniques remain faster for a given accuracy than Fourier methods, and are the de-facto choice.

9.5.2 Periodic boundary conditions

So far, we have ignored the periodic boundary conditions. For Fourier methods, these are easy and natural to implement since the Fourier transform *implicitly* applied periodicity. More complicated is applying periodic boundary conditions for tree codes. This is typically done using Ewald's method (Ewald 1921), which was originally invented for solid-state physics and imported to this field by Hernquist *et al.* 1991. We defer the interested reader to these texts (but note an error in their eq. 2.14b as pointed out by Klessen 1997).

9.5.3 Time integration

9.5.3.1 The Simple Euler integrator

This section largely follows our review article Dehnen and Read 2011. Having calculated the force on the particles, we must then evolve them forwards in time. It is tempting to use the simple 'Euler method'. Defining a *timestep* Δt_i for a particle i, we can update its position and velocity as:

$$\mathbf{x}_i(t+\Delta t) = \mathbf{x}_i(t) + \dot{\mathbf{x}}_i\Delta t_i \tag{9.42}$$

$$\dot{\mathbf{x}}_i(t + \Delta t) = \dot{\mathbf{x}}_i(t) + \ddot{\mathbf{x}}_i \Delta t_i \tag{9.43}$$

where $\ddot{\mathbf{x}}_{i,0}$ is the acceleration evaluated at t_0. However, while this is conceptually straightforward, such a scheme performs very poorly in practice. The Euler method is nothing more than a Taylor expansion in Δt about t to first order. Thus, the errors will be proportional to Δt^2. We can significantly improve on this at little additional computational cost by using either a *symplectic integrator* and/or a 'higher order' integrator – i.e. one that has an error that goes as Δt^n, with $n > 2$. Symplectic integrators precisely solve an approximate Hamiltonian and have the advantage that, as a result, energy is manifestly conserved in a time-averaged sense. This means that energy errors are bounded and will not grow even over many thousands of dynamical times. Higher order integrators give smaller errors for the same timestep, but do not necessarily conserve energy. Here we focus only on the symplectic "leap-frog" integrator that is the de-facto choice for cosmological simulations. Higher order and non-symplectic integrators are discussed in Dehnen and Read 2011.

9.5.3.2 The Leapfrog integrator

The leapfrog integrator is an example of a *symplectic integrator*. The idea is to replace the Hamiltonian H, with an approximate form:

$$\tilde{H} = H + H_{\rm err} \tag{9.44}$$

where $H_{\rm err}$ is the error Hamiltonian. Provided that $\tilde{H}$ and H are time-invariant, the energy error is bounded at all times (e.g. Yoshida 1993). The goal is to find a $\tilde{H}$ that can be solved *exactly* by simple numerical means and that minimises $H_{\rm err}$. Defining the combined phase-space coordinates $\mathbf{w} = (\mathbf{x}, \mathbf{p})$ we can re-write Hamilton's equations as:

$$\mathcal{H}\mathbf{w} = \dot{\mathbf{w}}, \tag{9.45}$$

where $\mathcal{H} \equiv \{\cdot, H\}$ (with $\{A, B\} \equiv \partial_{\mathbf{x}} A \cdot \partial_{\mathbf{p}} B - \partial_{\mathbf{x}} B \cdot \partial_{\mathbf{p}} A$ the Poisson bracket) is an *operator* acting on $\mathbf{w}$. Equation (9.45) has the formal solution

$$\mathbf{w}(t + \Delta t) = \mathrm{e}^{\Delta t\, \mathcal{H}}\, \mathbf{w}(t) \tag{9.46}$$

where we can think of the operator $\mathrm{e}^{\Delta t\, \mathcal{H}}$ as a symplectic map from t to $t + \Delta t$. This operator can be split into a succession of discrete but symplectic steps, each of which can be *exactly* integrated. The most common choice is to separate out the kinetic and potential energies, $H = T(\mathbf{p}) + V(\mathbf{x})$, such that we can split

$$\mathrm{e}^{\Delta t\, \mathcal{H}} = \mathrm{e}^{\Delta t\, (\mathcal{T} + \mathcal{V})} \simeq \mathrm{e}^{\Delta t\, \mathcal{V}}\, \mathrm{e}^{\Delta t\, \mathcal{T}} = \mathrm{e}^{\Delta t\, \tilde{\mathcal{H}}}. \tag{9.47}$$

Because the operators $\mathcal{T} \equiv \{\cdot, T\}$ and $\mathcal{V} \equiv \{\cdot, V\}$ are *non-commutative*, the central relation in equation (9.47) is only approximate. This operator splitting is extremely useful, because, while equation (9.45) has in general no simple solution, the equivalent equations for each of our new operators do:

$$\mathrm{e}^{\Delta t\, \mathcal{T}} \begin{bmatrix} \mathbf{x} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{x} + \Delta t\, \mathbf{p} \\ \mathbf{p} \end{bmatrix} \qquad \text{and} \qquad \mathrm{e}^{\Delta t\, \mathcal{V}} \begin{bmatrix} \mathbf{x} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ \mathbf{p} - \Delta t\, \nabla V(\mathbf{x}) \end{bmatrix}. \tag{9.48}$$

These operations are also known as *drift* and *kick* operations, because they only change either the positions (drift) or velocities (kick). Note that the drift step in (9.47) is identical to the simple Euler method (equation 9.42), while its kick step is *not* identical, because the acceleration is calculated using the drifted rather than the initial positions. The integrator that applies a drift followed by a kick (equation 9.47) is called *modified* Euler scheme and is symplectic.

It is clear from the similarity between modified and un-modified Euler schemes that both are only first order accurate. The error creeps in because of the approximation used to split the operators in equation (9.47). We can do better by concatinating many appropriately weighted kick and drift steps:

$$\mathrm{e}^{\Delta t\, \tilde{\mathcal{H}}} = \prod_i^N \mathrm{e}^{a_i \mathcal{V}}\, \mathrm{e}^{b_i \mathcal{T}} = \mathrm{e}^{\Delta t\, \mathcal{H} + \mathcal{O}(\Delta t^n)} \tag{9.49}$$

with coefficients a_i and b_i chosen to obtain the required order of accuracy n. From equation (9.49), we see that: (i) the approximate Hamiltonian $\tilde{H}$ is solved exactly by the successive application of the

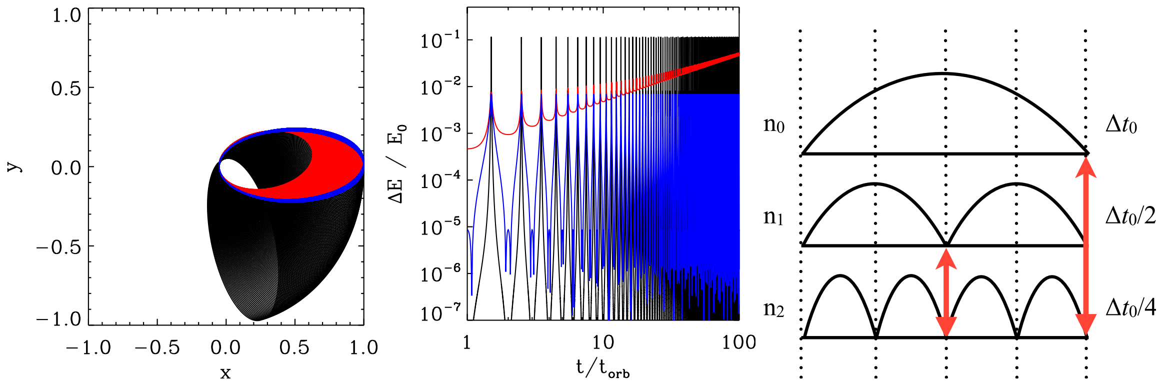


Figure 9.5: **Left** Comparison of the Leap Frog integrator (black); a 4th order non-symplectic Hermite scheme (red); and time symmetric Leap Frog using variable timesteps (blue) for the integration of an elliptic ($e = 0.9$) Kepler orbit over 100 periods. In the first two cases a fixed timestep of $\Delta t = 0.001\,t_{\rm orb}$ was used. **Middle** The fractional change in energy for the Kepler problem for various flavours of the Leap Frog integrator: fixed timesteps (black); variable timesteps (red); and symmetric variable timesteps (blue). **Right** A schematic diagram of a variable timestep scheme, where particles are arranged in a hierarchy of timestep rungs in powers of two. Particles can move between rungs at synchronous steps shown in red.

kick and drift operations (Yoshida 1993); and (ii) $\tilde{H}$ approaches H in the limit $\Delta t \to 0$, and/or the limit $n \to \infty$. At second order ($n = 2$), and choosing coefficients that minimise the error, we derive the *leapfrog* integrator: $\mathrm{e}^{\Delta t\,\mathcal{H}+\mathcal{O}(\Delta t^3)} = \mathrm{e}^{\frac{1}{2}\Delta t\,\mathcal{V}}\mathrm{e}^{\Delta t\,\mathcal{T}}\mathrm{e}^{\frac{1}{2}\Delta t\,\mathcal{V}}$. Writing out each of these operations using equations (9.48) we have (subscripts 0 and 1 refer to times t and $t + \Delta t$, respectively):

$$\begin{aligned}
\dot{\mathbf{x}}' &= \dot{\mathbf{x}}_0 + \tfrac{1}{2}\Delta t\,\ddot{\mathbf{x}}_0 && (9.50)\\
\mathbf{x}_1 &= \mathbf{x}_0 + \quad \Delta t\,\dot{\mathbf{x}}' && (9.51)\\
\dot{\mathbf{x}}_1 &= \dot{\mathbf{x}}' + \tfrac{1}{2}\Delta t\,\ddot{\mathbf{x}}_1 && (9.52)
\end{aligned}$$

where $\ddot{\mathbf{x}}_0 = -\nabla V(\mathbf{x}_0)$ and $\ddot{\mathbf{x}}_1 = -\nabla V(\mathbf{x}_1)$, while the intermediate velocity $\dot{\mathbf{x}}'$ serves only as an auxiliary quantity. Combining equations (9.50-9.52) we find

$$\begin{aligned}
\mathbf{x}_1 &= \mathbf{x}_0 + \Delta t\,\dot{\mathbf{x}}_0 + \tfrac{1}{2}\Delta t^2\,\ddot{\mathbf{x}}_0 && (9.53)\\
\dot{\mathbf{x}}_1 &= \dot{\mathbf{x}}_0 + \tfrac{1}{2}\Delta t\,(\ddot{\mathbf{x}}_0 + \ddot{\mathbf{x}}_1) && (9.54)
\end{aligned}$$

which are the familiar Taylor expansions of the positions and velocities to second order in Δt.

In principle, one may combine as many kick and drift operations as we choose to raise the order of the scheme. However, it is impossible to go beyond second order without having at least one a_i and one b_i coefficient in equation (9.49) be negative (Sheng; Suzuki 1989; 1991). This involves some *backwards* integration which is problematic when using varying timesteps—especially if time symmetry is required[4].

In practice, most modern codes also employ *variable timesteps*, typically in a hierarchy of timestep *rungs* (see Figure 9.5, right panel). This breaks the symplectic nature of the leap frog integrator and in principle time averaged energy conservation is no longer guaranteed. This can be circumvented by the use of *time-symmetric* variable timesteps (Dehnen and Read 2011). While not symplectic, time symmetric schemes also show excellent long term energy conservation.

In the left panel of Figure 9.5, we compare the integration of a simple Kepler orbit with an eccentricity of $e = 0.9$ for the leapfrog integrator with fixed (black) and variable (blue) timesteps, and a non-symplectic fourth order 'Hermite' integrator with fixed timesteps (red; see Dehnen and Read 2011 for details). The middle panel compares energy conservation for the leap-frog integrator with

[4] Recently, Chin and Chen 2005 have constructed fourth-order symplectic integrators which require only forwards integration. To achieve this, rather than eliminate all the errors by appropriate choice of the coefficients a_i and b_i, they *integrate* one of the error terms thus avoiding any backward step. Their method requires just two force and one force gradient evaluation per time step. It has not yet found wide application in N-body dynamics, but could be a very promising avenue for future research.

fixed (black), variable (red) and variable time-symmetric (blue) timesteps. For the leapfrog integrator with fixed timesteps (black), the energy fluctuates on an orbital time scale, but is perfectly conserved in the long term. This can be seen also in the orbit (left panel) that precesses, but does not decay. By contrast, the Hermite integrator that is not symplectic but is more accurate shows smaller phase error, but does decay with time. Best of all is the leapfrog integrator with variable symmetric timesteps (blue). This has very small orbital error (left panel), and excellent long-term error properties.

The time symmetric variable time step leapfrog used $\sim$ a quarter of the force calculations required for the fixed-step integration while giving over an order of magnitude better energy conservation. This is why variable timesteps are an essential ingredient in modern N-body calculations.

9.5.4 Initial conditions

These are usually set up by distorting a lattice of particles using the Zeldovich approximation (§9.2). Typically, this allows us to evolve the CMB fluctuations at $z \sim 1000$ down to $z \sim 50$. This is a huge advantage since the Universe is so close to being homogeneous at these early times that the tiny differences in force cause numerical problems, particularly for tree codes.

The initial power spectrum of perturbations is taken from the numerically calculated transfer function (equation 9.25) for an assumed Universe-composition. As already discussed, these fluctuations depend both on the physics of the very early Universe and on the nature of dark matter (c.f. Figure 8.1). We will focus here on the effect of the dark matter fluid on these initial fluctuations since this is the focus of this course. The effects cannot be strong on large scales or we would fail to successfully fit the CMB data. The effects on small scales are not well-probed by the CMB, but these evolve into the non-linear regime and so can be probed instead in the nearby Universe. This is why non-linear structure formation will give us unique information about the nature of dark matter. We will consider just one example of modified dark matter here: warm dark matter.

9.6 Warm versus cold dark matter

So far, we have assumed that dark matter is a collisionless non-relativistic fluid. But suppose that it starts out relativistic for a time and then undergoes a phase transition to a non-relativistic fluid. In this case, structure formation will be suppressed during the relativistic phase. Whether this happens or not depends on what dark matter is. If we suppose that it is some new particle, then heavy particles will be non-relativistic, while light particles can show some initial relativistic behaviour, depending on how they are produced (e.g. Boyarsky *et al.* 2009b). The energy of such a dark matter particle of mass m_χ is given by:

$$E^2 = m_\chi^2 c^4 + p^2 c^2 = \gamma^2 m_\chi^2 c^4 \tag{9.55}$$

which, rearranging, gives the particle velocity v_p as:

$$\frac{v_p}{c} = \frac{pc}{\sqrt{m_\chi^2 c^4 + p^2 c^2}} \tag{9.56}$$

Now, in an expanding FLRW metric Universe, the momentum scales inversely proportional to the scale factor a: $p \propto m_\chi / a$ (this just follows from Hubble's law). Thus, we may write $p = m_\chi c a_{nr}/a$, where a_{nr} is the scale factor at the moment when the dark matter switches from a relativistic to a non-relativistic equation of state (at a redshift z_{nr}). Substituting this, we obtain:

$$\frac{v_p}{c} = \frac{a_{nr}}{\sqrt{a^2 + a_{nr}^2}} \tag{9.57}$$

Which is called the 'free-streaming' velocity. The total distance that dark matter can free stream up to this moment is then given by the age of the Universe at z_{nr} times the free streaming velocity. This determines the 'free streaming' length R_f (e.g. Bode *et al.* 2001):

$$R_f \simeq 0.2(\Omega_\chi h^2)^{1/3} \left(\frac{m_\chi}{1\,\mathrm{keV}}\right)^{-4/3} \mathrm{Mpc} \tag{9.58}$$

where Ω_χ is the density parameter for the warm dark matter.

The above leads to a 'filtering mass scale' (e.g. Avila-Reese *et al.* 2001):

$$M_f \simeq \frac{4\pi}{3}\eta\rho_0 R_f^3 \tag{9.59}$$

where η is the over-density for the collapsing dark matter structure. Thus, we expect dark matter halo formation to be suppressed below $\sim M_f$ at a redshift z_{nr}.

A second effect that happens is that the free streaming velocity imprints some intrinsic velocity dispersion in the dark matter (hence the name *warm* dark matter). Assuming a dispersion of $\sigma \sim v_p$, this dispersion sets a maximum phase space density for the dark matter particles (Tremaine and Gunn 1979). Assuming a Maxwellian distribution of velocities, we have:

$$f(v) = \rho_0\left(\frac{1}{2\pi\sigma^2}\right)^{3/2}\exp\left(-\frac{v^2}{2\sigma^2}\right) \tag{9.60}$$

and the maximum phase space density is $f_{\mathrm{max}} = \rho_0\left(\frac{1}{2\pi\sigma^2}\right)^{3/2}$.

The maximum phase space density sets a 'core' radius for the warm dark matter halos that is given by (Tremaine and Gunn 1979):

$$r_c \gtrsim 32\left(\frac{10\,\mathrm{km\,s^{-1}}}{\sigma}\right)^{1/2}\left(\frac{\mathrm{keV}}{m_\chi}\right)^2\,\mathrm{pc} \tag{9.61}$$

Because of the above, it is often stated that warm dark matter 'leads to large cores in dark matter halos'. Let's examine this claim a little. Suppose we want a ~ 1 kpc core inside a galaxy of mass $10^{10}\,\mathrm{M}_\odot$ (rather like the Large Magellanic Cloud that we have encountered previously). Assuming a central dispersion of $\sigma = 10\,\mathrm{km/s}$ (which is conservative), equation 9.61 gives $m_\chi = 0.18$keV. Putting this into equation 9.58, we obtain $R_f \sim 0.9$ Mpc, which would entirely erase galaxies of the size of the LMC from the Universe! (For a more detailed version of this argument, see Macciò *et al.* 2012.) Thus, the 'core' effect in warm dark matter can only be significant if we can decouple the thermal relic velocity of the particle from its free streaming length. This is possible if more exotic warm dark matter models are used (e.g. Strigari *et al.* 2007).

Lecture 10

Key results from structure formation simulations

In this lecture we present the key results from structure formation simulations for the expected dark matter distribution in the Universe. We consider both warm and cold dark matter cosmologies, as well as discussing the effect of normal baryonic matter on the observed and actual dark matter distribution.

10.1 Key results from structure formation simulations (ignoring baryonic physics)

Now that we have an understanding for how such simulations are conducted, we present here the key results from structure formation simulations under various assumptions about the cosmology, initial conditions, and the nature of the dark matter fluid. We will start by ignoring the baryons in the Universe, modelling only the dark matter fluid and therefore only gravity. In §10.2, we will discuss the role baryons play and how much more complicated things then become – particularly on small scales. An example of an N-body simulation for our standard cosmological model is given in Figure 10.1 taken from Bode *et al.* 2001.

10.1.1 The halo mass function

The simplest thing we can measure from the simulations is the number of dark matter halos of a given mass within a given volume: the *halo mass function*. The halo mass function as a function of redshift for our standard cosmology, and for a warm dark matter mode, is given in Figure 10.2 (taken from Bode *et al.* 2001). If we imagine that at the centre of each dark matter halo there is a visible galaxy, then we can compare this predicted halo mass function with the number of galaxies of a given mass in the Universe. As we discuss in 10.2, however, this comparison is complicated by the physics of galaxy formation.

Note that the low mass end of the halo mass function in the warm dark matter simulation has a sudden upturn. This used to be taken as evidence for *fragmentation*, but is now understood to be a numerical error (Wang and White 2007; and see right panel of Figure 10.2). Raising the numerical resolution moves this feature to lower masses, but extremely slowly (Wang and White 2007). This has led to the development of new and more accurate N-body techniques (e.g. Hahn *et al.* 2013; Hobbs *et al.* 2016). However, these are still at the bleeding edge of current research at the time of writing and have not yet been used for large-scale simulations.

10.1.2 The dark matter density distribution

In addition to simply adding up all of the bound dark matter structures in a given volume (the mass function), we can also study the *internal* density distribution of these dark matter structures. Dubinski and Carlberg 1991 were the first to notice that dark matter halo density profiles can be reasonably well fit by a split power law profile:

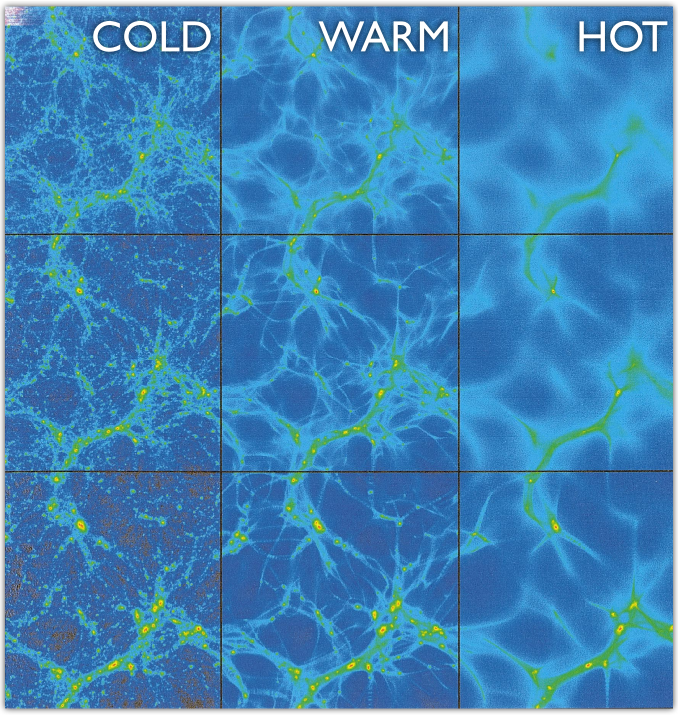


Figure 10.1: A comparison of three N-body simulations. From top to bottom the panels show the time evolution in redshift of the simulations: $z = 3, 2, 1$. From left to right, the dark matter 'temperature' is changed from cold to warm to hot dark matter (cold dark matter; $m_\chi = 350\,\mathrm{eV}$; $m_\chi = 175\,\mathrm{eV}$). Notice that as the dark matter temperature is increased, structure formation is delayed and the smallest structures are erased.

$$\rho(r) = \rho_0 \left(\frac{r}{r_0}\right)^{-\alpha} \left(1 + \frac{r}{r_0}\right)^{\alpha-\beta} \tag{10.1}$$

In the limit $r \ll r_0$, equation 10.1 tends to $\sim \rho_0 \left(\frac{r}{r_0}\right)^{-\alpha}$; in the limit $r \gg r_0$ it tends to $\sim \rho_0 \left(\frac{r}{r_0}\right)^{-\beta}$. Thus α describes in the inner logarithmic slope of the density profile and β the outer slope. Dubinski and Carlberg 1991 suggested $\alpha \sim 1$; $\beta \sim 4$ which is called a Hernquist profile (Hernquist 1990; and see Figure 10.3 left panel). Navarro *et al.* 1996b went on to show that the above split power law profile appears to be *universal*. That is, the same profile fits subhalos orbiting within a galaxy, galaxy halos, and cluster halos. They favoured a slightly shallower outer slope of $\beta = 3$ which has become known as the 'NFW' profile.

As the resolution of such simulations continues to improve it is becoming clear that a perfect split power law is not an adequate fit anymore (e.g. Merritt *et al.* 2006; Merritt *et al.* 2006; Stadel *et al.* 2009). Instead, it seems that the logarithmic slope continues to evolve smoothly as a function of radius with no clear large or small radius asymptote (see Figure 10.3, right panel).

10.1.3 The local dark matter phase space distribution function

A final interesting thing we can extract from such simulations is the *phase space density* of dark matter. This is important for experiments that hope to detect a dark matter particle in the laboratory. We will discuss this in more detail in later lectures.

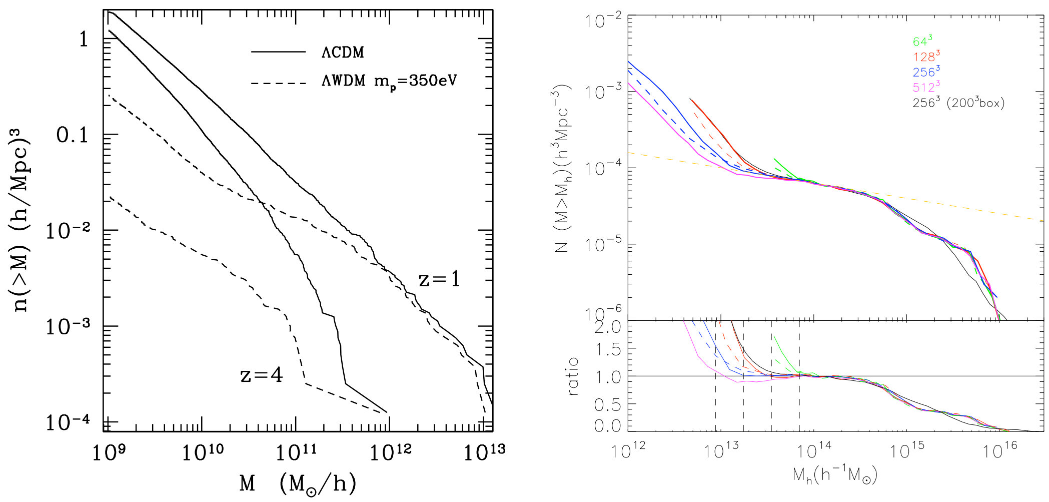


Figure 10.2: **Left:** The halo mass function: the cumulative number of halos of mass greater than M as a function of M (from Bode et al. 2001). The solid lines show results at redshifts $z = 1$ and $z = 4$ for cold dark matter (CDM); the dashed lines show the same for warm dark matter (WDM). The sudden upturn at $M \sim 5 \times 10^{10}$ for the warm dark matter simulation is a numerical error. **Right:** The halo mass function for a WDM simulation of increasing resolution (from Wang & White 2007). Notice that the upturn in the mass function moves very slowly to lower masses.

10.2 The importance of baryonic physics

So far we have discussed solving the non-linear growth of structures under gravity in an expanding FLRW spacetime. But the baryonic matter is also subject to pressure forces and other more complex physics like gas cooling, star formation, and energy injection due to exploding stars (supernovae). These processes become increasingly important as we move to smaller and smaller scales and can affect the observed and even the actual distribution of dark matter in the Universe. We can divide baryonic effects into two main branches: *observational effects* that change the way we see the Universe but do not fundamentally alter the underlying dark matter distribution; and *dynamical effects* that physical alter the dark matter structure. We discuss these in turn, next.

10.2.1 Observational effects

These baryonic effects do not physically alter the distribution of mass (mostly dark matter) in the Universe. Instead, they bias our view of the matter distribution by having a complex mapping between dark and visible structures. On very small scales, for example, inefficient star formation – that can be a function both of galaxy mass and environment – can makes small galaxies difficult to see in star light, or gas absorption/emission. If we simply count galaxies, we might expect then to see a rather different distribution that that expected from the dark Universe alone. We discuss this in more detail in the next lecture.

10.2.2 Dynamical effects

On the very smallest scales, the baryons can cool and actually dominate the gravitational potential. There it is conceivable that they actually physically alter the distribution of dark matter, even if dark matter and baryons interact only gravitationally. This is problematic since such alterations could erase information about the nature of the dark matter fluid. To explore the importance of this, let us suppose that we can crudely divide baryonic processes into two discrete events: *inflow* – i.e. how baryons get into the dark matter halos; and *outflow* – i.e. how they get out again. Furthermore, to keep things fully analytic, let us suppose that we can treat both the baryonic matter and the dark halo as point masses, and assume that the dark halo is constructed of particles moving on circular

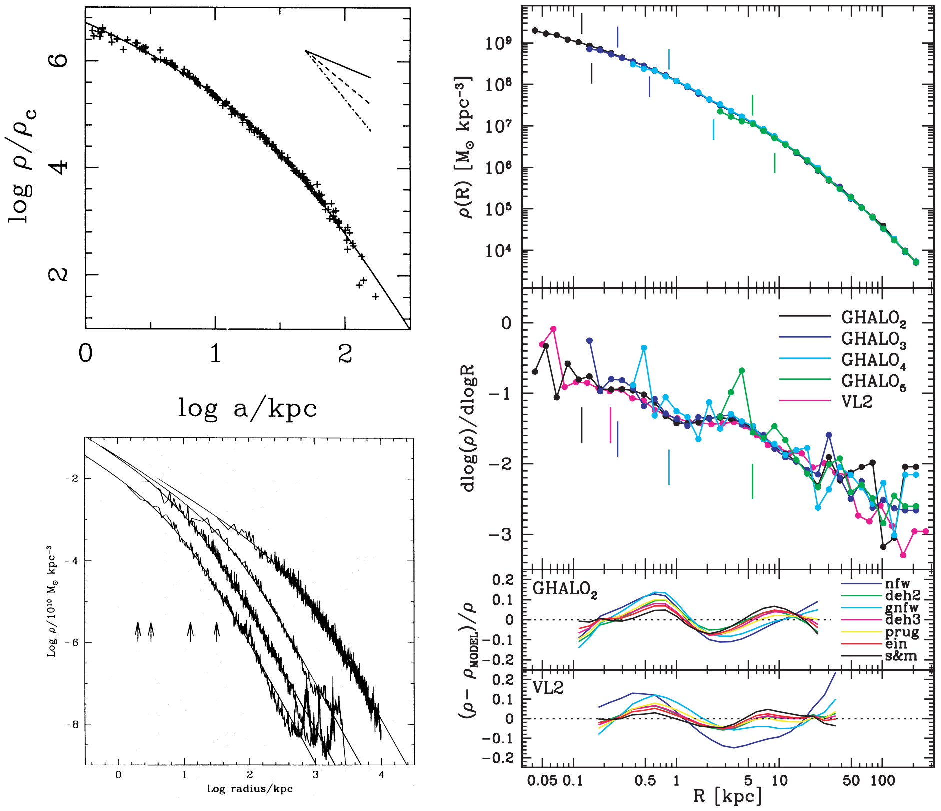


Figure 10.3: **Top left:** The dark matter halo density profile found numerically by Dubinski & Carlberg 1991. The solid, dashed and dotted lines show power law exponents of $-1, -2$ and -3, respectively. The solid line shows a fit using equation 10.1 with $\alpha = 1$; $\beta = 4$ (a Hernquist profile). The densities are given relative to the critical density of the Universe ρ_c. **Bottom left:** The same found numerically by Navarro, Frenk & White 1996 but for halos spanning four orders of magnitude in mass. They favour a fit (solid line) with $\beta = 3$. The same functional form gives a good representation (at the 10% level) to all of their halos. **Right:** The density profile found recently in the multi-billion particle 'G-halo' simulation from Stadel et al. 2009. Notice that the logarithmic slope has no clear asymptotes. None of the simple analytic forms proposed so far (bottom panel, coloured lines) give an excellent fit to the data.

orbits. These assumptions are rather crude, but illustrate the key principles. Now, let's consider the inflow and outflow phases separately.

10.2.2.1 Inflow

10.2.2.1.1 Adiabatic (slow) inflow One extreme is that the baryons flow in *adiabatically* – that is slowly with respect to the local dynamical time. In this case, dark matter particle orbits will conserve their adiabatically invariant *actions* (see Appendix G). Assuming point masses means – as for any spherically symmetric matter distribution – the *angular momentum* is an action. Now, let us imagine that the dark matter halo is constructed entirely of particles moving on circular orbits. In this case, we may write for the specific angular momentum j of a dark matter particle:

$$j^2 = GM_ir_i = GM_tr_t \tag{10.2}$$

where r_i is the initial circular orbit radius, r_t is the final orbit radius, and $M_t = M_i + M_b$ is the sum of the dark matter mass initially enclosed within r_i, and M_b is the mass in baryons adiabatically added. Rearranging gives us the final radius:

$$r_t = \frac{M_i r_i}{M_i + M_b} \tag{10.3}$$

Thus, the radii do indeed *contract* in response to the addition of baryonic matter. The effect is significant if $M_b \gtrsim M_i$ – i.e. the mass of baryons is similar to or greater than the enclosed dark matter mass within r_i.

Following early work by Young 1980 and Blumenthal *et al.* 1986, it was thought until the mid-90's that this inflow phase is the dominant effect that baryons can have on a dark matter halo. Thus, the thinking went, we should expect once baryons are added that the dark halo is more dense than the simple predictions from dark matter only simulations. As we shall see, however, both the inflow phase and a subsequent outflow phase can lead to *expansion* of the dark matter halo. As a result, the situation is not so clear. We must get the details of galaxy formation right if we wish to understand the distribution of dark matter inside galaxies.

10.2.2.1.2 Lumpy inflow The other extreme from adiabatic inflow is to have the baryons flow in in discrete lumps. These can lose energy and angular momentum to the dark halo via *dynamical friction* (see Appendix H), causing the halo to *expand*. The precise details of the dynamical friction are not actually important. If the processes is slow compared to the dynamical time (it is much slower), then the situation is adiabatic, as above, and a lump initially on a circular orbit will remain on a circular orbit. Let us imagine then that a baryonic lump of mass M_p falls to the centre, moving from one circular orbit to the next. It has initial specific orbital energy:

$$E_i = \frac{1}{2}v_i^2 - \frac{GM_i}{r_i} = -\frac{1}{2}\frac{GM_i}{r_i} \tag{10.4}$$

In moving from some radius r_i to a radius r_t, it looses specific energy:

$$\Delta E = -\frac{1}{2}GM_i\left(\frac{1}{r_t} - \frac{1}{r_i}\right) \tag{10.5}$$

And, thus the halo must absorb an energy $\Delta E M_p$ as the lump sinks. The process will be significant if this energy is comparable to the binding energy of the halo at r_t. Assuming virial equilibrium, this is:

$$E_h(r_t) = T + V = -\frac{V}{2} + V = \frac{V}{2} \sim -\frac{GM_i^2}{2r_t} \tag{10.6}$$

where we use M_i here since we have assumed a point mass halo model. Thus, the ratio of the energy lost by the baryonic lump to the halo binding energy is:

$$f_E \sim \frac{M_p}{M_i}\left(1 - \frac{r_t}{r_i}\right) \tag{10.7}$$

In the limit the lump doesn't move $r_i = r_t$ and there is no effect. In the limit the lump falls all the way to the centre, we release the maximum binding energy that is proportional to M_p/M_i. Thus, the effect is only significant if the mass in baryonic lumps is comparable to the enclosed dark matter mass.

Note that we can divide our baryons up into many discrete lumps each of which heat the dark matter halo a little: $M_p = \sum_n M_{p,n}$ and so each lump does not need to have an enormous mass. However, we cannot do this indefinitely, because it takes time for these lumps to fall to the centre via dynamical friction. This infall time goes as (see Appendix H):

$$t_{\rm fric} = \frac{2.64 \times 10^{11}}{\ln\Lambda}\left(\frac{r_i}{2\mathrm{kpc}}\right)^2\left(\frac{v_c}{250\mathrm{km/s}}\right)\left(\frac{10^6\mathrm{M}_\odot}{M_p}\right) \tag{10.8}$$

where $\ln\Lambda$ is the Coulomb logarithm that we encountered already in §1 and v_c is the circular velocity of the dark matter halo. Notice that equation 10.8 is inversely propotional to the lump mass M_p. For $r_i \sim 2\,$kpc, $v_c \sim 250\,$km/s and $\ln\Lambda \sim 10$, we require $M_p \sim 2 \times 10^6\,\mathrm{M}_\odot$ to fall to the centre within a Hubble time. Thus, we require rather massive baryonic lumps infalling very near to the centre of the dark matter halo for this mechanism to be effective.

Our simple calculation above matches well (qualitatively) with results from numerical simulations (e.g. El-Zant *et al.* 2001; Goerdt *et al.* 2010; Cole *et al.* 2011). Note that *any* angular momentum or energy transfer mechanism between baryons and the halo will produce a similar effect. A galactic bar, for example, can behave similarly (e.g. Binney and Evans 2001).

The main problem with both of the above inflow mechanisms is that, for a significant effect, they require a similar mass in baryons to flow in as the dark matter enclosed. This leaves us with extremely baryon dominated galaxies. Many such galaxies are observed in the Universe, but they are typically not interesting for dark matter studies because most of their mass (in their central regions) is in visible light. Far more interesting are systems that are *deficient* in baryons, like the smallest galaxies in the Universe: dwarf galaxies. Since these contain so few baryons, it is much easier to 'see' their dark matter and indeed some of these tiny galaxies even have mass to light ratios upwards of several hundred. For a long time, such dark matter dominated dwarfs were considered to be extremely simple systems; easy to model due to the insignificance of their baryons. However, their extremely low baryon content may suggest that they have in fact *lost* a significant fraction of their baryonic mass. We discuss this next.

10.2.2.2 Impulsive outflow

For the smallest galaxies, baryons not only flow in; they can also flow out. The energy from a supernova explosion is about $E_{SN} \sim 10^{44}$ J (e.g. Phillips 1999). The binding energy of a galaxy halo (from the virial theorem as above) is $E_b \sim -\frac{1}{2}\frac{GM_i^2}{r_i}$ and using $\sigma^2 \sim GM_i/r_i$ with $\sigma \sim 10\,\mathrm{km/s}$ and $r_i \sim 300\,\mathrm{pc}$ (for dwarf galaxies), we have $E_b \sim -\sigma^4 r_i/G \sim -10^{45}$ J. Thus a single supernovae releases $\sim 10\%$ of the binding energy of a dwarf galaxy! If a significant fraction of this energy heats the gas in the dwarf's interstellar medium, then the gas will become unbound.

If the gas is unbound adiabatically, then the effect is simply the reverse of adiabatic inflow:

$$r_f = \frac{M_i + M_b}{M_i} r_t \tag{10.9}$$

where r_f is the final radius of a dark matter particle orbit after a slow baryonic outflow. Assuming an adiabatic inflow, we then have trivially that $r_f = r_i$ and there is no net effect.

Suppose instead, however, that we have an *impulsive* outflow – i.e. all of the baryons are removed instantaneously. In this case, the angular momentum is no longer conserved. Instead, the kinetic energy will be instantaneously conserved (since forces have not yet had time to change it). The instantaneous specific energy after outflow of a dark matter particle is then:

$$E_f = \frac{1}{2}v_t^2 - \frac{G(M_t - M_b)}{r_t} \tag{10.10}$$

where v_t is the velocity of the dark matter particle after the inflow phase, but before outflow, and M_t is similarly the mass after inflow but before outflow. Using the fact that the particle is initially on a circular orbit, we may substitute for $v_t^2 = GM_t/r_t$ to give:

$$E_f = -\frac{1}{2}\frac{G(M_t - 2M_b)}{r_t} \tag{10.11}$$

And we see that the particle becomes *unbound* if $M_b \sim 0.5 M_t$ (this recovers the usual 'Hills' result; Hills 1980). Thus, we can expect this impulsive outflow to produce an irreversible halo expansion (c.f. Navarro *et al.* 1996a; Read and Gilmore 2005).

There is observational evidence for such outflows, both in the form of *observed* galactic winds from star forming dwarf galaxies, and in the observed extreme deficiency of baryons in many nearby dwarfs today (c.f. references and discussion in Read and Gilmore 2005 and Pontzen and Governato 2014).

10.2.2.3 Adiabatic inflow and impulsive outflow

We can now sew together the inflow and outflow phases. Assuming adiabatic inflow (the extreme case) and impulsive outflow (also extreme), we have for the final energy:

$$E_f = -\frac{1}{2}\frac{G(M_i - M_b)(M_i + M_b)}{M_i r_i} \tag{10.12}$$

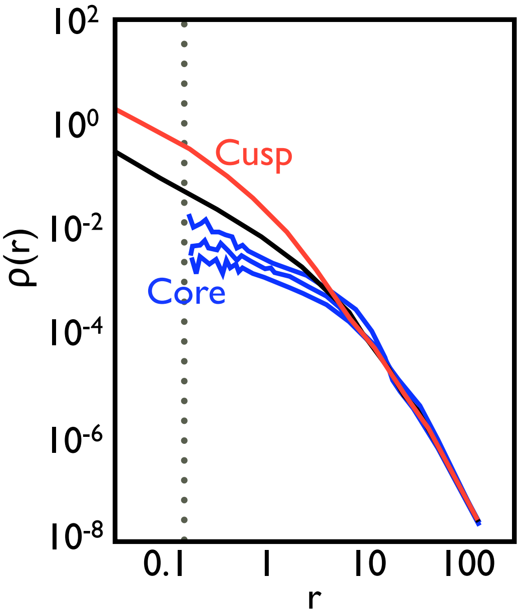


Figure 10.4: The effect of *repeated* baryonic inflow/outflow on a dark matter cusp. The initial dark matter distribution is shown in black. After the first phase of baryon inflow, the dark matter adiabatically contracts (red line). If the baryons are then impulsively removed, the dark matter expands again to the highest of the blue lines. This is remarkably similar to the initial conditions. However, if this inflow/outflow processes is repeated over several phases of bursty star formation, then the cusp is gradually transformed into a core (blue lines). The dotted grey vertical line marks the resolution limit of the simulation. The initial scale radius of the dark matter halo was 10 in simulation units, with a mass of 10. The baryonic material was assumed to have 10% of the mass of the dark matter halo and to collapse by a factor 10 in each star formation event.

and thus the ratio of initial to final energy is given by (using $E_i = -\frac{1}{2}GM_i/r_i$):

$$\frac{E_i}{E_f} = \frac{M_i^2}{(M_i - M_b)(M_i + M_b)} \tag{10.13}$$

The above equation, though crudely derived, contains the key insight: adiabatic inflow followed by impulsive outflow will produce a net *heating* effect on the underlying dark matter halo. For $M_b \sim M_i$ the effect is strong enough to actually unbind the dark matter halo.

Dwarf spheroidal galaxies orbiting the Milky Way have a mass to light ratio typically larger than ~ 100 (e.g. Mateo 1998). Let us assume that they once contained the universal baryon fraction of $f_b \sim 0.16$ (§9). They have a current mass of $\gtrsim 10^8\,\mathrm{M}_\odot$ with $\sim 10^7\mathrm{M}_\odot$ within $\sim 300\,\mathrm{pc}$. Thus, they would have had initially $M_b \sim M_i$ within 300 pc, while they currently have $M_b \sim 10^{-2}M_i$. Thus it is not unreasonable to imagine that exactly the above process occurred for these little galaxies: that they accreted a significant amount of gas, turned a small fraction of this into stars, expelled the rest and caused the central part of their dark matter halo to expand (Read and Gilmore 2005).

10.2.2.4 Numerical results

We now present the results of numerical calculations for the above processes. This takes us beyond the simple point mass approximation and assumed circular orbits that we have used so far. Navarro *et al.* 1996a were the first to consider the idea that adiabatic inflow followed by impulsive outflow could cause a dark matter halo to expand. However, they required collapse factors for the baryons of ~ 100 in order to see any strong effect. Gnedin and Zhao 2002 refined this earlier work, pointing out that angular momentum sets a barrier to collapse. The expected angular momentum gained through tidal torques in our cosmological model predicts a mean collapse factor of ~ 10. Thus, it appeared

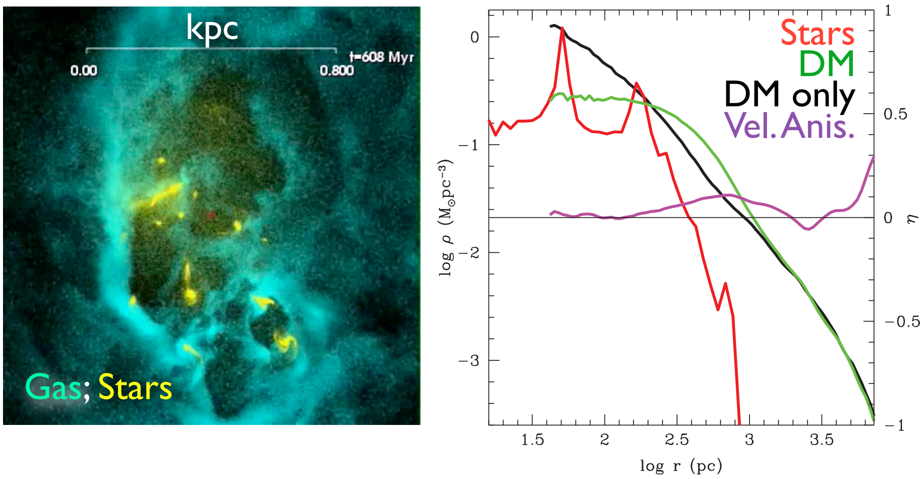


Figure 10.5: Cusp/core transformations are now seen in cosmological dwarf galaxy formation simulations. **Left:** The formation of a dwarf galaxy can be very violent with strong gas inflows/outflows and clumpy highly mobile star clusters. **Right** The dark matter profile at the end of the simulation (green), the stellar profile (red), and the velocity anisotropy of the dark matter (purple). If the same simulation is run without baryons the result is the black line that recovers the usual split-power law dark matter 'universal' profile. (Results taken from Mashchenko et al. 2008.)

that baryons could not effect dark matter halos in any significant way. However, there is a key flaw in this argument. Read and Gilmore 2005 showed that if inflow/outflow is *repeated* then we can avoid the problem of the angular momentum barrier. A dark matter halo can be gradually heated over many star formation events gradually transforming a dark matter cusp into a core, without requiring enormous collapse factors (see Figure 10.4). This work demonstrated that it is possible for baryons to alter the dark matter distribution within dwarf galaxies. For this reason, we must simulate the galaxy formation process in detail before we can be confident of the expected small scale dark matter distribution in our current cosmology.

Modern galaxy formation simulations now universally find such cusp/core transformations, if they reach a resolution where the multiphase interstellar medium is resolved (see Figure 10.5 taken from Mashchenko *et al.* 2008; and Governato *et al.* 2010; Pontzen and Governato 2012; Teyssier *et al.* 2013; Oñorbe *et al.* 2015; Read *et al.* 2016a). The mechanism at work in these simulations is potential fluctuations caused by repeated inflow/outflow events (Pontzen and Governato 2012; Pontzen *et al.* 2015). The very latest simulations have past a key resolution threshold where the effects of *individual* supernova explosions can be modelled. This is a milestone because it removes past sensitivity to the choice of 'sub-grid' numerical parameters (e.g. Oñorbe *et al.* 2015; Read *et al.* 2016a). These simulations find that dark matter is heated such that – given sufficient star formation – the central dark matter cusp is transformed to a constant density core of size $\sim R_{1/2}$, the projected half stellar mass radius. This is, then, a key prediction that we will confront with observations in the next lecture. Most importantly, it predicts that we should find 'pristine' dark matter halos either at radii $R > R_{1/2}$, or inside galaxies with truncated star formation.

10.3 A critique of the cosmological 'local Newtonian' approximation

We have derived the 'full' equations of motion – equations 9.5 and 9.6 – assuming that we can apply Newtonian gravity locally within an expanding spacetime. In fact, these equations follow from a proper linear perturbation theory of the FLRW metric in GR. Furthermore, the approximation appears to be valid also at second order (Noh and Hwang 2006). Higher order relativistic corrections appear at third order and ought to be small. That said, there is an inherent danger in our assumption that the equations of motion can be derived from perturbations of the FLRW metric. We might worry that small corrections add up coherently across the whole Universe on some intermediate scale, invalidating the perturbation approach. Such effects are called *back-reaction* terms and it has been claimed that this could conceivably explain away the accelerated expansion without the need for dark energy (e.g. Buchert 2011). However, it is now widely accepted that such effects must be small (e.g. Green and Wald 2011).

Lecture 11

The observed distribution of dark matter in the Universe

In this lecture confront the predicted dark matter distribution in the Universe with the observed distribution. We discuss the implications for the nature of dark matter.

11.1 Large scale structure

The first test we can make is to compare the observed and predicted large scale distribution of matter in the Universe. This has the advantage that baryonic effects are likely to be small, and so the predicted distribution is more robust (c.f. §10).

One key probe of the large scale distribution of matter comes from *damped Lyman-α* systems. The basic idea is to find bright distant galaxies called *quasars*. These are incredibly bright because (we think) their light is dominated by emission from gas falling onto a supermassive black hole (e.g. Lynden-Bell 1969). Being so bright, their light can make it across very large distances in the Universe; we see such objects out to redshifts of $z \sim 6-7$. As their light travels across the Universe to us, many of the photons are absorbed by intervening neutral hydrogen by the Lyman-α $n = 2$ to $n = 1$ electron transition. Since the gas has a broad range of redshifts, the result is the Lyman-α *forest* – many absorption lines shifted by the redshift of the absorbers. These absorption lines encode information about the structure of the intervening gas, and therefore about the composition of the Universe.

Seljak *et al.* 2006a and Viel *et al.* 2008 used the above to probe the free streaming length of dark matter – i.e. by comparing whether cold or warm dark matter give a better fit to the observed distribution of absorption lines. Viel *et al.* 2008 combine data from 55 high resolution quasar spectra between redshifts $z = 2-6.4$ (from `HIREZ`) and 3035 low resolution spectra from the Sloan Digital Sky Survey (SDSS) in the range $z = 2.2 - 4.2$. The *flux power spectrum* derived from the `HIREZ` spectra is given in Figure 11.1, where the 'flux power spectrum' is defined as the power $\Delta_k = P(k,z)k^3/2\pi^2$ in hydrogen absorption at a given wavenumber k. The matter power spectrum is related to the flux power spectrum in some complex model dependent non-linear way. It makes sense to compare data and models in the 'flux power' spectrum space since it is easier to transform the model (that comes from cosmological N-body simulations) than it is to transform the data. Notice that the 2.5 keV warm dark matter model gives a poor fit to the data that gets worse towards high redshift. This owes to the suppression of small scale structure at early times in warm dark matter models.

Note that, as emphasised in §9, warm dark matter models are often parameterised by the 'particle mass' in keV. In reality, all we constrain is the free streaming length (really the power spectrum of perturbations at recombination). For naive assumptions, this can be related to a particle mass through equation 9.58. In general, the relationship between free streaming length and particle mass is model dependent and complex (see e.g. Boyarsky *et al.* 2009a). With this caveat in mind, the latest constraints for thermally produced warm dark matter is $m_{\rm WDM} > 4.09$ keV at 95% confidence (**? ?**), suggesting that dark matter is quite cold.

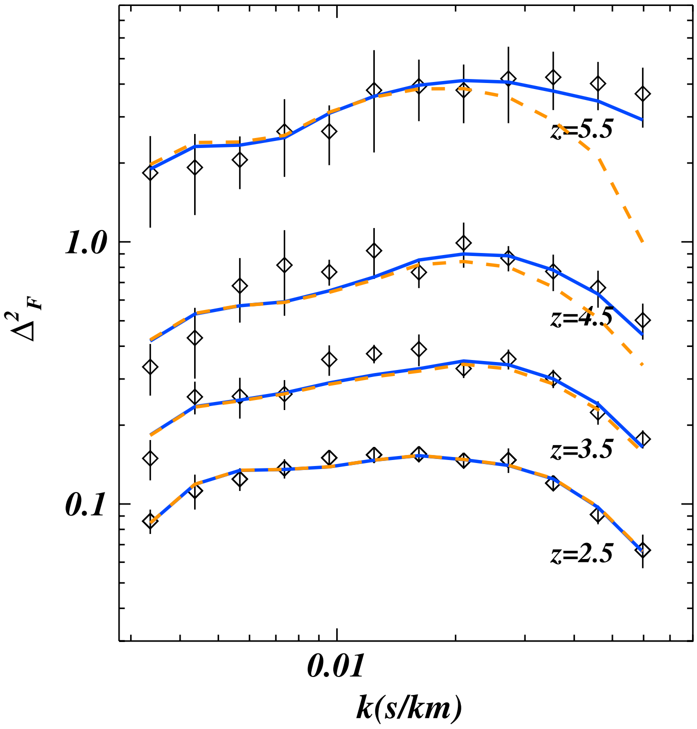


Figure 11.1: 'Flux power spectrum' derived from absorption along sight lines to 55 high resolution quasar spectra. The blue line shows a model for 8 keV warm dark matter (WDM); the red dashed line shows the same for 2.5 keV WDM.

11.2 Strong gravitational lensing

We can measure the distribution of mass in galaxy clusters and massive galaxies using strong gravitational lensing. In §4, we derived the basic lensing equations assuming a Schwarzshchild lens. These can be generalised to any projected mass distribution by realising that the Newtonian weak-field GR equations *are* the Schwarzschild metric with $GM/r \to \Phi$, where Φ is the Newtonian gravitational field (see §3). Thus, we may generalise the deflection angle:

$$\boldsymbol{\delta\alpha} = \nabla_\theta \psi \tag{11.1}$$

where the *lensing potential* ψ is given by:

$$\psi(\boldsymbol{\theta}) = \frac{D_{LS}}{D_L D_S} \frac{2}{c^2} \int \Phi(r) dz \tag{11.2}$$

and $\Phi(r = D_L\boldsymbol{\theta})$ is the Newtonian potential, and we have assumed as previously a *thin lens* (i.e. that we can treat the lens as a infinitesimal sheet of mass).

Figure 11.2 shows the lensing cluster Abel 1703 with the *PixeLens* non-parametric strong lensing mass map overlaid in white (Saha and Read 2009). The blue dots show the lensed images. Notice that the mass contours trace the underlying galaxy distribution, yet *PixeLens* used only the observed images as input. This is, then, an actual *image* of the distribution of mass (mostly dark matter) in this lens.

We can go further and deproject this two dimensional mass distribution, creating a spherically averaged 3D distribution to compare with the expected density profile from dark matter only simulations (§10). A unique deprojection requires us to assume spherical symmetry, which gives the Abel transformation:

$$\rho(r) = -\frac{1}{\pi} \int_r^\infty \frac{d\Sigma(R)}{dR} \frac{dR}{\sqrt{R^2 - r^2}}. \tag{11.3}$$

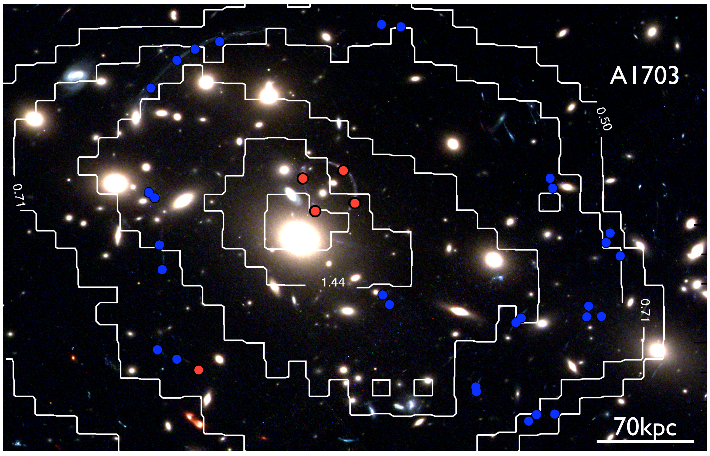


Figure 11.2: The strong lensing galaxy cluster Abel 1703. The images are marked in blue. Five are marked in red, indicating that these come from a source at a very different redshift to the others – $z = 0.88$ and compared to $z = 2.2 - 3$ for the other images (the cluster itself is at a redshift $z = 0.28$ ~ 1 Gpc away). The derived *PixeLens* 2D lensing mass map is shown in white. These iso-density contours include the mass from both the galaxies and the dark matter. Notice that they are elongated in a manner aligned with the galactic light. This is striking since the lensing analysis did not use the galaxy distribution as input data, only the images.

The result is shown in Figure 11.3. In the left panel, we perform the deprojection using all of the images except for the 5 shown in red in Figure 11.2. These five images are special because they are at a significantly different *redshift* than all of the others – $z = 0.88$ as compared with $z = 2.2 - 3$ for the other images. Notice that the mass distribution is poorly constrained. This is because even very many sources at a single redshift give us constraints only on the enclosed mass within the Einstein radius, but not the mass distribution. To obtain the mass distribution we must sample different Einstein radii, which requires multiple sources at multiple redshifts (or other constraints - e.g. time delays or kinematic information). The right panel shows the same result but now including the 'quint' images. Now the mass profile is rather well constrained. The derived distribution over scales $\sim 10 - 200$ kpc is consistent with predictions from structure formation simulations that assume that dark matter is a cold collisionless fluid (i.e. that the density profile goes as r^{-1} - a 'cusp'; see §10).

The above result for the density profile in A1703 appears to hold also for other galaxy clusters with good data, suggesting that dark matter really does behave on these large scales like a cold collisionless fluid (e.g. Saha *et al.* 2006). However, the dark matter distribution in lower mass galaxy-scale strong lenses departs from the pure dark matter predictions. This is shown in Figure 11.4, taken from Bruderer *et al.* 2016. This Figure shows the surface density of 11 strong lensing galaxies with excellent data, in units of the critical surface density for strong lensing. Notice that in all cases, the surface density profile (green) is either in excellent agreement with an NFW profile (dashed line), or steeper. Where it agrees well, the lenses are more massive, with outermost images at radii $\gtrsim 10$ kpc. Where steeper, the lenses are less massive, with images at radii $\lesssim 10$ kpc. This diversity of dark matter profiles is strong evidence for particle dark matter that contracts (on these mass scales) in response to the build-up of stars.

11.3 Weak lensing and 'self-interacting' dark matter

Another key probe of the nature of dark matter is weak gravitational lensing. This is like strong lensing, but the images are only distorted rather than split into multiple copies on the sky. In this case, we can no longer determine if an individual galaxy is being lensed (is it elliptical distorted, or intrinsically distorted?). But, we can determine statistically if a collection of galaxies in a given patch of the sky are all distorted in a similar manner. In this way, we can build up a map of the distortion due to a lens on the sky. This map is much lower resolution than that we obtain from strong lensing,

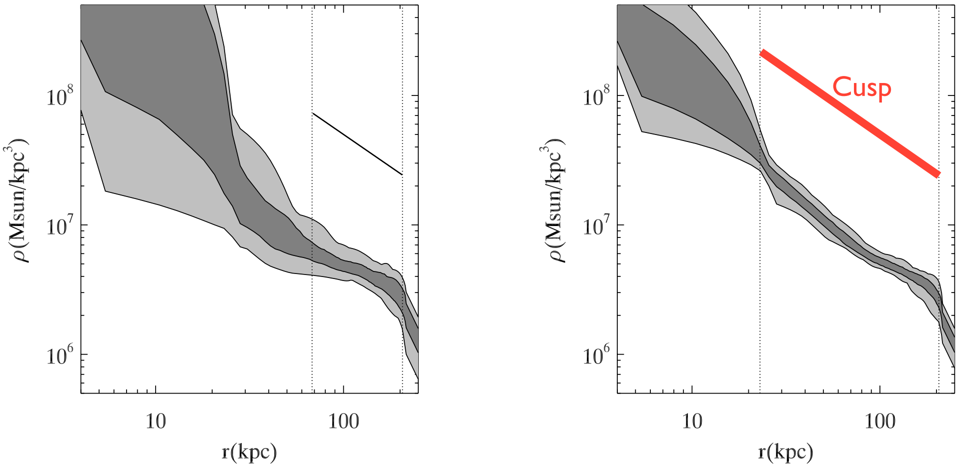


Figure 11.3: The deprojected density as a function of radius for the lensing galaxy cluster Abel 1703. **Left:** The deprojection is performed *without* the five images that are at a different redshift ($z = 0.88$) to the others ($z = 2.2 - 3$). **Right:** As for the left panel, but including the 'quint' – the five images at $z = 0.88$. Notice that, as derived analytically in §**??**, we require sources with a wide redshift separation in order to constrain the mass *distribution* in the lens. The derived distribution over scales $\sim 10 - 200$ kpc is consistent with predictions from structure formation simulations that assume that dark matter is a cold collisionless fluid (i.e. that the density profile goes as r^{-1} - a 'cusp'; see §10).

but covers a much larger area. This opens up the possibility of imaging the mass distribution in galaxy cluster mergers, like the famous 'bullet cluster'.

In Figure 11.5, I show recent work from Harvey *et al.* 2015 who have performed a weak lensing analysis of 30 merging cluster systems with over 70 mergers in total (many systems are undergoing multiple mergers). They used these data to place new limits on the the self-interaction strength of dark matter. The idea is that if dark matter is collisional, then it will experience additional drag forces as compared to the galaxies orbiting within galaxy clusters that are an almost perfect collisionless fluid. Thus, in a merger the visible part of galaxy clusters will pass through one another like ghosts, but the dark matter – like the hot X-ray emitting gas – might get left behind. This would lead to a measurable offset between light and dark in cluster mergers. Harvey *et al.* 2015 detect no such offset in their large sample, using this to place a new limit on the dark matter self-interaction cross section of $\sigma_{\rm DM} < 0.47\,{\rm cm}^2/{\rm g}$ at 95% confidence.

11.4 Near-field cosmology

In near-field cosmology, we study the tiniest galaxies in the Universe that live in our cosmic 'back yard'. These dwarf galaxies are too small to produce any significant strong lensing signal. Instead, we must probe their dark matter distribution using *kinematic* tracers – much as Zwicky derived the mass of the Coma cluster over 70 years ago. In §2, we used the kinematics of stars and gas in galaxies and galaxy clusters as a probe of dark matter. There we were interested only in demonstrating that dark matter exists – at least as a gravitational phenomenon. Here, we want to go further and use the observed kinematics to map out the *distribution* of dark matter inside galaxies. We will focus on galaxies that have enormous mass to light ratio: nearby gas rich 'dwarf irregular' galaxies and gas poor 'dwarf spheroidal' galaxies. These have the advantage that their gravitational potential is almost entirely due to dark matter.

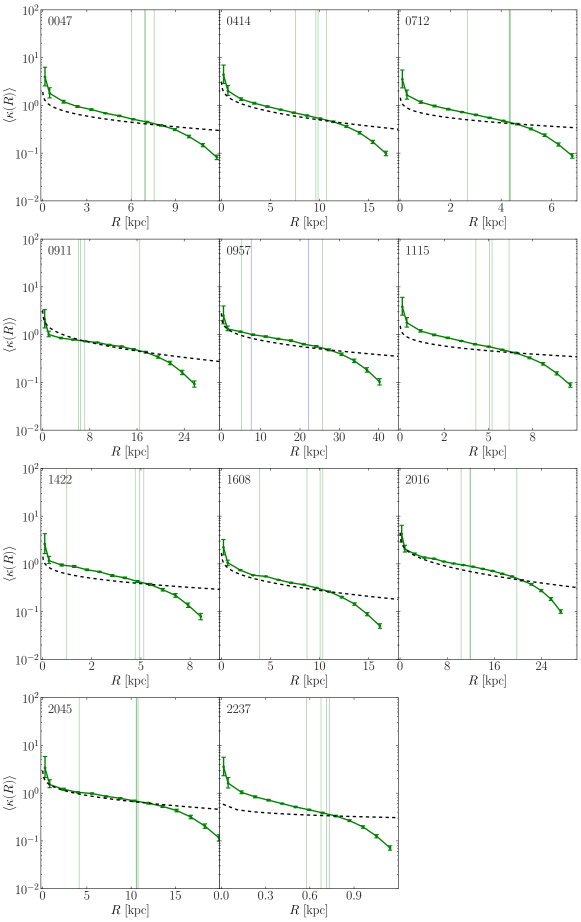


Figure 11.4: Projected surface density for 11 strong lensing galaxies with excellent data, in units of the critical surface density for strong lensing (taken from Bruderer et al. 2016). Notice that in all cases, the surface density profile (green) is either in excellent agreement with an NFW profile (dashed line), or steeper. Where it agrees well, the lenses are more massive, with outermost images at radii $\gtrsim 10$ kpc. Where steeper, the lenses are less massive, with images at radii $\lesssim 10$ kpc. This diversity of dark matter profiles is strong evidence for particle dark matter that contracts (on these mass scales) in response to the build-up of stars.

11.4.1 The dark matter distribution in gas rich dwarfs

For a rotating gas disc, we expect the gas to move on circular orbits since this is the lowest energy state of the system (c.f. §2). With this assumption, and assuming that the potential is spherical, mass modelling is especially simple. We find, from the balance of centripetal and gravitational forces, that:

$$v_c^2 = \frac{GM(r)}{r} \tag{11.4}$$

and thus the gaseous rotation curve directly gives us the mass distribution[1].

In Figure 11.6 I show the latest data and models for the tiny 'WLM' dwarf irregular galaxy (Read *et al.* 2016c). In the left panel, I attempt to fit an NFW dark matter profile (grey contours) to the

[1] Things are more complex if the mass distribution is not spherical. The rotation curve (by symmetry) gives us only dynamical information in the plane. An infinitely flattened mass distribution could produce the same dynamical effect as a spherical one if only the rotation curve is considered (see problem sheets). There has also been significant debate in the literature on observational problems in rotation curve modelling: accounting for the finite resolution of observational instruments ('beam-smearing'); non-circular gas motions; and non-spherical potentials. Recent work has demonstrated, however, that modern mass modelling techniques can recover the correct $M(r)$ even in the face of all of these problems (Kuzio de Naray and Kaufmann 2011; Read *et al.* 2016c).

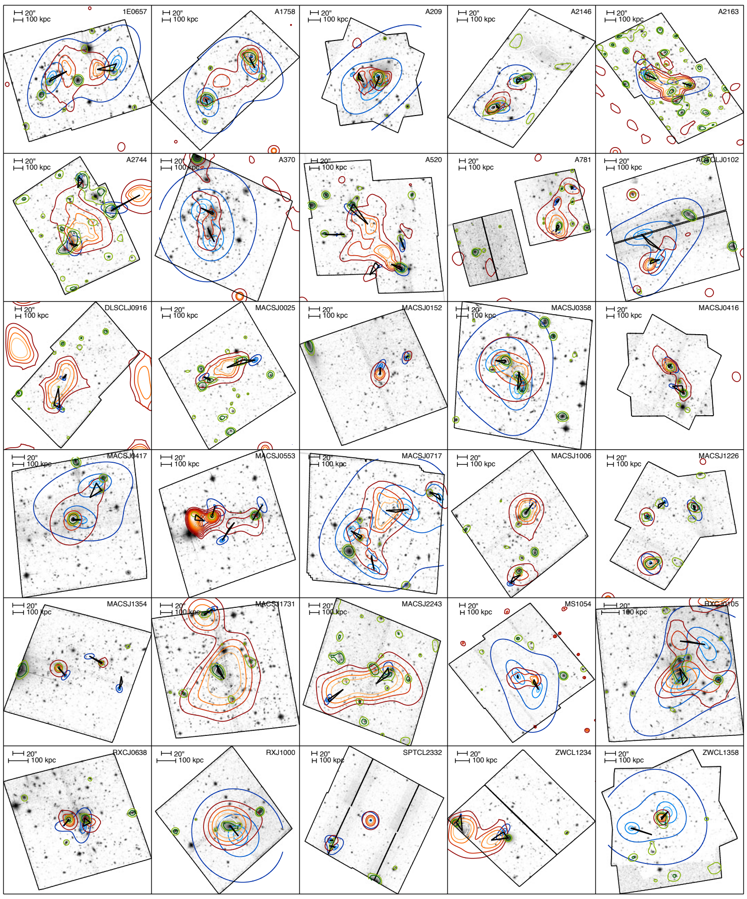


Figure 11.5: Thirty cluster merger systems studied by Harvey et al. 2015, showing the distribution of galaxies (green); gas (red); and dark matter (blue). Notice that in almost every case, there is a clear offset between the gas (red) and the galaxies and dark matter. This is because the gas experiences pressure forces and drag during the merger that the galaxies and dark matter do not feel. Harvey et al. 2015 searched for a similar offset between the galaxies and dark matter that would indicate some dark matter self-interaction force. They found no statistically significant offset, concluding that any dark matter self interaction must have a cross section of $\sigma_{\rm DM} < 0.47\,{\rm cm}^2/{\rm g}$ at 95% confidence.

data (red). As can be seen, the fit is very poor, recovering the long-standing 'cusp-core' problem (Flores and Primack 1994; Moore 1994). This discrepancy has long been suggested as evidence for self-interacting or fluid dark matter, or other beyond-ΛCDM physics (e.g. Moore 1994). However, as discussed in §10, stellar feedback leads to bursty star formation that heats dark matter at the centre of dwarf galaxies like WLM, changing the inner dark matter distribution from a cusp to a core. The right panel of Figure 11.6 shows a similar fit, but using the 'coreNFW' profile from Read *et al.* 2016a. This is derived from simulations that model such stellar feedback at high resolution, finding that it leads (after a Hubble time of star formation) to a dark matter core inside $R \gtrsim R_{1/2}$, the projected half stellar mass radius. Since $R_{1/2}$ is known observationally for WLM, the coreNFW profile has the same two free parameters as the NFW profile. Yet it gives a remarkably good fit to the data. This is the case also for *all* known low mass dwarf irregulars studied to date (Read *et al.* 2016b). It remains to be seen if other models like self interacting dark matter can fit these data too, once 'baryonic effects' are properly accounted for.

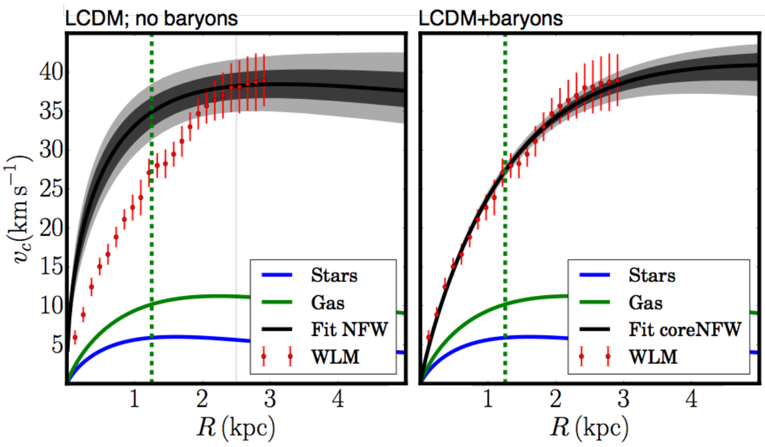


Figure 11.6: Fitting the rotation curve of the isolated dwarf irregular galaxy, WLM (Figures from Read et al. 2016b,c). The left panel shows a fit using the NFW profile. As can be seen, the fit is very poor, recovering the long-standing 'cusp-core' problem. Theshows a similar fit, but using the 'coreNFW' profile from Read et al. 2016a. This is derived from simulations that model stellar feedback at high resolution, finding that it leads (after a Hubble time of star formation) to a dark matter core inside $R \gtrsim R_{1/2}$, the projected half stellar mass radius. Now the agreement is excellent.

11.4.2 Abundance matching and new constraints on the temperature of dark matter

The excellent fits to the rotation curves of isolated dwarf irregulars opens up a new cosmological probe on the very smallest scale on which galaxies can form. Read *et al.* 2016c use mock data to show that they can recover the halo Virial masses of isolated dwarfs with good quality rotation curve data. Applying this to a large sample of nearby dwarfs, Read *et al.* 2016b show that their halo masses so-derived are very tightly correlated with their stellar masses. This means that *abundance matching* – monotonically mapping galaxies to dark matter halos of like number density – must work, if the cosmological model is correct. This test is shown in Figure 11.7. The left panel illustrates how abundance matching works. The red line shows the cumulative number of galaxies of a given stellar mass, normalised to a Mpc volume, as measured using the Sloan Digital Sky Survey (SDSS) data. The black line shows the same but for dark matter halos taken from the 'Bolshoi' ΛCDM structure formation simulation. If there is a tight monotonic relation between stellar mass and halo mass, then the galaxies marked by the blue dashed arrows should inhabit dark matter halos of the same number density, and similarly for both more and less massive galaxies. The stellar mass-halo mass relation of isolated dwarfs is shown in the right panel (purple data points) and is observed to be monotonic with little scatter. Thus, we expect abundance matching to work, if the cosmological model is correct. The expected relation from abundance matching in ΛCDM is shown by the blue solid lines and is in excellent agreement with the purple data points. What agrees less well are the equivalent warm dark matter models that fail at 68% confidence for $m_{\mathrm{WDM}} > 2\,\mathrm{keV}$ (blue dashed lines). This is not yet competitive with the Lyman-α forest constraints, but it is entirely independent. If we can find more loss mass isolated galaxies, this new method has the potential to probe beyond $m_{\mathrm{WDM}}4\,\mathrm{keV}$.

11.4.3 The dark matter distribution in dwarf spheroidal galaxies

Dwarf spheroidal galaxies are extremely interesting because they are the most dark matter dominated systems in the Universe. They also have, in many cases, very truncated star formation – most likely due to ram pressure stripping on infall to the Milky Way (e.g. Gatto *et al.* 2013). This means that they are expected to retain 'pristine' dark matter cusps, a key prediction of ΛCDM (Read *et al.* 2016a). However, this prediction has proven difficult to test because these tiny galaxies are devoid of gas. We must estimate their dark matter distribution using stars alone and this presents a problem. Stars, unlike gas, can have a wide range of orbit distributions. Their orbits can cross

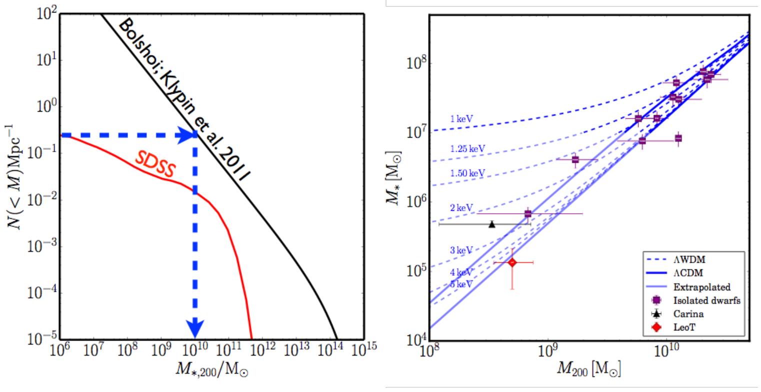


Figure 11.7: The stellar mass-halo mass relation of isolated dwarf galaxies: a new cosmological probe (taken from Read et al. 2016c). The left panel illustrates how abundance matching works. The red line shows the cumulative number of galaxies of a given stellar mass, normalised to a Mpc volume, as measured using the Sloan Digital Sky Survey (SDSS) data. The black line shows the same but for dark matter halos taken from the 'Bolshoi' ΛCDM structure formation simulation. If there is a tight monotonic relation between stellar mass and halo mass, then the galaxies marked by the blue dashed arrows should inhabit dark matter halos of the same number density, and similarly for both more and less massive galaxies. The stellar mass-halo mass relation of isolated dwarfs is shown in the right panel (purple data points) and is observed to be monotonic with little scatter. Thus, we expect abundance matching to work, if the cosmological model is correct. The expected relation from abundance matching in ΛCDM is shown by the blue solid lines and is in excellent agreement with the purple data points. What agrees less well are the equivalent warm dark matter models that fail at 68% confidence for $m_{\rm WDM} > 2\,$keV (blue dashed lines).

without any consequences, allowing them to form a fluid with potentially large velocity 'anisotropy'. This introduces severe model degeneracies when only one component of the velocity is available. (We can only measure the velocity along the line of sight using the Doppler shift of stellar spectral lines; proper motions can be used to measure the other velocity components, but even with Gaia this is not yet possible for the Milky Way's companion galaxies.) I now explain the theory behind mass modelling such 'stellar systems', and how this key degeneracy comes about in detail.

A system of many particles is described by its *distribution function*, $f(\mathbf{x}, \mathbf{v}, t)$, which is the number density of particles in phase space $(\mathbf{x}, \mathbf{v})$. Remember that this is not the *normal space density*, which is given by integrating over the velocities:

$$\rho(\mathbf{x}) = \int f d^3\mathbf{v} \tag{11.5}$$

Using the chain rule, the absolute time derivative of f is given by:

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x_i}\dot{x}_i + \frac{\partial f}{\partial v_i}\dot{v}_i \tag{11.6}$$

where we use the summation convention (see C.2), as usual.

In Appendix G.3, we prove an important theorem: Liouville's theorem. This states that phase space evolves as an incompressible fluid. In its most powerful incarnation, the theorem applies to 6N dimensional phase space and is valid for any system of particles which obey Hamilton's equations. However, in the limit of a *collisionless* system, each particle trajectory is *independent* of all of the others. This means that Liouville's theorem must also apply in 6D phase space. This is much much more useful. It means that, for a collisionless fluid, every time particles leave a small patch of phase space, they are replenished by other particles flowing in, such that the phase space density is a constant. Since f *is* the phase space density, we conclude that for collisionless systems $\frac{df}{dt} = 0$.

This leads to the collisionless Boltzmann equation:

$$\frac{\partial f}{\partial t} + \frac{\partial f}{\partial x_i}\dot{x}_i + \frac{\partial f}{\partial v_i}\nabla_i \Phi(\mathbf{x}) = 0 \tag{11.7}$$

where we have used the face that the acceleration $\dot{v}_i$ is given by the gradient of the gravitational potential $\nabla_i \Phi(\mathbf{x})$.

We now have everything we need to mass model galaxies. In principle 'all' we have to do is to measure the distribution function of stars in the galaxy $f(\mathbf{x}, \mathbf{v})$. Assuming that the system is in a steady state, $\frac{\partial f}{\partial t} \sim 0$, then we can then solve equation 11.7 to derive the gravitational potential $\nabla \Phi(\mathbf{x})$. Subtracting off the observed visible light, we are left with the gravitational potential due to the dark matter. Unfortunately, this is problematic for two reasons. Firstly, we need full 6D phase space space information for the stars in a galaxy – i.e. the 3D positions and 3D velocities. As explained in §1, this is very hard to measure in practice. Secondly, even if we could measure this, 6D space is just enormous. Even with a million stars, we would sample the space with only ten stars per dimension. And we need to take *derivatives* of f in this space!

11.4.3.1 Distribution function modelling

The above problems have led to two main approaches in the literature. One is to assume a particular functional form for f motivated by some theoretical prior, but general enough not to overly bias the solution. Derivatives then follow (semi)-analytically and the data may be compared directly with f to obtain a probability that the data are consistent with the model. Even very sparse discrete data can be compared in this way which makes such methods – called *distribution function modelling* – very appealing. The downside, however, is that we are forced to specify a form for f that could be wrong. If our guess for f does not bracket the true solution, then we will never obtain the correct answer for Φ (for an example of where this can be problematic see e.g. Garbari *et al.* 2011).

11.4.3.2 Jeans modelling

A second approach is to take instead *moments* of the distribution function:

- **Zeroth moment (spatial density):**

$$\nu(\mathbf{x}) = \int f(\mathbf{x}, \mathbf{v}) d^3\mathbf{v} \tag{11.8}$$

- **First moments (mean velocity):**

$$\overline{v}_i(\mathbf{x}) = \frac{1}{\nu(\mathbf{x})} \int \mathbf{v}_i f(\mathbf{x}, \mathbf{v}) d^3\mathbf{v} \tag{11.9}$$

- **Second moments (root mean square velocities):**

$$\overline{v_i v_j}(\mathbf{x}) = \frac{1}{\nu(\mathbf{x})} \int \mathbf{v}_i \mathbf{v}_j f(\mathbf{x}, \mathbf{v}) d^3\mathbf{v} \tag{11.10}$$

- ... and higher order moments.

which allows us to define the *velocity dispersion tensor*:

$$\sigma_{ij} = \overline{v_i v_j} - \overline{v}_i \overline{v}_j \tag{11.11}$$

This allows us to 'integrate out' some of the dimensions in the problem. Since galaxies are often roughly spherical, spherical polar coordinates are a natural choice. In this case, the steady state collisionless Boltzmann equation becomes:

$$\dot{r}\frac{\partial f}{\partial r} + \dot{\theta}\frac{\partial f}{\partial \theta} + \dot{\phi}\frac{\partial f}{\partial \phi} + \dot{v}_r \frac{\partial f}{\partial v_r} + \dot{v}_\theta \frac{\partial f}{\partial v_\theta} + \dot{v}_\phi \frac{\partial f}{\partial v_\phi} = 0 \tag{11.12}$$

where $v_r = \dot{r}$ is the velocity along r, $v_\theta = \dot{\theta} r$, and $v_\phi = \dot{\phi} r \sin\theta$.

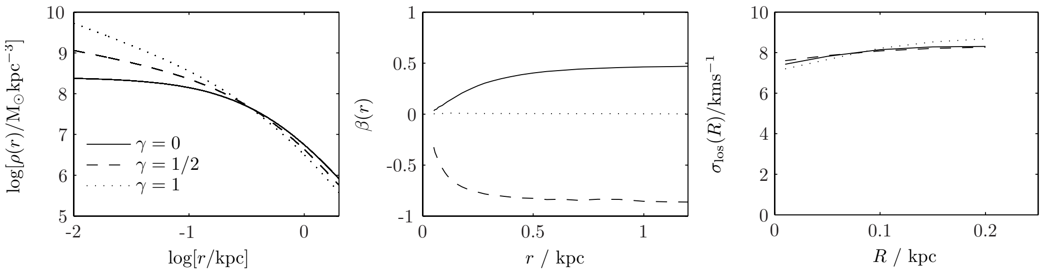


Figure 11.8: The velocity anisotropy-mass degeneracy (Figures courtesy of Mark Wilkinson). The left panel shows three different density profiles of interest: cuspy (dotted), cored (solid) and something in-between (dashed). Three different anisotropy profiles $\beta(r)$ are chosen in each case (middle panel) such that the projected velocity dispersion (that we can measure) is almost identical in each case (right panel).

Now, we can multiply through by powers of each of the velocity components v_r, v_θ, v_ϕ and integrate over velocity to obtain *moment equations* called the Jeans equations (Binney and Tremaine 2008). Assuming spherical symmetry, the 'radial' second order moment equation is given by:

$$\frac{1}{\nu}\frac{\partial}{\partial r}\left(\nu\sigma_{rr}^2\right) + \frac{2\left(\sigma_{rr}^2 - \sigma_{tt}^2\right)}{r} = -\frac{\partial\Phi}{\partial r} = -\frac{GM(r)}{r^2} \qquad (11.13)$$

where by symmetry $\sigma_{tt} = \sigma_{\theta\theta} = \sigma_{\phi\phi}$.

Now, these moment equations have the key advantages that (i) we *do not need to specify the form of f*; instead it is constrained entirely by its moments; and (ii) we have now significantly reduced the dimensionality of the problem. The above assumption of spherical symmetry can be relaxed, of course. But there is a more fundamental problem: the hierarchy of Jeans equations is not closed. If the true distribution function f is a Gaussian, then we are fine; f is fully specified by its first and second moment. But in general, f can require an infinite number of moments to be fully specified, with an associated infinity of Jeans equations! Luckily, f is typically quite close to Gaussian and so the lowest order Jeans equations usually suffice.

There is one final point worth noting. Let us define a *velocity anisotropy parameter*:

$$\beta(r) = 1 - \frac{\sigma_{tt}}{\sigma_{rr}} \qquad (11.14)$$

which is a measure of how much stars are moving tangentially (σ_{tt}) versus how much they move radially (σ_{rr}).

Now the spherical, radial, Jeans equation (equation 11.13) becomes:

$$\frac{1}{\nu}\frac{\partial}{\partial r}\left(\nu\sigma_{rr}^2\right) + \frac{2\sigma_{rr}^2\beta(r)}{r} = -\frac{GM(r)}{r^2} \qquad (11.15)$$

and we derive an important result. Typically, we measure only the velocity of stars along the *line of sight* which is some projection of σ_{rr}. Let us suppose we could measure σ_{rr} *perfectly*. In this case, we would still be unable to determine $M(r)$ since we do not know $\beta(r)$. This is a fundamental degeneracy in mass modelling that is illustrated in Figure 11.8.

The good news is that we do not perfectly measure σ_{rr}, we measure its projection along the line of sight. Thus, we also must measure some projection of σ_{tt}. The situation is messy, but given enough stellar tracers – and using higher order moments – we can hope to measure both σ_{tt} and σ_{rr} (see e.g. Łokas 2009). This is still under the assumption of spherical symmetry, however.

The anisotropy mass degeneracy has meant that until very recently, we could not reliably measure the mass distribution within dwarf galaxies. This is illustrated for the Fornax dwarf spheroidal in Figure 11.9 (taken from Walker and Peñarrubia 2011). The top left panel shows the projected velocity dispersion data for this dwarf. Several models are overlaid (lines) demonstrating that a wide variety of models – including both 'cuspy' and 'cored' models fit the data.

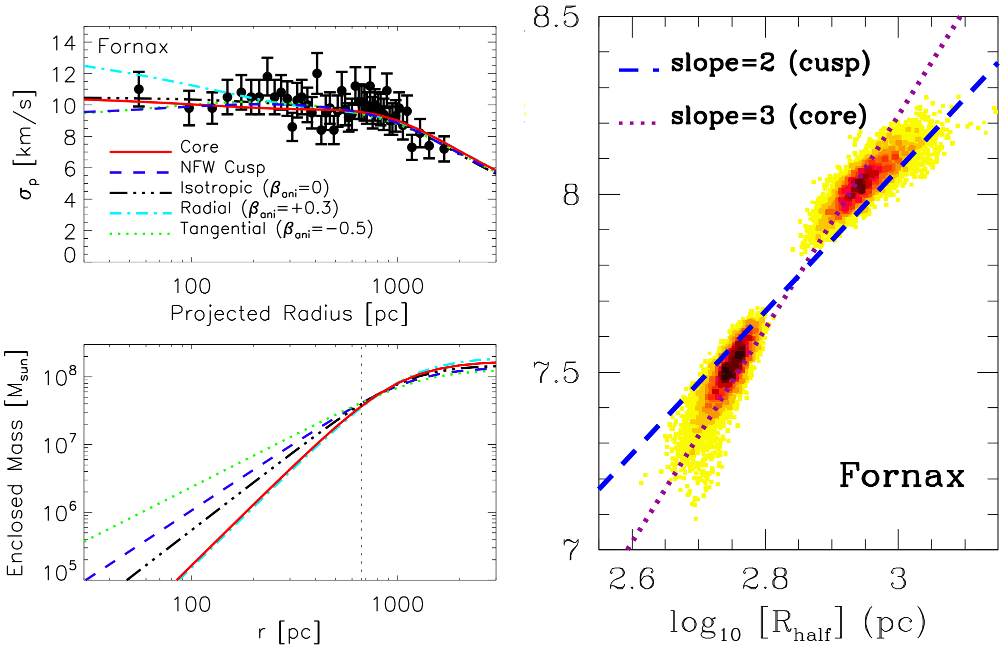


Figure 11.9: The observed mass distribution in the Fornax dwarf spheroidal galaxy (taken from Walker & Penarrubia 2011). **Top left:** The projected velocity dispersion as a function of radius for Fornax (data points). Several models are overlaid (lines) demonstrating that a wide variety of models – including both 'cuspy' and 'cored' models fit the data. **Bottom left:** The enclosed mass as a function of radius for the same models shown in the top left panel. Notice that all models cross at a critical point shown by the dotted black line: the half light radius. **Right:** Splitting the stars in Fornax into a metal rich and a metal poor population improves the constraints on the mass profile. Each population gives a constraint on the enclosed mass at a different radius shown by the two dense clouds. The data appear to favour a cored model (magenta dotted line) over a cusped model (blue dashed line).

However, the are ways to break this degeneracy. The bottom left panel shows the enclosed mass as a function of radius for the same models shown in the top left panel. Notice that all models cross at a critical point shown by the dotted black line: the half light radius. Battaglia *et al.* 2008 pointed out that we can split the stars in dwarf galaxies into a metal rich and a metal poor population that have *different scale lengths* (see also Walker and Peñarrubia 2011; Amorisco and Evans 2011). Thus, we can obtain a reliable mass estimate at two different radii: one for each population. This breaks the mass anisotropy degeneracy, giving a constraint on the mass profile. This is shown for Fornax in the right panel of Figure 11.9 (taken from Walker and Peñarrubia 2011). Each population gives a constraint on the enclosed mass at a different radius shown by the two dense clouds. The data appear to favour a cored model (magenta dotted line) over a cusped model (blue dashed line), an argument further strengthened by *indirect* but compelling evidence for a dark matter core from Fornax's Globular Cluster distribution (Goerdt *et al.* 2006; Cole *et al.* 2012).

However, the case for cores in the other Milky Way dwarf spheroidals is less compelling. Battaglia *et al.* 2008 and Walker and Peñarrubia 2011 find – using the same analysis technique as for Fornax – that the Sculptor dwarf spheroidal also favours a core over a cusp, though its core is much less statistically significant. All other dwarfs analysed to date are degenerate between cusps and cores because split populations as powerful as those in Fornax have not yet been found.

Work continues in ernest on the dwarfs, however, because the prize is large. Many of them, like the Draco dwarf, have formed so few stars that it should contain a near-pristine dark matter cusp, a key prediction of 'Cold Dark Matter' that has yet to be tested.

Appendix A

Common constants in astrophysics

Constant	*Value in S.I. units*
Gravitational constant	$G = 6.672(4){\times}10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$
Speed of light	$c = 2.99792458{\times}10^{8}\,\mathrm{m\,s^{-1}}$ (by definition)
Solar mass	$\mathrm{M}_\odot = 1.989(2){\times}10^{30}\,\mathrm{kg}$
Earth mass	$\mathrm{M}_\oplus = 5.976(4){\times}10^{24}\,\mathrm{kg}$
Solar bolometric luminosity	$\mathrm{L}_\odot = 3.826(8){\times}10^{26}\,\mathrm{j\,s^{-1}}$
Stefan-Boltzmann constant	$\sigma = 5.670 \times 10^{-8}\mathrm{J\,K^{-4}\,m^{-2}\,s^{-1}}$

Unit conversions	*Value in S.I. units*
Astronomical unit	$1\ \mathrm{a.u.} = 1.49597892(1){\times}10^{11}\,\mathrm{m}$
Parsec	$\mathrm{pc} = 3.08567802(2){\times}10^{16}\,\mathrm{m}$
Light year	$\mathrm{lyr} = 9.4605284{\times}10^{15}\,\mathrm{m}$
Erg	$\mathrm{erg} = 10^{-7}\,\mathrm{j}$
Minute of arc	$\mathrm{arcmin} = 2\pi/360/60\,\mathrm{rad}$
Second of arc	$\mathrm{arcsec} = 2\pi/360/60/60\,\mathrm{rad}$

Appendix B

Key results from Vector Calculus

B.1 Curvilinear coordinates

Life is easy working in Cartesian coordinates: (x, y, z). However, as we shall see time and again throughout this course, problems are often much simpler if we exploit inherent *symmetries*. It helps then to work in coordinate systems which share the same symmetry as the problem we are looking at. In practise, this means working typically in cylindrical polar coordinates: (R, ϕ, z), or spherical polar coordinates: (r, θ, ϕ). Here, we briefly summarise the mathematical machinery required to transform between general coordinate systems. For a much more detailed account see e.g. [Arfken and Weber, 2005].

Suppose we switch from Cartesian coordinates to some general coordinates: (q_1, q_2, q_3). From the chain rule we have:

$$dx = \frac{\partial x}{\partial q_1}dq_1 + \frac{\partial x}{\partial q_2}dq_2 + \frac{\partial x}{\partial q_3}dq_3 \tag{B.1}$$

and similarly for y and z. Thus the distance between two points (q_1, q_2, q_3) and $(q_1+dq_1, q_2+dq_2, q_3+dq_3)$, is given by:

$$\begin{aligned} ds^2 &= dx^2 + dy^2 + dz^2 \\ &= \sum_l \frac{\partial x_l}{\partial q_i}\frac{\partial x_l}{\partial q_j}dq_i dq_j \\ &= h_{ij} dq_i dq_j \end{aligned} \tag{B.2}$$

where we have employed the *summation convention* (repeated indices are summed over), and h_{ij} is known as the *metric tensor* - you may be familiar with this from General Relativity. In *orthogonal* coordinate systems, h_{ij} is diagonal $\Rightarrow ds^2 = h_{ii}dq_i^2 = h_i^2 dq_i^2$. This last equality is a notation usually used to avoid confusion since for $h_{ii}dq_i dq_i$ it may not be clear what is really summed over. In the above definition, we have that $h_1 = \sqrt{h_{11}}$ and similarly for the other components. Finally, note that we have used the notation x_l for the lth component of the vector $\underline{x} = (x, y, z)$.

As an example, consider spherical polar coordinates. Here we have:

$$x = r\sin\theta\cos\phi; \qquad y = r\sin\theta\sin\phi; \qquad z = r\cos\theta \tag{B.3}$$

Thus, we have:

$$\begin{aligned} h_1^2 = h_{11} &= \sum_l \frac{\partial x_l}{\partial q_1}\frac{\partial x_l}{\partial q_1} \\ &= \sin^2\theta\cos^2\phi + \sin^2\theta\sin^2\phi + \cos^2\theta \\ &= 1 \end{aligned} \tag{B.4}$$

Similarly, we find $h_2 = r$, $h_3 = r\sin\theta$.

B.2 Divergence operator

The *Divergence operator*, also called *grad*, or *nabla*, in Cartesian coordinates is given by:

$$\underline{\nabla} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \tag{B.5}$$

The above notation is commonly used, but is potentially dangerous. More formally, we should write:

$$\underline{\nabla} = \hat{\underline{x}} \frac{\partial}{\partial x} + \hat{\underline{y}} \frac{\partial}{\partial y} + \hat{\underline{z}} \frac{\partial}{\partial z} \tag{B.6}$$

where $\hat{\underline{x}}, \hat{\underline{y}}, \hat{\underline{z}}$ are unit vectors pointing along each of the Cartesian coordinate axes. In Cartesian coordinates, where each unit vector is a function only of one coordinate ($\underline{\nabla} \cdot \hat{\underline{x}} = \frac{\partial}{\partial x}$, etc.), this distinction is not so important. However, in more general orthogonal coordinates, we *must remember that nabla acts also on the unit vectors themselves.*

In a general, orthogonal, coordinate system: (q_1, q_2, q_3), $\underline{\nabla}$ is given by:

$$\underline{\nabla} = \frac{\hat{\underline{e}}_1}{h_1} \frac{\partial}{\partial q_1} + \frac{\hat{\underline{e}}_2}{h_2} \frac{\partial}{\partial q_2} + \frac{\hat{\underline{e}}_3}{h_3} \frac{\partial}{\partial q_3} \tag{B.7}$$

where $\hat{\underline{e}}_1, \hat{\underline{e}}_2, \hat{\underline{e_3}}$ are unit vectors pointing along each of the general coordinate axes. Note that we do not concern ourselves here with covariant and contravariant forms since these only come into play when we consider *non-orthogonal* coordinate systems.

B.3 Divergence & Curl

The *Divergence* in Cartesian coordinates is defined:

$$\underline{\nabla} \cdot \underline{F} = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z} \tag{B.8}$$

And in a general, orthogonal, coordinate system: (q_1, q_2, q_3), it is:

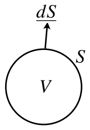
$$\underline{\nabla} \cdot \underline{F} = \frac{1}{h_1 h_2 h_3} \left(\frac{\partial}{\partial q_1}(h_2 h_3 F_1) + \frac{\partial}{\partial q_2}(h_3 h_1 F_2) + \frac{\partial}{\partial q_3}(h_1 h_2 F_3) \right) \tag{B.9}$$

Similar results may be derived for the curl, $\underline{\nabla} \times \underline{F}$, in general coordinate systems (see e.g. [Arfken and Weber, 2005]).

The Divergence and Curl may be better understood physically through the following theorems:

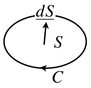
The Divergence Theorem:

$$\int_V \underline{\nabla} \cdot \underline{F} dV = \int_S \underline{F} \cdot d\underline{S} \tag{B.10}$$



Stoke's Theorem:

$$\oint_C \underline{F} \cdot d\underline{l} = \int_S (\underline{\nabla} \times \underline{F}) \cdot d\underline{S} \tag{B.11}$$



The above two theorems give us *physical insight.* Suppose that the field $\underline{F}$ represents the force per unit mass of a gravitational field: $\underline{F} = -\underline{\nabla}\Phi$. Then $\underline{\nabla} \cdot \underline{F} = 0$ tells us that there is no net force pointing in or out of a surface bounding some volume around the gravitational field, V. No force means no mass to produce that force. Not surprisingly, then, we have from Poisson's equation $\underline{\nabla} \cdot \underline{F} = \underline{\nabla}^2 \Phi = 4\pi G \rho = 0$. Similarly, $\underline{\nabla} \times \underline{F} = 0$ tells us something important about the gravitational field. It means that the integral around a closed loop of $\underline{F} \cdot d\underline{l} = 0$. But this is just the *work done* – the energy expended in moving around that closed loop. It means that the field is *conservative* and that particles moving in that field conserve energy. Since $\underline{\nabla} \times \underline{\nabla}\Phi = 0$ for any scalar field Φ [exercise], we have that gravity must be a conservative force.

Appendix C

Some useful mathematical functions

C.1 The Dirac Delta function

The *Dirac Delta* function is defined as:

$$\delta(\underline{x}) = 0; \underline{x} \neq 0 \tag{C.1}$$

$$\int f(\underline{x})\delta(\underline{x})d^3\underline{x} = f(\underline{0}) \tag{C.2}$$

C.2 Functions for use in tensor calculus

The Dirac Delta function is not to be confused with the *Kronecker delta* used in tensor calculus:

$$\delta_{ij} = \begin{cases} 1, & \text{if} \quad i = j \\ 0, & \text{if} \quad i \neq j \end{cases} \tag{C.3}$$

Another useful object in tensor calculus is the *Levi-Civita* pseudo-tensor:

$$\epsilon_{ijk} = \begin{cases} 1, & \text{if} \qquad (i,j,k) = (1,2,3), (2,3,1), \text{ or } (3,1,2) \\ -1, & \text{if} \qquad (i,j,k) = (3,2,1), (2,1,3), \text{ or } (1,3,2) \\ 0, & \text{otherwise} : \quad i = j, j = k, \text{ or } k = i \end{cases} \tag{C.4}$$

It is used to define the *cross product* of two vectors:

$$\underline{c} = \underline{a} \times \underline{b} \equiv c_i = \epsilon_{ijk} a_j b_k \tag{C.5}$$

where we have employed the *summation convention*:

$$c_i = \epsilon_{ijk} a_j b_k \equiv \sum_{j,k} \epsilon_{ijk} a_j b_k \tag{C.6}$$

You should be able to convince yourself, using the definition of ϵ_{ijk}, that the above is indeed the usual cross product (which many students like to remember it as a determinant of a 3x3 matrix).

Note that ϵ_{ijk} is a *pseudo-tensor*. The result of a cross product is not actually a vector (as is sometimes taught), but a *pseudo-vector*. Pseudo vectors transform just like normal vectors under a rotation, but not under an inversion followed by a rotation (where they gain an extra sign-flip). This is easy to see for the cross product by considering a coordinate inversion where all vectors change sign: $\underline{a} \to -\underline{a}$, $\underline{b} \to -\underline{b}$. But the pseudo-vector $\underline{c} = -\underline{a} \times -\underline{b}$ remains unchanged. Pseudo-tensors may be similarly defined. They are, in general, of limited use because they are not (unlike normal tensors) coordinate invariant.

Appendix D

The Taylor expansion

We use the Taylor expansion a lot throughout this course; for completeness, we derive it here. We may write any function as an infinite power law series:

$$f(x) = \sum_{n=0} a_n (x-a)^n; \qquad (x-a) < 1 \tag{D.1}$$

The $(x - a) < 1$ is required to ensure the series converges. If a function may be represented by a finite number of terms (for example if $f(x)$ is really a polynomial), then this criteria may be dropped. The terms, a_n, may be obtained by differentiation. Notice that:

$$f(x)' = \sum_{n=1} n a_n (x-a)^{n-1} \tag{D.2}$$

$$f(x)'' = \sum_{n=2} n(n-1) a_n (x-a)^{n-2} \tag{D.3}$$

$$\vdots$$

$$f(x)^n = n! a_n \tag{D.4}$$

We can now find the a_n by setting $x = a$:

$$f^n(a) = n! a_n \tag{D.5}$$

and we derive the Taylor series:

$$f(x) = \sum_{n=0} \frac{f^n(a)}{n!}(x-a)^n; \qquad (x-a) < 1 \tag{D.6}$$

The above refers to a Taylor series in x about a point a. This can be a source of confusion for students because, more commonly, people want to expand a Taylor series in some small quantity δx about a point x. This means that in the above formula, we must substitute: $x \to x + \delta x$ and $a \to x$. This confusing use of notation is common, unfortunately, in most math methods books. Switching variables, as above, we obtain:

$$f(x + \delta x) = \sum_{n=0} \frac{f^n(x)}{n!} \delta x^n; \qquad \delta x < 1 \tag{D.7}$$

which is the form of the Taylor expansion most commonly used in physics. It may be simply generalised to vectors of more than one variable to give:

$$f(\underline{x} + \delta \underline{x}) = \sum_n \frac{1}{n!} \left(\delta \underline{x} \cdot \underline{\nabla}\right)^n f(\underline{x}); \qquad |\delta \underline{x}| < 1 \tag{D.8}$$

see e.g. [Arfken and Weber, 2005].

Appendix E

Solving Poisson's and Laplace's equations

Two very important equations in physics are Poisson's equation:

$$\underline{\nabla}^2\Phi = 4\pi G\rho \tag{E.1}$$

and, the special case, Laplace's equation:

$$\underline{\nabla}^2\Phi = 0 \tag{E.2}$$

Here we outlined the basic strategy for solving Poisson's equation: reduce $\underline{\nabla}^2\Phi = 4\pi G\rho$ to solving $\underline{\nabla}^2\Phi = 0$ inside and outside of infinitesimal spherical shells, subject to suitable boundary conditions; then sum over these infinitesimal shells. In this appendix we work through a concrete example of this in cylindrical polar coordinates (R, ϕ, z).

So, step one is to solve Laplace's equation in cylindrical coordinates. Before solving it, we must first recall what Laplace's equation *is* in cylindrical polar coordinates. In Cartesian coordinates it is straightforward from the definition of $\underline{\nabla}$, also called *grad*[1] (see Appendix B):

$$\underline{\nabla}^2 = \underline{\nabla}\cdot\underline{\nabla} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{E.3}$$

In more general coordinate systems, we must remember to correctly transform each of the Cartesian coordinates (see e.g. Appendix B and [Arfken and Weber, 2005]). $\underline{\nabla}^2$ then looks quite different. Substituting for $\underline{F} = \underline{\nabla}\Phi$ in equation B.9 and noting that in cylindrical coordinates we have $h_R = 1$, $h_\phi = R$ and $h_z = 1$, we recover:

$$\underline{\nabla}^2\Phi = \frac{1}{R}\frac{\partial}{\partial R}\left(R\frac{\partial\Phi}{\partial R}\right) + \frac{1}{R^2}\frac{\partial^2\Phi}{\partial\phi^2} + \frac{\partial^2\Phi}{\partial z^2} = 0 \tag{E.4}$$

The key to solving Laplace's equation is the method of *separation of variables.* It is important to note that this is only possible in some special coordinate systems – notably Cartesian, cylindrical polars, spherical polars and oblate spherical coordinates. In more general coordinates things get more difficult.

Separation of variables works by writing: $\Phi(R, \phi, z) = A(R)B(\phi)C(z)$. Notice that we may now rearrange equation E.4 to give:

$$\underbrace{\frac{1}{AR}\frac{\partial}{\partial R}\left(R\frac{\partial A}{\partial R}\right) + \frac{1}{R^2 B}\frac{\partial^2 B}{\partial\phi^2}}_{f(R,\phi)} + \underbrace{\frac{1}{C}\frac{\partial^2 C}{\partial z^2}}_{g(z)} = 0 \tag{E.5}$$

[1] $\underline{\nabla}$ and $\underline{\nabla}^2$ are often referred to as *operators* because they operate on the variable which comes after them. In this case the operation is differentiation.

Notice that the left two terms are a function of R and ϕ only, while the right term is a function only of z. This is the key to the separation of variables method. Now the term on the right is a *constant* as far as the left two terms are concerned. We may write:

$$-\frac{1}{C}\frac{\partial^2 C}{\partial z^2} = \text{const.} = m^2 \tag{E.6}$$

which gives us $C(z) = C_m e^{mz}$, where m is a *complex number*.

Now we may play the same game with the left two terms. Re-arranging, we obtain:

$$\frac{R}{A}\frac{\partial}{\partial R}\left(R\frac{\partial A}{\partial R}\right) - m^2R^2 = -\frac{1}{B}\frac{\partial^2 B}{\partial \phi^2} \tag{E.7}$$

Now the left term is a function only of R, while the right is a function only of ϕ. As above, we may introduce another constant and find: $B(\phi) = B_l e^{l\phi}$ (l is also a complex number). This leaves just the R equation, which may be rearranged to give:

$$\frac{1}{R}\frac{\partial}{\partial R}\left(R\frac{\partial A}{\partial R}\right) - (m^2 + \frac{l^2}{R^2})A = 0 \tag{E.8}$$

This is Bessel's equation. Its solutions may not be obtained analytically. This is usually 'swept under the carpet' by simply labelling the functions which solve the above equation as 'Bessel functions'; these may then be calculated whenever they are required using numerical techniques (see e.g. [Press *et al.*, 1992]). The full solution to Laplace's equation may now be written as:

$$\Phi(R,\phi,z) = \sum_{m,l} a_{ml}A_{ml}(R)e^{mz}e^{l\phi} \tag{E.9}$$

where $A_{ml}(R)$ are the Bessel functions. Poisson's equation may now be solved from the above by applying boundary conditions at infinity, zero and the surface of the thin shell; and summing over all such infinitesimal shells.

For a disc galaxy, we may assume that it is symmetric in ϕ. This is, of course, just an approximation. Real galaxies show beautiful spiral arm features which are clearly not symmetric in ϕ. However, using this assumption, we have: $B(\phi) = \text{const.}$ and $l = 0$. Thus equation E.8 reduces to:

$$\frac{1}{R}\frac{\partial}{\partial R}\left(R\frac{\partial A}{\partial R}\right) - m^2 A = 0 \tag{E.10}$$

where $A \equiv J_m(R)$ are cylindrical Bessel functions of order m.

As a final postscript, note that there are a whole other independent set of solutions to equation E.10 usually denoted $J_{-m}(R)$ which is a bit confusing since this does not mean that m is negative. In special cases, we also need to consider Bessel functions of the second kind – called Neumann functions. Finally, there are a set of related functions which solve a very similar equation called modified Bessel functions. You can read about all of these and more in any good math methods textbook (e.g. [Arfken and Weber, 2005]).

Appendix F

Spherical harmonics

Spherical harmonics are an orthonormal basis function set, defined on the surface of a sphere (θ, ϕ). They are a natural choice for systems at or close to spherical symmetry. A function can in general be written as some sum over the basis set (c.f. Fourier series):

$$f(r, \theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{m=l} f_{lm}(a) Y_l^m(\theta, \phi) \tag{F.1}$$

where (r, θ, ϕ) are the familiar spherical polar coordinates, $f_{lm}(a)$ are coefficients, and $Y_l^m(\theta, \phi)$ are the spherical harmonic basis functions (see below for the first few of these). The coefficients may be derived similarly to Fourier series coefficients by multiplying through by the conjugate basis set $Y_l^{m*}(\theta, \phi)$ and integrating:

$$f_{lm} = \int_0^{\pi} \sin\theta d\theta \int_0^{2\pi} d\phi Y_l^{m*}(\theta, \phi) f(r, \theta, \phi) \tag{F.2}$$

The first few spherical harmonic terms are given below for reference. A graphical representation of these is given in Figure F.1.

$$\begin{array}{lll}
Y_0^0(\theta, \phi) = \frac{1}{\sqrt{4\pi}} & & \\
Y_1^0(\theta, \phi) = \sqrt{\frac{3}{4\pi}} \cos\theta & Y_1^{\pm 1}(\theta, \phi) = \pm\sqrt{\frac{3}{8\pi}} \sin\theta e^{\pm i\phi} & \\
Y_2^0(\theta, \phi) = \sqrt{\frac{5}{16\pi}} (3\cos^2\theta - 1) & Y_2^{\pm 1}(\theta, \phi) = \pm\sqrt{\frac{15}{8\pi}} \sin\theta \cos\theta e^{\pm i\phi} & Y_2^{\pm 2}(\theta, \phi) = \sqrt{\frac{15}{32\pi}} \sin^2\theta e^{\pm 2i\phi}
\end{array}$$

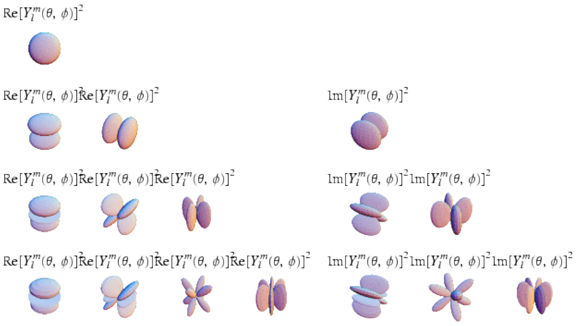

$\mathrm{Re}[Y_l^m(\theta, \phi)]^2$
$\mathrm{Re}[Y_l^m(\theta, \phi)]^2$ $\mathrm{Re}[Y_l^m(\theta, \phi)]^2$
$\mathrm{Im}[Y_l^m(\theta, \phi)]^2$
$\mathrm{Re}[Y_l^m(\theta, \phi)]^2$ $\mathrm{Re}[Y_l^m(\theta, \phi)]^2$ $\mathrm{Re}[Y_l^m(\theta, \phi)]^2$
$\mathrm{Im}[Y_l^m(\theta, \phi)]^2$ $\mathrm{Im}[Y_l^m(\theta, \phi)]^2$
$\mathrm{Re}[Y_l^m(\theta, \phi)]^2$ $\mathrm{Re}[Y_l^m(\theta, \phi)]^2$ $\mathrm{Re}[Y_l^m(\theta, \phi)]^2$ $\mathrm{Re}[Y_l^m(\theta, \phi)]^2$
$\mathrm{Im}[Y_l^m(\theta, \phi)]^2$ $\mathrm{Im}[Y_l^m(\theta, \phi)]^2$ $\mathrm{Im}[Y_l^m(\theta, \phi)]^2$

Figure F.1: The real and imaginary parts of the first few spherical harmonic basis functions. They may look familiar to you from chemistry class. The monopole term is what physical chemists would call an ‘s’ orbit, the dipole term is what is referred to as a ‘p’ orbit. The sum over many terms in the spherical harmonic series, each with different weight can reproduce any smooth function of (θ, ϕ). Hence they are referred to as orthogonal basis functions. Fourier series are another example of a set of basis functions you may have come across before.

Appendix G

Lagrangian & Hamiltonian mechanics

In this appendix we briefly review Lagrangian and Hamiltonian mechanics. It is important to state up-front that neither of these methods will do anything that you can't already do with Newtonian mechanics. In fact, they do a little bit less as you will see on the problem sheet. However, they often make hard problems in Newtonian mechanics very simple. As an example, we will use them in this appendix to solve two powerful theorems: Noether's theorem and Liouville's theorem. You will see many other examples on the problem sheet and throughout the course.

G.1 Lagrangian mechanics

In classical Newtonian mechanics, a system of gravitating particles evolves under *Newton's laws.* These are the familiar: a body continues its motion unchanged unless acted on by a force; force is the rate of change of momentum; and every action has an equal and opposite reaction[1]. Lagrangian mechanics is really just a reworking of Newton's first law. The central idea is summed up in Figure G.1. A particle starts at a position $\underline{x}_1(t_1)$ and moves to $\underline{x}_2(t_2)$. If no forces acted it would move in a straight line. However, in the presence of forces (in this case gravity), the particle's motion will be more complex. The central idea in Lagrangian mechanics is that this deviation from a straight line path will be as small as possible. Particles will move on the shortest possible path between points 1 and 2 given the *constraint* that they are acted on by forces. In more mathematical language, we may write the path length between 1 and 2 as:

$$S = \int_{\underline{x}_1, t_1}^{\underline{x}_2, t_2} L(\underline{x}, \dot{\underline{x}}, t) dt \tag{G.1}$$

This defines the *Lagrangian*, $L(\underline{x}, \dot{\underline{x}}, t)$, which now contains all of the *physics.*

Now, suppose that we know that the path, S, is the shortest given the physical constraints. This means that if we pick a path infinitesimally close to S, S', then $\delta S = S' - S = 0$. This is shown in Figure G.1. Along the path S, the particle motion is given by the function $\underline{x}(t)$; along S' it is given by $\underline{x}'(t) = \underline{x} + \delta\underline{x}$. Now, if we can solve for $\underline{x}(t)$, then we have solved for the *dynamics* of the system; we know what path the particle will take given the *boundary conditions*: $\underline{x}_1(t_1), \underline{x}_2(t_2)$. The solution is a key result from the Calculus of Variations and we will now derive it:

$$\begin{aligned}
\delta S &= S' - S \\
&= \delta \int_{\underline{x}_1, t_1}^{\underline{x}_2, t_2} L dt \\
&= \int_{\underline{x}_1, t_1}^{\underline{x}_2, t_2} \left[L([\underline{x} + \delta\underline{x}], [\dot{\underline{x}} + \delta\dot{\underline{x}}], t) - L(\underline{x}, \dot{\underline{x}}, t) \right] dt
\end{aligned} \tag{G.2}$$

[1] Of course, Newton himself never actually phrased the laws in this way.

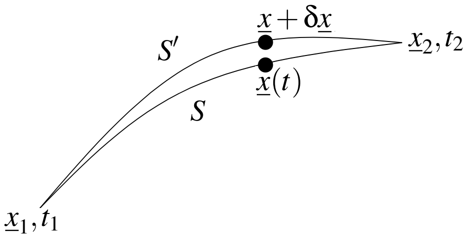


Figure G.1: The principle of least action: a particle will move on the extremum path between two fixed points.

Taylor expanding the left term[2], and using the *summation convention*[3], gives:

$$\begin{aligned}\delta S &= \int_{\underline{x}_1,t_1}^{\underline{x}_2,t_2} \left[L + \frac{\partial L}{\partial x_i}\delta x_i + \frac{\partial L}{\partial \dot{x}_i}\delta \dot{x}_i + O(\delta^2) - L\right] dt \\ &= \int_{\underline{x}_1,t_1}^{\underline{x}_2,t_2} \left[\frac{\partial L}{\partial x_i}\delta x_i + \frac{\partial L}{\partial \dot{x}_i}\delta \dot{x}_i\right] dt \qquad \text{(G.3)}\end{aligned}$$

where x_i is one component of the vector $\underline{x}$.

The second term may be dealt with by integrating by parts and noting that $\delta x(t_1, t_2) = 0$. Thus, we have:

$$\delta S = \int_{\underline{x}_1,t_1}^{\underline{x}_2,t_2} \left[\frac{\partial L}{\partial x_i} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}_i}\right)\right]\delta x_i dt = 0 \qquad \text{(G.4)}$$

and we derive the *Euler-Lagrange equations*:

$$\frac{\partial L}{\partial x_i} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}_i}\right) = 0 \qquad \text{(G.5)}$$

The above equations now allow us to solve for $\underline{x}(t)$ given the Lagrangian, L. But what *is* L? In practice, L, is just whatever mathematical function recovers the correct dynamics equations – in this case Newton's laws. For classical mechanics, we have $L = T - V$ where T is the kinetic energy, and V is the potential energy. It is tempting to ascribe some physical meaning to the above, but this would be a mistake. The Lagrangians for special relativity and electromagnetism do not have such simple forms and so we should think of it only as a coincidence.

Putting $L = T - V = \frac{1}{2}m\dot{x}_i^2 - m\Phi$ into equation G.5 gives:

$$m\left(\frac{\partial \Phi}{\partial x_i} + \ddot{x}_i\right) = 0 \qquad \text{(G.6)}$$

and we recover Newton's second law!

We appear to have put lots of work into developing some mathematical machinery which has just given us (by design) Newton's second law. So what is the point of the above exercise? In the following sections, we shall use the Euler-Lagrange equations to derive some very powerful theorems. We will see that the above is effectively a very useful mathematical trick which makes some problems much easier to solve. However, as you will see on the problem sheet, it can make some problems harder to solve; and some, impossible.

G.1.1 Holonomic constraints

If you remember back to all of those mechanics classes you sat through in your youth, you may remember that one of the most confusing aspects of Newtonian mechanics is getting the reactionary

[2] You should have seen this many times by now. If you need to refresh your memory see Appendix D.

[3] This convention, due to Einstein, means that all repeated indices are summed over: $\frac{\partial L}{\partial x_i}\delta x_i \equiv \sum_{i=1}^{3}\frac{\partial L}{\partial x_i}\delta x_i$.

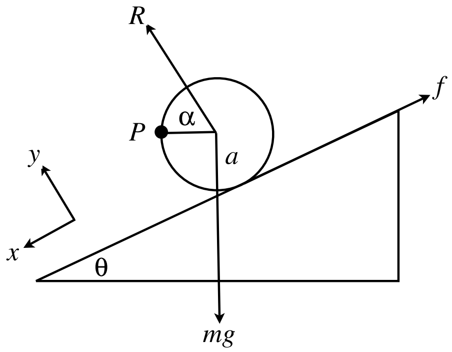


Figure G.2: An example of Newtonian v.s. Lagrangian mechanics: a ball rolling ($f \neq 0$) or sliding ($f = 0$) down an inclined plane. The reaction force and friction forces from the plane, R and f respectively, are marked; neither is necessary in the Lagrangian approach.

forces right; for example the force of reaction of an inclined plane on a ball falling under gravity. Such problems are completely avoided in the Lagrangian approach which derives the equations of motion just from the kinetic and potential energy. However, unfortunately, we do not get something for nothing as we shall show in this section.

Let us consider the inclined plane problem to help to illustrate what is going on. The familiar set-up is shown in Figure G.2: a ball of mass m, radius a, slides down a plane of angle θ. The reactionary force from the plane is marked, R. We assume, for now, that the ball slides *without rolling* ($f = 0$).

The reactionary force must be supposed to exist because otherwise the ball would fall directly through the plane – the force from gravity points downwards, after all! So the reactionary force is a form of *constraint* – the ball is constrained to move on the surface of the plane. Mathematically, we may write the constraint as: $y = 0$. Constraints which may be written in this form ($g(x_i) = 0$) are called *holonomic* constraints. Such constraint equations reduce the *degrees of freedom* for the ball: the number of independent directions the ball can move in.

The entire motion of the ball may be described using just the x coordinate along the plane. This is known as a *generalised coordinate*. It describes the motion of the ball only within the space it is constrained to move.

Just for illustrative purposes, let us now derive the equations of motion for the ball. First using Newton, and then again with the new Lagragian technique. Newton is straightforward: balancing forces along the plane and perpendicular to the plane we have:

$$mg\cos\theta = R \tag{G.7}$$

$$mg\sin\theta = m\ddot{x} \tag{G.8}$$

Now using Lagrangian mechanics:

$$L = \frac{1}{2}m\dot{x}^2 - mgx\sin\theta \tag{G.9}$$

which from equation G.5 gives:

$$m\ddot{x} - mg\sin\theta = 0 \tag{G.10}$$

and notice that we no longer need to even introduce the concept of the reactionary force!

Lagrangian mechanics is great for systems where we can work in generalised coordinates which describe the motion of some particles subject to some *holonomic* constraints. However, we run into problems when we cannot describe the constraints in holonomic form.

Imagine now the same problem as above, but with friction between the plane and the ball ($f \neq 0$). If the ball *rolls without slipping*, the no-slippage constraint gives us: $dx = ad\alpha$ (the angle, α, is defined in Figure G.2). A constraint of this form is *non-holonomic*. Consider a representative point on the surface of the sphere, P. It now necessarily moves in two dimensions. There is no simple constraint which reduces the degrees of freedom of the problem.

We can still solve the problem using Newton and Lagrange, however. First Newton:

$$m\ddot{x} = mg\sin\theta - f \tag{G.11}$$

$$I\ddot{\alpha} = \frac{I}{a}\ddot{x} = fa \tag{G.12}$$

where $I = \frac{2}{5}ma^2$ is the *moment of inertia* of a sphere[4]; and we have used the no-slippage constraint equation to eliminate $\ddot{\alpha}$. Eliminating for the friction force, f, then gives us the equation of motion:

$$m\ddot{x} = mg\sin\theta - \frac{2}{5}m\ddot{x} \tag{G.13}$$

Notice that we have once again required an additional force – the friction, f. In the Lagrangian approach this is not necessary; we can derive the equation of motion directly from the Lagrangian:

$$L = \frac{1}{2}m\dot{x}^2 + \frac{1}{2}I\dot{\alpha}^2 - gmx\sin\theta \tag{G.14}$$

But, wait a minute! We know from Newton that the equation of motion has only one variable, x (see equation G.13 – remember that θ is a constant). Yet the Lagrangian as written above is an equation in two variables: x and α. We may *substitute* α for x using the no-slippage constraint. But we cannot chose a new coordinate system which forces the Lagrangian to have only one generalised coordinate. The above illustrates a key point. There is no holonomic constraint equation which can be applied.

The non-slip condition is an example of a velocity constraint equation: $\dot{x} = a\dot{\alpha}$; these are always non-holonomic. Of course, we can still solve the problem by substituting for $\dot{\alpha}$ and then using equation G.5. This recovers equation G.13 as expected [exercise].

You will try some more astrophysical examples of this on the problem sheet.

G.1.2 Noether's Theorem

We have seen some of the strengths and short-falls of the Lagrangian approach. Now it is time to see its full potential. Consider a Lagrangian which is invariant under a translation: $L(x_i + \delta x_i) = L(x_i)$. Then we can Taylor expand the left term:

$$L(x_i + \delta x_i) = L(x_i) + \delta x_i \frac{\partial L}{\partial x_i} + O(\delta x_i^2) \tag{G.15}$$

Breaking the translation up into a series of infinitesimal steps ($\lim \delta x \to 0$), we recover the definition of differentiation:

$$\frac{L(x_i + \delta x_i) - L(x_i)}{\delta x_i} = \frac{\partial L}{\partial x_i} = 0 \tag{G.16}$$

and from the Euler-Lagrange equation, (equation G.5), we obtain:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}_i}\right) = 0 \tag{G.17}$$

$\Rightarrow$

$$\frac{\partial L}{\partial \dot{x}_i} = m\dot{x}_i = \text{const.} \tag{G.18}$$

A Lagrangian which does not depend on the position of the system *as a whole*, requires linear momentum conservation! This is actually quite a deep result. We observe momentum conservation in the Universe as an empirical fact. However, the above tells us that we must have linear momentum conservation if the laws of physics (which are completely described by the Lagrangian) are to be the same everywhere in space.

[4] Recall that the moment of inertia of a body is a second moment of the density. The first moment is the centre of mass times the mass: $M\overline{\underline{x}} = \int \underline{x}\rho(\underline{x})d^3\underline{x}$. The moment of inertia is the second moment involving the *perpendicular distance* to a given axis. It must in general be a *tensor*. About the centre of mass we have: $I_{ij} = \int (r^2\delta_{ij} - x_i x_j)\rho(\underline{x})\mathrm{d}^3\underline{x}$. Hence $I\ddot{\alpha}$ describes a *torque*. In this problem, we use the moment of inertia for the sphere. In this case, by symmetry, we have $I_{xx} = I_{yy} = I_{zz} = I = \int (x^2 + y^2)\rho d^3\underline{x} = \frac{2}{5}ma^2$. All other terms are zero.

We may perform a similar exercise for angular momentum. We work in cylindrical polars (R, ϕ, z) for simplicity. Consider a Lagrangian invariant under a rotation: $L(\phi + \delta\phi) = L(\phi)$. As above, we can Taylor expand to show that $\frac{\partial L}{\partial \phi} = 0$. Then the Euler-Langrange equations give us:

$$\begin{aligned} \frac{\partial L}{\partial \dot{\phi}} &= \frac{\partial}{\partial \dot{\phi}}\left(\frac{1}{2}mR^2 + \frac{1}{2}mR^2\dot{\phi}^2 + \frac{1}{2}m\dot{z}^2 - V(R,\phi,z)\right) \\ &= mR^2\dot{\phi} \\ &= \text{const.} \end{aligned} \tag{G.19}$$

So if physics is the same whichever way we are facing, then we must have angular momentum conservation.

Finally, we can look at what happens if the Lagrangian is invariant in time. Now we have $\frac{\partial L}{\partial t} = 0$. Thus:

$$\begin{aligned} \frac{dL}{dt} &= \frac{\partial L}{\partial x_i}\dot{x}_i + \frac{\partial L}{\partial \dot{x}_i}\ddot{x}_i \\ &= \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}_i}\dot{x}_i\right) \end{aligned} \tag{G.20}$$

Thus:

$$\begin{aligned} H &= \frac{\partial L}{\partial \dot{x}_i}\dot{x}_i - L \\ &= T + V \\ &= \text{const.} \end{aligned} \tag{G.21}$$

Where H is the *Hamiltonian* of the system and is the total energy of the system.

We have shown that if physics is the same from one moment to the next, is the same independent of direction and is the same from one place to the next, then we must have conservation (globally) of momentum, angular momentum and energy. This is why these three conservation laws are so central to modern physics. They are difficult to get away from – particularly in astronomy. If physics is different from one location to the next, or from one time to the next, then the whole exercise of astrophysics is made extremely difficult. We could no longer reliably apply terrestrial physics to the cosmos!

G.1.3 Rotating reference frames

As another example of Lagrangian mechanics making life easier, let's derive the equations of motion for a *general* rotating reference frame. You may think that you have derived this before using Newton's laws. But little did you know that you really only treated a special case: that of frames rotating at constant angular speed. We will now derive the general result using Lagrangian mechanics with the angular velocity of the frame free to be a function of time: $\underline{\Omega} = \underline{\Omega}(t)$. We will see that there are, in general, more fictitious forces than just the centrifugal and coriolis terms you will have seen before. Sounds hard? Watch as the Euler-Lagrange equations make it easy...

First let's write down the Lagrangian:

$$L = T - V = \frac{1}{2}m|\underline{v}_{\text{in}}|^2 - m\Phi(\underline{x}) \tag{G.22}$$

where $\underline{v}_{\text{in}}$ is the total velocity as observed from an inertial frame. This velocity then includes that of the rotating frame. Thus we have:

$$\underline{v}_{\text{in}} = \underline{\dot{x}} + \underline{\Omega} \times \underline{x} \tag{G.23}$$

where $\underline{\Omega}(t)$ is the angular velocity of the rotating frame.

Now, hopefully you will have seen tensor notation before. Vectors are all very well, but tensor notation makes life so much easier. We will, as above, use the summation convention throughout and from now on x_i refers to an element in the vector $\underline{x}$, and similarly for other quantities. The main reason why this makes life easier is that all quantities then become *scalars* (e.g. x_i is a scalar element of the vector $\underline{x}$) and they commute, add, subtract, multiply and divide just like normal numbers.

In tensor notation the Lagrangian becomes:

$$L = \frac{1}{2}m\left(\dot{x}_i + \epsilon_{ijk}\Omega_j x_k\right)^2 - m\Phi(\underline{x}) \tag{G.24}$$

Hopefully, you will have seen the *Levi-Civita pseudo-tensor*, ϵ_{ijk}, before. Primarily it is used to define the cross product in tensor notation. If you are not familiar with it, have a look in Appendix C. Note also that the notation, $\Phi(\underline{x})$, just means that Φ is a function of all three position coordinates.

Now all we have to do is put the above Lagrangian into the Euler-Lagrange equations. Simple! We have:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}_l}\right) = \frac{d}{dt}\left[m\left(\dot{x}_i + \epsilon_{ijk}\Omega_j x_k\right)\delta_{il}\right] \tag{G.25}$$

where δ_{il} is the Kronecker delta (see Appendix C). An important point to note here. Remember when differentiating in tensor notation to always introduce a new subscript (in this case l). *Don't use any of the subscripts already in use.* You will end up in a horrible mess that way. Remember that all of the other subscripts in the Lagrangian are being summed over! (they are called 'dummy' indices).

Now for the next part of the Euler-Lagrange equations:

$$\frac{\partial L}{\partial x_l} = m\left(\dot{x}_i + \epsilon_{ijk}\Omega_j x_k\right)\epsilon_{ipq}\Omega_p\delta_{ql} - m\frac{\partial \Phi}{\partial x_l} \tag{G.26}$$

Putting the above together, and returning to vector notation, we recover:

$$m\left(\ddot{\underline{x}} + \dot{\underline{\Omega}} \times \underline{x} + 2\underline{\Omega} \times \dot{\underline{x}} + \underline{\Omega} \times \underline{\Omega} \times \underline{x}\right) + m\underline{\nabla}\Phi \tag{G.27}$$

We now see that, in general, there are *three* fictitious forces. The term on the left is the new one: the *inertial force of rotation*, the middle term is the coriolis force, and the right term is the centrifugal force. We have been using the centrifugal force throughout the course so far. We see that in the special case of circular orbits, $\underline{\Omega} = \Omega\hat{\underline{z}}$, and the centrifugal term reduces to the familiar: $m\Omega^2 R$ (in cylindrical coordinates: (R, ϕ, z)).

G.2 Hamiltonian mechanics

At the end of section G.1.2, we introduced a new quantity – the Hamiltonian, H, of a system. This describes the total energy of the system and we may use it to form Hamiltonian mechanics. This is really just another way of formulating Lagrangian mechanics. But as we shall see, it is also useful for quickly solving some otherwise difficult problems.

We first define more rigourously what we mean by *generalised coordinates.* We mentioned these briefly in section G.1.1. If we have a generalised position coordinate, q_i, then we may define the generalised momentum:

$$p_i \equiv \left(\frac{\partial L}{\partial \dot{q}_i}\right) \tag{G.28}$$

We now have generalised coordinates which completely describe the particle distribution at a given moment: q_i, p_i. Recall that these are the coordinates which describe the space constrained by the holonomic constraint equations. However, we may think of them as 'position' and 'momentum'. The space they describe is called *phase space.* In general this space has 6 dimensions per particle; in practice, the constraint equations reduce this dimensionality.

Now, we wish to reformulate the Euler Lagrange equations in terms of the generalised coordinates. This will give us *Hamilton's equations* of motion.

Recall from section G.1.2, we defined the Hamiltonian as:

$$H = p_i\dot{q}_i - L \tag{G.29}$$

and we proved that if L does not depend explicitly on time then H represents the total energy of the system and is conserved.

Using the Euler-Lagrange equations, it is straightforward to prove that:

$$\frac{\partial H}{\partial q_i} = -\dot{p}_i; \qquad \frac{\partial H}{\partial p_i} = \dot{q}_i \tag{G.30}$$

These are *Hamilton's* equations. They are a reworking of the Euler-Lagrange equations, but now in the generalised coordinates: (q_i, p_i).

So the above is all just fancy mathematics at the moment. What do we gain from it? Well, generalised coordinates are very useful. The Euler-Lagrange equations are somewhat hampered by the fact that we must choose a coordinate, q_i, and its time derivative, $\dot{q}_i$ to represent the system. Now, we have no such limitation. We can represent our system using *any* two independent coordinates q_i and p_i. Notice the *symmetry* in Hamilton's equations. This will help you understand what we have achieved by using generalised coordinates. We can now completely swap $p_i \rightarrow q_i$ and $q_i \rightarrow -p_i$. Although p_i is called a generalised momentum, we can quite happily use it to represent position: $p_i = x_i$, with the generalised coordinate representing momentum $q_i = -m\dot{x}_i$, if we like.

G.2.1 Canonical transformations

We will see how far these generalised coordinates can take us in the following sections. However, first we must define a *canonical coordinate transformation.* This will allow us to transform between different generalised coordinates, while ensuring that Hamilton's equations still hold true.

Substituting the Hamiltonian (equation G.29) into equation G.2, we can now define a variational principle for Hamiltonian mechanics – just as we did for Lagrangian mechanics:

$$\delta S = \delta \int (p_i \dot{q}_i - H)\, dt = 0 \tag{G.31}$$

Now, imagine we switch to some new phase space coordinates, (Q_i, P_i), with a new associated Hamiltonian, F. For the new coordinates, we may also write:

$$\delta S = \delta \int \left(P_i \dot{Q}_i - F\right) dt = 0 \tag{G.32}$$

The key point here is that we are only interested in transformations which *preserve the dynamics of the system.* This means that if I travel in a loop from one fixed point in phase space to another, and back again, the physical path taken in each coordinate system must be the same. Mathematically speaking this means that:

$$\oint \left[(p_i \dot{q}_i - H) - \left(P_i \dot{Q}_i - F\right)\right] dt = 0 \tag{G.33}$$

The above will always be true if we state that the integrand is given by the absolute time derivative of some function, G:

$$(p_i \dot{q}_i - H) - \left(P_i \dot{Q}_i - F\right) = \frac{dG}{dt} \tag{G.34}$$

such that:

$$\oint \frac{dG}{dt} dt = \oint dG = 0 \tag{G.35}$$

Coordinate transformations of the above sort are called *canonical transformations*, and we now prove an important property of them.

Splitting up G in the following way:

$$dG = p_i dq_i - P_i dQ_i + dG' \tag{G.36}$$

with $dG' = (-H + F)dt$, we find that:

$$\oint p_i dq_i - P_i dQ_i = 0 \tag{G.37}$$

since the contribution from dG' cancels in integrating around a closed loop. This proves an important property of canonical transformations: $\oint p_i dq_i$ is conserved. We will return to why this is important shortly.

But how do we actually perform such a transformation? We can understand this by asserting some form for G. The function G is arbitrary, but a useful choice (and we shall see why below) is:

$$G = S(P_i, q_i, t) - P_i Q_i \tag{G.38}$$

From equation G.34, and substituting equation G.38, we obtain:

$$p_i dq_i - H dt - P_i dQ_i + F dt - d(S - P_i Q_i) = 0 \tag{G.39}$$

gathering terms together, we have:

$$\left(p_i - \frac{\partial S}{\partial q_i}\right) dq_i + \left(\frac{\partial S}{\partial P_i} - Q_i\right) dP_i + \left(F - H - \frac{\partial S}{\partial t}\right) dt = 0 \tag{G.40}$$

and we obtain:

$$p_i = \frac{\partial S}{\partial q_i}; \qquad Q_i = \frac{\partial S}{\partial P_i}; \qquad F = H + \frac{\partial S}{\partial t} \tag{G.41}$$

We can now use the above equations to transform from (q_i, p_i) to (Q_i, P_i). The function, S, is called the *generating function*, and it defines the canonical coordinate transform.

G.2.2 The Hamilton-Jacobi equation

Now it's time to start showing you *why* we went to so much trouble to pose our dynamics equations in such abstract terms. The key is understanding a cunning trick due to Jacobi. He realised that if we can find some phase space coordinates (Q_i, P_i) in which the Hamiltonian vanishes, $F = 0$, then Hamilton's equations become:

$$\dot{P}_i = 0; \qquad \dot{Q}_i = 0 \tag{G.42}$$

Both P_i and Q_i are now constants of the motion. In one step, we have completely solved the problem. The transformation which achieves this may be written down directly from equations G.41. The result is the Hamilton-Jacobi (H-J) equation:

$$H(\frac{\partial S}{\partial q_i}, q_i, t) + \frac{\partial S}{\partial t} = 0 \tag{G.43}$$

All we need to do is to solve the above equation for the generating function, S. We then obtain Q_i from $Q_i = \frac{\partial S}{\partial P_i}$, and the P_i are the integration constants of equation G.43. This is a more significant step than Hamilton's equations themselves. Recall that Hamilton's equations are ultimately just a reworking of Newton's laws. But if we can solve the H-J equation, then we have actually solved Newton's equations of motion completely.

In practice, however, solving the H-J equation is not that useful because $S = S(t)$. Our transformation which solves the problem evolves with time! But, if we assume from the start that S is some simple function of time: $S(t) = \alpha t + \beta$, then we can reduce the H-J equation to an even simpler form:

$$H(P_i) = \text{const.} \tag{G.44}$$

The above equation is really useful. Now Hamilton's equations reduce to:

$$P_i = \text{const.}; \qquad Q_i = At + B \tag{G.45}$$

As above, if we can find the generating function, S, then our dynamics problem is fully solved. But now S takes a much simpler time independent form.

In general, there will be many many cases where there are *no solutions* to the H-J equation. However, as we shall see next, even these cases tell us something important about the dynamics of the system.

G.2.3 Actions & integrals

Solving equation G.44 gives us the generating function, S which transforms $(q_i, p_i) \to (Q_i, P_i)$ such that $P_i = \text{const.}$. Such constants are arbitrary – they are just constants of integration which come out of the H-J equation. As such, we may further constrain S by *choosing* our constants of motion in advance. A natural choice is the conserved loop integral we encountered earlier. We define the *action* of a system as being:

$$A_i = \frac{1}{2\pi} \oint_{\gamma_i} p_i dq_i \tag{G.46}$$

where γ_i describes an independent path through phase space. In general, there will be *three* such independent paths – one for each $P_i = A_i$.

Actions are useful for two main reasons. Firstly, they are *isolating integrals* of the motion. An *integral* of the motion is conserved along the trajectory of the particle. An isolating integral is even more useful: it lowers the dimensionality of the phase space available to the particle. An example you have already come across is the energy, which is an isolating integral for any static potential. (This is a direct result of gravity being a conservative force.) If no other isolating integrals exist, then the particle is free to roam throughout all of phase space – as far as its energy will allow. Such an orbit will be *chaotic* and it will not be a solution of the H-J equation. The other extreme is an orbit with five isolating integrals of motion. Such an orbit will be completely confined to a line in phase space. The Kepler orbit is the classic example of this.

Actions are particularly special isolating integrals. This is because, having found the actions, the motion of the particle is then fully parameterised by the other phase space coordinates, Q_i. Recall that, by construction, these evolve linearly with time: $Q_i = At + B$ (see equation G.45). By using the actions as our canonical momenta, the phase space coordinates, Q_i, have a simple physical meaning. They are *angles* which describe the motion of the particle in time around the phase space loop which defines the associated action. This is probably quite difficult to picture right now. But don't worry, I promise you will see this much more clearly, when we consider a specific an example, below.

The second advantage of using actions is that they are *adiabatic invariants* of the motion. They remain unchanged if the gravitational potential changes slowly enough (strictly speaking, this means infinitely slowly). This can be understood if you think of an infinitesimal change to the Hamiltonian: $H \to H + \delta H$, which is the result of the change in the gravitational potential. We may equate $\delta H = \frac{\partial S}{\partial t}$ with the generating function (see equation G.41). Now we see that the adiabatic change is just a canonical map (a canonical coordinate transformation). But we have already proved that canonical maps conserve the actions, and so we have now proven that actions must be adiabatically invariant.

G.2.4 A worked example: the simple harmonic oscillator

As a worked example, let's consider the favourite hobby-horse of physics: the simple harmonic oscillator. It is a toy model which students encounter time and time again – on first dynamics courses, right through to quantum mechanics. But often students are not often told *why* such a simple system could be important in so many areas of physics.

Consider the minimum of some 1D potential well[5] which takes the form $\Phi(x)$, where the minimum is at $x = 0$. Here Φ can represent any scalar field, but we can think of it, in keeping with this course, as being a gravitational field. Now, suppose that we displace the particle from its equilibrium, by some small amount, x. We may now Taylor expand about the minimum to obtain:

$$\Phi(x) = \Phi(0) + x\Phi'(0) + \frac{x^2}{2!}\Phi''(0) + O(x^3) \tag{G.47}$$

Since we are at a minimum, the potential at $x = 0$ must be stationary, $\Phi'(0) = 0$; while $\Phi(0)$ and $\Phi''(0)$ are just constants. Thus, to order x^3, we recover the potential of a simple harmonic oscillator[6]:

[5] We work in one dimension for simplicity, but all of these arguments are equally valid in three dimensions [exercise].

[6] Just in case you've never seen this potential before, it is trivial to solve the dynamics using Newton's second law. We have $\ddot{x} = -\frac{d\Phi}{dx} = -2Ax$ – the equation of motion for a simple harmonic oscillator!

$$\Phi(x) = Ax^2 + B \tag{G.48}$$

Now we see why the simple harmonic oscillator pops up time and again in physics. It is because a small perturbation about *any minima* will produce approximate simple harmonic motion.

For the following example, let's assume that $A = \frac{1}{2}$ and $B = 0$, since the constants are arbitrary. Thus we obtain a Hamiltonian:

$$H = \frac{p^2}{2} + \frac{x^2}{2} = E \tag{G.49}$$

where $p = m\dot{x}$ is the momentum, we assume $m = 1$, and E is the energy as usual.

Let's solve the problem first using Hamilton's equations and then again using our new H-J equation. This will help you to understand what all of the above mathematics really *means* in practice.

First Hamilton. Hamilton's equations give:

$$\frac{\partial H}{\partial x} = -\dot{p} = x; \qquad \frac{\partial H}{\partial p} = \dot{x} = p \tag{G.50}$$

which gives, for the equation of motion:

$$\ddot{x} = -x \tag{G.51}$$

The familiar equation of motion for a simple harmonic oscillator. Integrating we obtain:

$$x = A\sin(t + B) \tag{G.52}$$

where A and B are constants of the integration.

All straightforward so far. Now let's try again using the H-J equation. In some sense, this is a bit like using a sledge hammer to play the piano, but it will give you a better understanding of how to really use the results we have derived above (the problem sheet will also help you to develop this further).

We know that the Hamiltonian doesn't depend explicitly on time and so must be the (conserved) energy: $H = E$. Now the aim is to find new generalised coordinates, (Q_i, P_i), which make $H(P_i) =$ const.. We are working in 1 dimension, so will drop the subscript, i, from now on – there can only possibly be one action since there is only one momentum coordinate, p, in the first place.

The H-J equation is now:

$$\frac{1}{2}\left(\frac{\partial S}{\partial x}\right)^2 + \frac{1}{2}x^2 = E \tag{G.53}$$

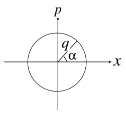
where all we have done is substitute for $p = \frac{\partial S}{\partial x}$ using equations G.41. It is now a simple matter to solve equation G.53 for the generating function, S:

$$S = \int \sqrt{2E - x^2} dx \tag{G.54}$$

Now, you can see that there is an arbitrary integration constant which will come from equation G.54. We can fix this by deciding that our conserved momentum – P will be an *action*. Recall that this decision is arbitrary, but that actions are more physically meaningful than other choices because they are *isolating integrals*. From the definition of an action (equation G.46) we have:

$$P = \frac{1}{2\pi}\oint p dx \tag{G.55}$$

Now, we require a bit of care here. We are integrating around a closed loop in phase space. This is best done in plane polar coordinates since this will make it clear what the integration limits should be. We substitute: $x = q\cos\alpha, p = q\sin\alpha$ (see the diagram in the margin). Now the integration limits range from $0 \leq \alpha \leq 2\pi$ in these coordinates, and, by convention, we take the loop integral in a *clockwise* direction.



In general, $q = q(\alpha)$ in these new coordinates. We may calculate the shape of the phase loop, $q(\alpha)$, by substituting our new coordinates into the Hamiltonian. Rearranging equation G.49, we obtain:

$$\begin{aligned}p(x) &= \sqrt{2E - x^2} \Rightarrow \\ q^2 \sin^2 \alpha &= 2E - q^2 \cos^2 \alpha \Rightarrow \\ q &= \sqrt{2E}\end{aligned} \tag{G.56}$$

In this case, the loop in phase space is a circle of radius $\sqrt{2E}$. Using $dx = dq \cos \alpha - q \sin \alpha d\alpha = -q \sin \alpha d\alpha$, we have in our new coordinates:

$$\begin{aligned}P &= \frac{1}{2\pi} \oint p dx \\ &= -\frac{1}{2\pi} \int_0^{2\pi} -q^2 \sin^2 \alpha d\alpha \\ &= \frac{2E}{2\pi} \int_0^{2\pi} \frac{1 - \cos(2\alpha)}{2} d\alpha \\ &= E\end{aligned} \tag{G.57}$$

Note the minus sign in the second line which comes from the clockwise integration direction.

We find that, for the special case of the harmonic oscillator, the action is just the energy, E. But this is not true in general, so take note! It does highlight an important point, however. The actions provide all of the isolating integrals required to describe a system. We know that in any static potential the energy will be an isolating integral (remember, this is just because gravity is a conservative force). So it is not surprising that, when we can have only one action (as in this 1D case), it is simply related to the energy.

Now, all that is left is to calculate the other canonical phase space coordinate:

$$Q = \frac{\partial S}{\partial P} = \frac{\partial S}{\partial E} = \int \frac{1}{\sqrt{2E - x^2}} dx \tag{G.58}$$

and, using a standard trigonometric substitution, we recover the equation of motion 'directly':

$$x = \sqrt{2E} \sin(Q - Q_0) \tag{G.59}$$

where Q_0 is just an integration constant.

But what does G.59 mean? We have solved the problem in the new coordinates $P = E$ and Q, but we must now think about what these coordinates are physically. The simplest way to do this is to plug our solution back into the Hamiltonian. This allows us to solve also for p and gives:

$$p = \sqrt{2E} \cos(Q - Q_0) \tag{G.60}$$

which satisfies $x^2 + p^2 = 2E$. Now we can see the meaning of P and Q. The trajectory of the particle is a circle in phase space, (x, p) (see Figure G.3). Q is now an *angle* in that phase space, while P is $\frac{1}{2\pi} \times \{\text{integral of phase space area}\}$. That is, the radius of the circle in phase space is $\sqrt{2E}$, and its area is then $\pi 2E = 2\pi P$.

Recall that, by construction, we know $Q = At + B$, where A and B are arbitrary constants (remember, we chose to transform to phase space coordinates where this is so!). This now proves that equations G.52 and G.59 are indeed the same solution.

The above exercise illustrates a final useful thing about choosing to use actions as our conserved canonical momenta. The Q coordinate is now an *angle* in the phase space (x, p). By using actions, this will always be the case. The actions, P_i, represent integrals over phase space area in the coordinates, (q_i, p_i), while the angle variables, Q_i, represent angles around the loop integral. Can you think of what shape three action-angle coordinates would trace out in phase space? [Hint: each action coordinate must be independent, and you know the answer in 1D is a circle].

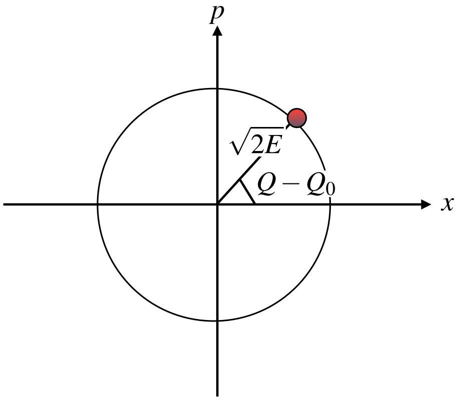


Figure G.3: The trajectory of a 1D harmonic oscillator in phase space: (x, p). Marked also are the new canonical coordinates, (Q, P), for which $H = H(P) = \text{const}$. By fixing P to be an action, the physical meaning of these new coordinates is now clear. Q represents an *angle* around the particle trajectory in phase space; while P is the area of the particle trajectory divided by 2π. Q now fully parameterises the position of the particle in phase space.

G.3 Phase space and Liouville's Theorem

We conclude this chapter will one final proof which we be of great importance later on in the course. Recall that in general, phase space is filled with N particles, each with six phase space coordinates – a 6N dimensional space! We can write this in the following compact form:

$$\omega_{i,j} = (q_{i,j}, p_{i,j}) \qquad 1 \leq i \leq 3\,;\, 1 \leq j \leq N \tag{G.61}$$

were $\omega_{0,j}$, $\omega_{1,j}$ and $\omega_{2,j}$ form three independent vector fields. These act like three separate 'axes' for our 6N dimensional phase space. Each contains $2N$ elements representing an independent component of the phase position for each particle.

With the above compact notation, we can now write down a 'velocity' of our particles in phase space. I have carefully put the word 'velocity' in quotes because really I mean the rate of change of the canonical coordinates. This should not be confused with 'velocity' in the old Newtonian sense of the word. We have:

$$\dot{\omega}_{i,j} = (\dot{q}_{i,j}, \dot{p}_{i,j}) \qquad 1 \leq i \leq 3\,;\, 1 \leq j \leq N \tag{G.62}$$

Now, things are probably getting a little hairy right now. How on earth are we to visualise 4 dimensional space – let alone 6N dimensional space? One trick to help you with this is to exploit our old friend: *symmetry*. For example, we know that a sphere is a three dimensional object. However, we only need one number to completely describe a perfect sphere. As long as we know that it is a perfect sphere, all we need is the radius. The same is true of an N-sphere in higher dimensions. Although it becomes impossible to visualise the sphere itself, we know it is an object which is fully described by one radius and that radius will be a magnitude in N-dimensional space: $r^2 = x_1^2 + x_2^2 + \cdots x_n^2$.

Bearing the above in mind, imagine that our 6N dimensional phase space fills some 6N dimensional 'volume'. Again this is not a volume in the usual sense, and hence the quotes. The particles in phase space need not have any symmetry whatsoever, but picture it in your mind as a sphere for now. Now, since $\dot{\omega}_{1,j}$ is just a 2N-dimensional vector field (and similarly for $\dot{\omega}_{2,j}$ and $\dot{\omega}_{3,j}$), we may write down the 2N-dimensional analogue of the *divergence theorem*:

$$\int_V \underline{\nabla}_i \underline{\dot{\omega}}_i dV = \int_S \underline{\dot{\omega}}_i \cdot d\underline{S}_i \tag{G.63}$$

where the above vectors represent the $1 \leq j \leq N$ components of $\dot{\omega}_{i,j}$.

The integral on the right represents the flow of particles in phase space through the 'surface' S, bounding the 'volume' V. Now it becomes clear why imagining the sphere helps. Because we are picturing a flow through a 'surface' one dimension lower than the 'volume', we really can imagine the problem. The flow of particles through the 6N-1 hypersurface into the 6Nth dimension can be understood intuitively by imagining the flow through a 2D surface bounding a sphere. Note also that we have substituted our slightly dodgy 'surface' for the correct terminology: *hypersurface*.

Now notice that:

$$\begin{aligned} \underline{\nabla}_i \underline{\dot{\omega}}_i &= \frac{\partial}{\partial q_{i,j}}\dot{q}_{i,j} + \frac{\partial}{\partial p_{i,j}}\dot{p}_{i,j} \\ &= \frac{\partial^2 H}{\partial q_{i,j}\partial p_{i,j}} - \frac{\partial^2 H}{\partial p_{i,j}\partial q_{i,j}} \\ &= 0 \end{aligned} \tag{G.64}$$

where the second line follows by substituting Hamilton's equations (equations G.30).

The above proves *Liouville's theorem*:

Hamiltonian flow preserves phase space volume (and therefore density) for any region of phase space.

Another way of expressing the above is that the particles evolve in phase space as an incompressible fluid. Without solving any equations, or doing any dynamics the above tells us two very important things:

1. A boundary in phase space always encloses the same group of particles.

2. From 1. we see that phase trajectories don't cross.

But we must be careful. It can be very difficult to picture 6N dimensional space. In projections of this space onto lower dimensional subspaces, particles can appear to cross all the time. Again, we can gain some intuition for this by imagining lower-dimensional analogues. In three dimensions, for example, a simple example is two sets of lines which are confined to two parallel planes. The lines in one plane will never cross the lines in the other plane, since the two planes are parallel. But they will *appear* to cross one another when viewed along some projections.

In discussing projections onto lower dimensional spaces, there is a common confusion which is worth highlighting. A system of 6N particles is just *one point* in 6N dimensional phase space. When we say that trajectories don't cross in 6N dimensional phase space, we really refer to different copies of our original system, each copy being another point in 6N dimensional space. When viewed like this, Liouville's theorem seems less useful. Much more useful is the 6 dimensional *special case* which applies only to collisionless systems. If particles are collisionless and therefore do not interact with one another at all (they do not exchange energy, angular momentum nor any other property), then Liouville's theorem also applies in 6 dimensional phase space[7]. This more restricted version of the theory means that individual phase trajectories of particles in a collisionless system will not cross; each point in phase space represents a unique trajectory. The theorem now applies to individual particle orbits, rather than copies of the system as a whole and is, correspondingly, more useful.

[7] Hopefully it is clear why. If particles do not interact then we immediately remove $N-1$ degrees of freedom per particle.

Appendix H

Dynamical friction

A related physical effect to relaxation is that of *dynamical friction*. Dynamical friction is of much broader relevance in astrophysics than relaxation since it proceeds much faster. The basic idea is outlined in Figure H.1. A larger body, or 'satellite', moves through a medium of smaller bodies. Through successive scattering events, the larger body loses energy to the smaller ones and slows down. The physics of the interaction is the same as in relaxation: scattering. However, because of the disparity in mass between the two bodies, dynamical friction proceeds much faster, as we shall prove.

Dynamical friction likely affects the distribution of globular clusters in our Galaxy, the distribution of galaxies in a cluster of galaxies and the rate of accretion of satellite galaxies in the Universe. Figure H.2, for example, shows the stream of stars coming of the Sagittarius dwarf galaxy. This small companion to the Milky Way was accreted in the last few gigayears. It is a fascinating galactic interaction which is helping us to learn about the shape of our Galactic potential, about dark matter v.s. alternative gravity and about the role of galaxy interactions in galaxy formation in general. Its rate of infall is undoubtedly affected by dynamical friction between the galaxy and the background stars and dark matter.

H.1 The Chandrasekhar approach

The first description of dynamical friction was that due to Chandrasekhar 1943. It is still an extremely accurate rule of thumb some sixty years later and a good place to start. The derivation is very similar to that of the relaxation time presented in Lecture 1. Things are made a little more complicated, however, by the different mass of the infalling body and the background particles. Our strategy is as follows:

- Derive the effect of one scattering event.
- Integrate over all impact parameters.
- Integrate over all particles.

We will assume an infinite, homogeneous background distribution of particles. Later we will call this assumption to question and we shall see that it does miss out some very important physics in special cases.

First consider the two body interaction. The two body problem may always be transformed into a Kepler problem for a fictitious *reduced particle* moving a *reduced* Kepler potential. The equation of motion is:

$$\frac{mM}{m+M}\ddot{\underline{r}} = -\frac{GMm}{r^2}\hat{\underline{r}} \tag{H.1}$$

where m and M are the masses of the background particles and the infalling body, respectively, and $\underline{r} = \underline{x}_m - \underline{x}_M$; $m\dot{\underline{x}}_m + M\dot{\underline{x}}_M = 0$.

Let us define the change in velocity of the reduced particle due to the interaction as: $\Delta\dot{\underline{r}}$. From the above definitions, we have:

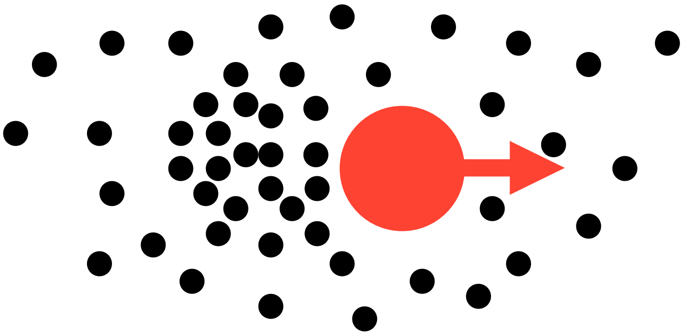


Figure H.1: Schematic diagram of Chandrasekhar dynamical friction. The larger body marked in red moves in the direction marked by the arrow, like a marble falling through honey. It scatters stars in front of it leading to an overdensity of stars behind it – a *wake*. The momentum transfer between the larger body and the background due to these gravitational scattering events causes the larger body to slow down. This is the dynamical friction.

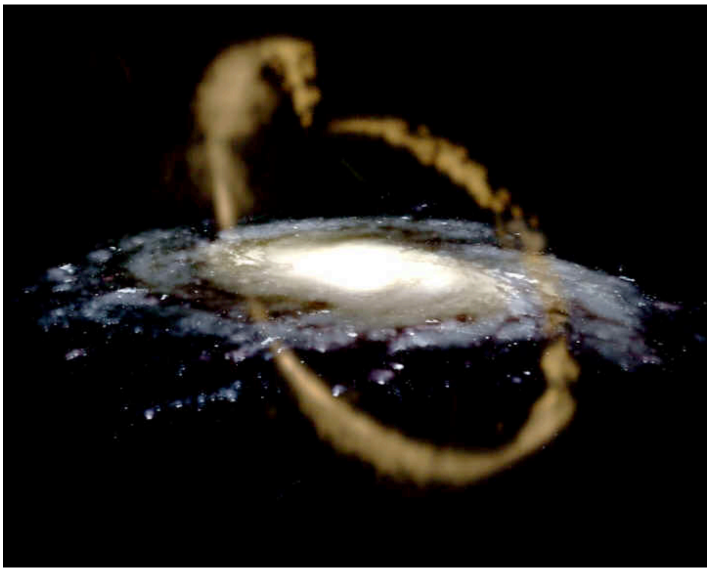


Figure H.2: A view of the Sagittarius stream as it would be seen by an alien sitting outside our Galaxy. The stream is tidal debris torn off of the Sagittarius dwarf galaxy as it fell into the Milky Way little over a billion years ago. The rate of infall of the satellite was undoubtedly governed by dynamical friction.

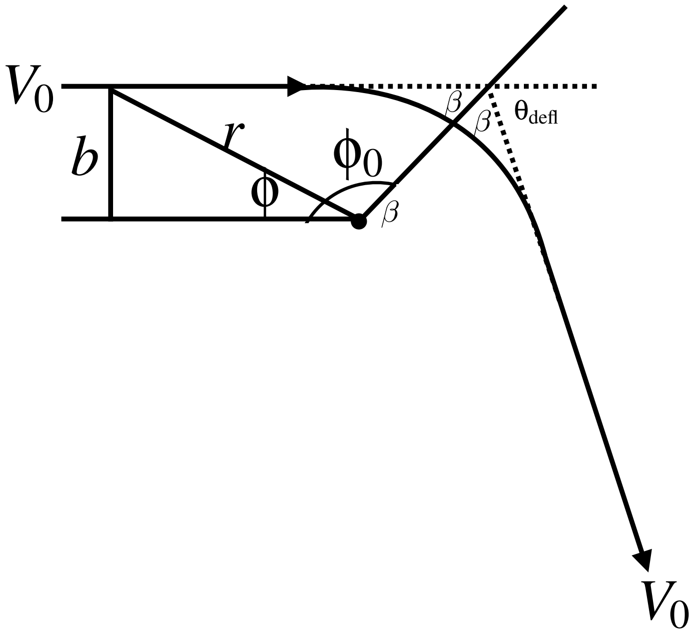


Figure H.3: Schematic of the two-body scattering event, see text for details. Note that b marks the *impact parameter*, while ϕ_0 is the angle at which the two bodies are at closest approach.

$$\Delta\underline{\dot{r}} = \Delta\underline{\dot{x}}_m - \Delta\underline{\dot{x}}_M \tag{H.2}$$

$$m\Delta\underline{\dot{x}}_m + M\Delta\underline{\dot{x}}_M = 0 \tag{H.3}$$

Eliminating $\Delta\underline{\dot{x}}_m$, we obtain:

$$-\Delta\underline{\dot{r}}\left(\frac{m}{m+M}\right) = \Delta\underline{\dot{x}}_M \tag{H.4}$$

Now we wish to solve for the change in velocity $\Delta\underline{\dot{r}}$, which we do using the solution for Kepler orbits. To help you picture the geometry, the scattering event is shown in Figure H.3.

For a Kepler orbit we have:

$$\frac{1}{r} = C\cos(\phi - \phi_0) + \frac{G(M+m)}{L^2} \tag{H.5}$$

where L is the specific angular momentum. The above equation is just the standard Kepler solution relating the radius and angle (it is usually written in terms of eccentricity and semi-major axis, however).

From Figure H.3, we can see that the angular momentum is given by: $L = V_0 b$, while we know that $L = r^2\dot{\phi}$. Differentiating equation H.5 with respect to time and using the above substitutions, we obtain:

$$\dot{r} = CbV_0\sin(\phi - \phi_0) \tag{H.6}$$

At the start of the interaction we have $r = -\infty$, $\phi = 0$ which gives:

$$-V_0 = CbV_0\sin(-\phi_0) \tag{H.7}$$

$$0 = C\cos(-\phi_0) + \frac{G(M+m)}{b^2V_0^2} \tag{H.8}$$

and eliminating for C, we obtain:

$$\tan\phi_0 = \frac{-bV_0^2}{G(M+m)} \tag{H.9}$$

Now we need a little geometry, which is given also in Figure H.3 (a good Figure often helps a lot!!). Notice that the scattering event is *symmetrical* about the point of closet approach, $\phi = \phi_0$. Thus the final perpendicular and parallel components of the velocity are given by:

$$V_{f,\mathrm{perp}} = V_0 \sin\theta_{\mathrm{defl}} = V_0\sin(\pi - 2\beta) = V_0\sin(2\phi_0 - \pi) = -V_0\sin(2\phi_0) \tag{H.10}$$

and similarly:

$$V_{f,\mathrm{para}} = V_0\cos(2\phi_0) \tag{H.11}$$

Thus, using equations H.9 and H.4 and some trig substitutions, the *change* in the velocity components of $\underline{\dot{x}}_M$ are given by:

$$|\Delta\underline{\dot{x}}_{M,\mathrm{perp}}| = \frac{2mbV_0^3}{G(M+m)^2}\left[1 + \frac{b^2V_0^4}{G^2(M+m)^2}\right]^{-1} \tag{H.12}$$

$$|\Delta\underline{\dot{x}}_{M,\mathrm{para}}| = \frac{2mV_0}{M+m}\left[1 + \frac{b^2V_0^4}{G^2(M+m)^2}\right]^{-1} \tag{H.13}$$

Now that was one interaction. Over many interactions, the mean effect of perpendicular encounters will average to zero (they can have either sign). However, the important point is that the parallel encounters will *always reduce the velocity of the heavier object.* This is the dynamical friction effect. Of course, the perpendicular encounters do not vanish when you consider the *root mean square effect.* This is just the relaxation we already derived in Lecture 1. Over many many encounters, the heavy particle will random walk in the perpendicular component of its velocity. But this is a *tiny* effect compared to dynamical friction which will *always* reduce the parallel velocity component.

It is worth a brief paragraph at this point to discuss exactly *why* it is that the parallel component always points in one direction (i.e. decelerates the infalling body). What we have done, above, is to analyse the problem from the centre of mass frame of M and m. Now, in the limit $m = M$, $\Delta\underline{\dot{x}}_m = -\Delta\underline{\dot{x}}_M$: the particles receive equal and opposite momentum kicks, as expected. Provided that we have a 'typical' particle – i.e. not moving at significantly greater velocity than the mean of the background, then the net effect of such encounters will be zero in the mean (although relaxation will still occur through a random walk). Once the masses of the particles are significantly different, however, an asymmetry enters into the problem. We transform into the centre of mass frame which is now close to the rest frame of the massive particle. In this frame, the massive particle is nearly stationary and sees a 'head wind' from the background particles which must be moving towards the massive body. Hence the friction force is always a deceleration.

To obtain the final friction force, as in Lecture 1, we must now integrate over all impact parameters and all particles. First the impact parameters. The *rate* at which the infalling body encounters background particles with velocity density $f(\underline{\dot{x}}_m)$ within the range $b \to b + db$ is given by:

$$2\pi b db V_0 f(\underline{\dot{x}}_m) d^3\underline{\dot{x}}_m \tag{H.14}$$

Thus we obtain:

$$\left.\frac{d\underline{\dot{x}}_M}{dt}\right|_{\underline{\dot{x}}_M} = \underline{V}_0 f(\underline{\dot{x}}_m) d^3\underline{\dot{x}}_m \int_0^{b_{\max}} \frac{2mV_0}{M+m}\left[1 + \frac{b^2V_0^4}{G^2(M+m)^2}\right]^{-1} 2\pi b db \tag{H.15}$$

$$= 2\pi\ln(1+\Lambda^2)G^2m(M+m)f(\underline{\dot{x}}_m)d^3\underline{\dot{x}}_m\frac{(\underline{\dot{x}}_m - \underline{\dot{x}}_M)}{|\underline{\dot{x}}_m - \underline{\dot{x}}_M|^3} \tag{H.16}$$

Remember that we assume an *infinite* homogeneous medium. To avoid divergences in the integral, just like for the relaxation time, we have to add some cut-off which is the 'maximum size of the system': $b_{\max}$. We also come across, once again, the *Coulomb Logarithm* which we encountered in deriving the relaxation time:

$$\Lambda \equiv \frac{b_{\rm max} V_0^2}{G(M+m)} \tag{H.17}$$

and, since $\Lambda \gg 1$, typically:

$$\ln(1+\Lambda^2) \simeq 2\ln\Lambda \tag{H.18}$$

Now we integrate over all of the particles. Here we assume isotropy and use a cunning trick! Notice how the right hand part of equation H.16 looks like an integral over a velocity density times a function which looks like a 'gravitational potential', but in the velocity. If the velocity is isotropic, then just like a spherical gravitational potential, we can use the Newton II theorem (Binney and Tremaine 2008):

$$\int f(\dot{\underline{x}}_m)\frac{(\dot{\underline{x}}_m - \dot{\underline{x}}_M)}{|\dot{\underline{x}}_m - \dot{\underline{x}}_M|^3} d^3\dot{\underline{x}}_m = \frac{4\pi\int_0^{\dot{x}_M} f(\dot{x}_m)\dot{x}_m^2 d\dot{x}_m}{\dot{x}_M^2}\hat{\dot{\underline{x}}}_M \tag{H.19}$$

If the above looks strange, just imagine $\dot{\underline{x}}_m \to \underline{x}$ and you'll see that it just amounts to: $\int \rho\frac{(\underline{x}-\underline{x}')}{|\underline{x}-\underline{x}'|^3}d^3\underline{x}' = \frac{1}{G}\underline{\nabla}\Phi = \frac{M(r)}{r^2}\hat{\underline{x}}$, which is the familiar Newton II.

Putting it all together, we obtain the Chandrasekhar dynamical friction formula:

$$\frac{d\dot{\underline{x}}_M}{dt} = -16\pi^2\ln\Lambda G^2(M+m)\frac{\int_0^{\dot{x}_M} mf(\dot{x}_m)\dot{x}_m^2 d\dot{x}_m}{\dot{x}_M^2}\hat{\dot{\underline{x}}}_M \tag{H.20}$$

We may further simplify this if we assume a Maxwellian distribution of velocities:

$$f(\dot{x}_m) = \frac{n}{(2\pi\sigma^2)^{\frac{3}{2}}}e^{-\frac{\dot{x}_m^2}{2\sigma^2}} \tag{H.21}$$

where n is the local number density of particles and σ is the velocity dispersion. Using the above, $M \gg m$, $\rho = nm$, and $\dot{x}_M = v_M$ the Chandrasekhar friction formula now becomes:

$$\frac{d\underline{v}_M}{dt} = -\frac{4\pi G^2 M\ln\Lambda}{v_M^3}\rho\left[\mathrm{erf}(X) - \frac{2X}{\sqrt{\pi}}e^{-X^2}\right]\underline{v}_M \tag{H.22}$$

where $X = v_M/(\sqrt{2}\sigma)$.

Now lets spend a little time thinking about what equation H.22 actually *means*. The main points are the following:

- Particles moving faster than v_M do not contribute to the friction. This statement is only true for isotropic background distributions and is a result of the limit in the integral of equation H.20.
- Notice that the friction falls off as $1/v_M^2$: fast moving bodies experience little friction.
- Particles close in velocity to the heavier object contribute most to the friction. The is most clearly seen from equation H.16. Notice the divergence in the friction force for $\dot{\underline{x}}_m \to \dot{\underline{x}}_M$.
- Particles at radii larger than $b_{\rm min} = \frac{G(M+m)}{V_0^2}$ contribute almost all of the friction. We can prove that this is so by considering the integral in equation H.15:

$$\begin{aligned} I &= \int_0^{b_{\rm max}}\left[1 + \frac{b^2 V_0^4}{G^2(M+m)^2}\right]^{-1} bdb \\ &= \frac{G^2(M+m)^2}{2V_0^4}\ln\left(1+\frac{b_{\rm max}^2 V_0^4}{G^2(M+m)^2}\right) \\ &\simeq \frac{G^2(M+m)^2}{V_0^4}\ln\Lambda \end{aligned} \tag{H.23}$$

where $\Lambda = b_{\rm max}/b_{\rm min}$ using the above definition of $b_{\rm min}$.

Now imagine that there is a minimum impact parameter such that only encounters $b \gg b_{\rm min}$ contribute to the dynamical friction effect. The integral now becomes:

$$
\begin{aligned}
I &\simeq \int_{b_{\min}}^{b_{\max}} \left[1+\frac{b^2V_0^4}{G^2(M+m)^2}\right]^{-1} bdb \\
&\simeq \int_{b_{\min}}^{b_{\max}} \frac{G^2(M+m)^2}{V_0^4}\frac{1}{b}db \\
&= \frac{G^2(M+m)^2}{V_0^4}\ln\Lambda
\end{aligned}
\tag{H.24}
$$

The above proves that if $\Lambda \gg 1$ (as is the case for almost any astronomical system of interest) then there is a minimum impact parameter such that $b \gg b_{\min} = \frac{G(M+m)}{V_0^2}$. Particles closer than this to the heavier infalling body do not significantly contribute to the dynamical friction. This is a very important point. Figure H.3 might give the misleading impression that particles are actually bouncing off the infalling satellite. This is absolutely not correct. It is *long range* slow interactions which contribute most of the friction force.

- Finally, notice that the drag force is proportional to the mass *density* ρ rather than the individual particle masses (provided $M \gg m$). Thus the friction force from dark matter particles of $\sim$ proton mass is identical to the friction force from the same *mass density* of stars.

H.2 Resonance: what Chandrasekhar misses

There is something important which the Chandrasekhar approach misses. Equation H.22 suggests that *outside* of a galaxy where $\rho = 0$ there should be no friction, while if $\rho =$ const. one would expect friction as per normal. Both of these statements are inconsistent with numerical N-body experiments. Indeed, we should perhaps expect the first to be wrong since we have already shown that it is the *long range* interactions that provide most of the friction force.

The above failures occur because we have implicitly assumed that the infalling satellite sees each background particle only once. This is what we see in the schematic diagram of Figure H.3: the heavy particle moves in a straight line. For satellites orbiting in real galaxies, however, the satellite will scatter the same *resonant* particles on each orbit. Tremaine and Weinberg 1984 explore such ideas in more detail and derive a dynamical friction formula for spherical systems. This gives the new and important insight that it is particles that resonate with the infalling satellite that provide almost all of the friction force. Thus, friction does not cease simply because the local $\rho \to 0$.

Given the above it is surprising that Chandrasekhar works at all! However, for most gravitational potentials the infalling satellite sinks faster than resonances can be excited. Thus, to a good approximation, it is always encountering new particles – or a least new resonances. This, combined with the fact that the formula depends only logarithmically on the arbitrary parameters wrapped up inside Λ is why Chandrasekhar usually provides an excellent match to numerical experiments. An important exception is the constant density harmonic potential mentioned above. In this case, numerical experiments find that the friction force is momentarily much stronger than Chandrasekhar predicts, after which there is no observed friction at all! For a solution to this interesting problem, have a look in Read *et al.* 2006. For now, if you want a clue as to what is going on, have a think about which orbital frequencies are permitted in the harmonic potential. If resonant particles drive most of the friction force, what is special about the harmonic potential?

H.3 The dynamical friction timescale and the connection to relaxation

We have stated so far without proof that dynamical friction is more important in the Universe than relaxation because it proceeds faster. We may obtain a rough estimate of the dynamical friction time if we imagine an infalling satellite on a circular orbit. Let us assume it remains on a perfectly circular orbit throughout its infall, and that the background distribution is a spherical isothermal distribution given by:

$$\rho(r) = \frac{v_c^2}{4\pi G r^2} \tag{H.25}$$

where v_c is the circular speed and is a constant. This is a good model for galaxies which show flat rotation curves (see Lectures 1).

Equation H.22 now reduces to (using the fact that for an isothermal sphere $\sigma = v_c/\sqrt{2}$):

$$\begin{aligned} M\frac{dv_M}{dt} = F &= -\frac{4\pi \ln\Lambda G^2 M^2 \rho(r)}{v_c^2}\left[\mathrm{erf}(1) - \frac{2}{\sqrt{\pi}}e^{-1}\right] \\ &= -0.428\frac{\ln\Lambda G M^2}{r^2} \end{aligned} \tag{H.26}$$

The infalling satellite loses specific angular momentum L at a rate:

$$\frac{dL}{dt} = \frac{Fr}{M} \simeq \frac{-0.428GM}{r}\ln\Lambda \tag{H.27}$$

and since its orbit remains circular and $v_c = \mathrm{const.}$, we have that $L = rv_c$ at all times. Substituting this into the the above gives:

$$r\frac{dr}{dt} = -\frac{0.428GM}{v_c}\ln\Lambda \tag{H.28}$$

and solving gives us the dynamical friction timescale:

$$t_{\mathrm{fric}} = \frac{2.64\times 10^{11}}{\ln\Lambda}\left(\frac{r_i}{2\mathrm{kpc}}\right)^2\left(\frac{v_c}{250\mathrm{km/s}}\right)\left(\frac{10^6\mathrm{M}_\odot}{M}\right) \tag{H.29}$$

which recalling that $\ln\Lambda \sim 10$ is typically shorter than a Hubble time for infalling galaxies, star clusters and massive globular clusters. The relaxation time for all but the centre of globular clusters is by contrast many Hubble times (see Lecture 1).

H.4 Wakes

The effect of the scattering is that it puts more particles behind, rather than in front of the satellite (see Figure H.3). This creates a *wake* behind the satellite. It is possible to derive the Chandrasekhar friction formula by considering the gravitational pull of this wake on the satellite (see e.g. Mulder 1983). Hence, this is an alternative way of understanding the friction effect.

H.5 Mass segregation

The dynamical friction effect is of importance inside globular clusters because it causes *mass segregation*: the heavier stars to sink towards the centre, while the lighter stars move out to the edge. This process could be very important in seeding the supermassive black holes observed to reside at the centres of galaxies. We do not understand fully yet how such black holes form. It is not too difficult to grow them from 10^3 to $10^6\mathrm{M}_\odot$ through gas accretion. But black holes are born from the end-phase of the collapse of massive stars; they start out at a mere $40\mathrm{M}_\odot$ at best. Bridging the gap from 40 to $10^3\mathrm{M}_\odot$ is a significant challenge in theoretical astronomy, even today. Dynamical merging of massive stars and black holes at the centre of globular clusters could be one solution. Theoretically, we expect such massive stars and black holes to reside at the centres of globular clusters as a result of mass segregation and, indeed, mass segregation has been observed in some nearby clusters.

H.6 Collisionless relaxation and friction

So far we have discussed relaxation and dynamical friction as *collisional* processes. Yet we have loosely talked also about "dynamical friction on satellite galaxies". If a satellite galaxy mostly comprises dark matter particles then both it and the galaxy it is falling into are undoubtedly *collisionless*. So how is it then than dynamical friction can proceed? We discuss this in detail in the next lecture where we consider galaxy-galaxy interactions.

Bibliography

[Aguirre *et al.*, 2001] A. Aguirre, J. Schaye, and E. Quataert. Problems for Modified Newtonian Dynamics in Clusters and the Lyα Forest? *ApJ*, 561:550–558, November 2001. 53

[Alcock *et al.*, 1993] C. Alcock, C. W. Akerlof, R. A. Allsman, T. S. Axelrod, D. P. Bennett, S. Chan, K. H. Cook, K. C. Freeman, K. Griest, S. L. Marshall, H.-S. Park, S. Perlmutter, B. A. Peterson, M. R. Pratt, P. J. Quinn, A. W. Rodgers, C. W. Stubbs, and W. Sutherland. Possible gravitational microlensing of a star in the Large Magellanic Cloud. *Nature*, 365:621–623, October 1993. 45

[Amorisco and Evans, 2011] N. C. Amorisco and N. W. Evans. Dark matter cores and cusps: the case of multiple stellar populations in dwarf spheroidals. *MNRAS*, page 1606, October 2011. 111

[Anderson and Bregman, 2010] M. E. Anderson and J. N. Bregman. Do Hot Halos Around Galaxies Contain the Missing Baryons? *ApJ*, 714:320–331, May 2010. 42

[Angus *et al.*, 2006] G. W. Angus, B. Famaey, and H. S. Zhao. Can MOND take a bullet? Analytical comparisons of three versions of MOND beyond spherical symmetry. *MNRAS*, 371:138–146, September 2006. 54, 55

[Angus *et al.*, 2007] G. W. Angus, H. Y. Shan, H. S. Zhao, and B. Famaey. On the Proof of Dark Matter, the Law of Gravity, and the Mass of Neutrinos. *ApJ*, 654:L13–L16, January 2007. 53

[Arfken and Weber, 2005] G. B. Arfken and H. J. Weber. Mathematical methods for physicists 6th ed. *Materials and Manufacturing Processes*, 2005. 113, 114, 116, 117, 118

[Aubourg *et al.*, 1993] E. Aubourg, P. Bareyre, S. Bréhin, M. Gros, M. Lachièze-Rey, B. Laurent, E. Lesquoy, C. Magneville, A. Milsztajn, L. Moscoso, F. Queinnec, J. Rich, M. Spiro, L. Vigroux, S. Zylberajch, R. Ansari, F. Cavalier, M. Moniez, J.-P. Beaulieu, R. Ferlet, P. Grison, A. Vidal-Madjar, J. Guibert, O. Moreau, F. Tajahmady, E. Maurice, L. Prévôt, and C. Gry. Evidence for gravitational microlensing by dark objects in the Galactic halo. *Nature*, 365:623–625, October 1993. 45

[Avila-Reese *et al.*, 2001] V. Avila-Reese, P. Colín, O. Valenzuela, E. D'Onghia, and C. Firmani. Formation and Structure of Halos in a Warm Dark Matter Cosmology. *ApJ*, 559:516–530, October 2001. 91

[Babcock, 1939] H. W. Babcock. The rotation of the Andromeda Nebula. *Lick Observatory Bulletin*, 19:41–51, 1939. 18

[Bardeen, 1980] J. M. Bardeen. Gauge-invariant cosmological perturbations. *Phys. Rev. D*, 22:1882–1905, October 1980. 71

[Barnes and Hut, 1986] J. Barnes and P. Hut. A Hierarchical O(NlogN) Force-Calculation Algorithm. *Nature*, 324:446–449, December 1986. 86, 87

[Battaglia *et al.*, 2008] G. Battaglia, A. Helmi, E. Tolstoy, M. Irwin, V. Hill, and P. Jablonka. The Kinematic Status and Mass Content of the Sculptor Dwarf Spheroidal Galaxy. *ApJ*, 681:L13–L16, July 2008. 111

[Bekenstein and Milgrom, 1984] J. Bekenstein and M. Milgrom. Does the missing mass problem signal the breakdown of Newtonian gravity? *ApJ*, 286:7–14, November 1984. 48

[Bekenstein, 2004] J. D. Bekenstein. Relativistic gravitation theory for the modified Newtonian dynamics paradigm. *Phys. Rev. D*, 70(8):083509–+, October 2004. 48, 49, 50

[Binney and Evans, 2001] J. J. Binney and N. W. Evans. Cuspy dark matter haloes and the Galaxy. *MNRAS*, 327:L27–L31, October 2001. 97

[Binney and Merrifield, 1998] J. Binney and M. Merrifield. *Galactic astronomy*. Galactic astronomy / James Binney and Michael Merrifield. Princeton, NJ : Princeton University Press, 1998. (Princeton series in astrophysics) QB857 .B522 1998 ($35.00), 1998. 40

[Binney and Tremaine, 2008] J. Binney and S. Tremaine. *Galactic dynamics*. Princeton, NJ, Princeton University Press, 2008, 747 p., 2008. 1, 86, 110, 137

[Blumenthal *et al.*, 1986] G. R. Blumenthal, S. M. Faber, R. Flores, and J. R. Primack. Contraction of dark matter galactic halos due to baryonic infall. *ApJ*, 301:27–34, February 1986. 96

[Bode *et al.*, 2001] P. Bode, J. P. Ostriker, and N. Turok. Halo Formation in Warm Dark Matter Models. *ApJ*, 556:93–107, July 2001. 90, 92

[Bond *et al.*, 1991] J. R. Bond, S. Cole, G. Efstathiou, and N. Kaiser. Excursion set mass functions for hierarchical Gaussian fluctuations. *ApJ*, 379:440–460, October 1991. 83

[Bosma and van der Kruit, 1979] A. Bosma and P. C. van der Kruit. The local mass-to-light ratio in spiral galaxies. *A&A*, 79:281–286, November 1979. 18

[Boyarsky *et al.*, 2009a] A. Boyarsky, J. Lesgourgues, O. Ruchayskiy, and M. Viel. Lyman-α constraints on warm and on warm-plus-cold dark matter models. *JCAP*, 5:12, May 2009. 101

[Boyarsky *et al.*, 2009b] A. Boyarsky, O. Ruchayskiy, and M. Shaposhnikov. The Role of Sterile Neutrinos in Cosmology and Astrophysics. *Annual Review of Nuclear and Particle Science*, 59:191–214, November 2009. 90

[Brownstein and Moffat, 2007] J. R. Brownstein and J. W. Moffat. The Bullet Cluster 1E0657-558 evidence shows modified gravity in the absence of dark matter. *MNRAS*, 382:29–47, November 2007. 55

[Bruderer *et al.*, 2016] C. Bruderer, J. I. Read, J. P. Coles, D. Leier, E. E. Falco, I. Ferreras, and P. Saha. Light versus dark in strong-lens galaxies: dark matter haloes that are rounder than their stars. *MNRAS*, 456:870–884, February 2016. 103

[Buchert, 2011] T. Buchert. Toward physical cosmology: focus on inhomogeneous geometry and its non-perturbative effects. *Classical and Quantum Gravity*, 28(16):164007, August 2011. 100

[Chandrasekhar, 1943] S. Chandrasekhar. Dynamical Friction. I. General Considerations: the Coefficient of Dynamical Friction. *ApJ*, 97:255–+, March 1943. 133

[Chin and Chen, 2005] S. Chin and C. Chen. Forward symplectic integrators for solving gravitational few-body problems. *Celestial Mechanics and Dynamical Astronomy*, 91:301–322, 2005. 10.1007/s10569-004-4622-z. 89

[Chwolson, 1924] O. Chwolson. Über eine mögliche Form fiktiver Doppelsterne. *Astronomische Nachrichten*, 221:329–+, June 1924. 31

[Clowe *et al.*, 2006] D. Clowe, M. Bradač, A. H. Gonzalez, M. Markevitch, S. W. Randall, C. Jones, and D. Zaritsky. A Direct Empirical Proof of the Existence of Dark Matter. *ApJ*, 648:L109–L113, September 2006. 54

[Cole *et al.*, 2011] D. R. Cole, W. Dehnen, and M. I. Wilkinson. Weakening dark matter cusps by clumpy baryonic infall. *MNRAS*, 416:1118–1134, September 2011. 97

[Cole *et al.*, 2012] D. R. Cole, W. Dehnen, J. I. Read, and M. I. Wilkinson. The mass distribution of the Fornax dSph: constraints from its globular cluster distribution. *MNRAS*, 426:601–613, October 2012. 111

[Combes, 1991] F. Combes. Distribution of CO in the Milky Way. *ARA&A*, 29:195–237, 1991. 42

[Dehnen and Read, 2011] W. Dehnen and J. I. Read. N-body simulations of gravitational dynamics. *European Physical Journal Plus*, 126:55–+, May 2011. 83, 84, 87, 88, 89

[Dehnen, 2000] W. Dehnen. A Very Fast and Momentum-conserving Tree Code. *ApJ*, 536:L39–L42, June 2000. 87

[Dehnen, 2001] W. Dehnen. Towards optimal softening in three-dimensional N-body codes - I. Minimizing the force error. *MNRAS*, 324:273–291, June 2001. 84, 86

[Dicke *et al.*, 1965] R. H. Dicke, P. J. E. Peebles, P. G. Roll, and D. T. Wilkinson. Cosmic Black-Body Radiation. *ApJ*, 142:414–419, July 1965. 73

[Dodelson, 2011] S. Dodelson. The Real Problem with MOND. *International Journal of Modern Physics D*, 20:2749–2753, 2011. 55, 77

[Dong *et al.*, 2007] S. Dong, A. Udalski, A. Gould, W. T. Reach, G. W. Christie, A. F. Boden, D. P. Bennett, G. Fazio, K. Griest, M. K. Szymański, M. Kubiak, I. Soszyński, G. Pietrzyński, O. Szewczyk, Ł. Wyrzykowski, K. Ulaczyk, T. Wieckowski, B. Paczyński, D. L. DePoy, R. W. Pogge, G. W. Preston, I. B. Thompson, and B. M. Patten. First Space-Based Microlens Parallax Measurement: Spitzer Observations of OGLE-2005-SMC-001. *ApJ*, 664:862–878, August 2007. 45

[Dubinski and Carlberg, 1991] J. Dubinski and R. G. Carlberg. The structure of cold dark matter halos. *ApJ*, 378:496–503, September 1991. 92, 93

[Einstein, 1916] A. Einstein. Die Grundlage der allgemeinen Relativitätstheorie. *Annalen der Physik*, 354:769–822, 1916. 20

[Einstein, 1936] A. Einstein. Lens-Like Action of a Star by the Deviation of Light in the Gravitational Field. *Science*, 84:506–507, December 1936. 31

[Eisenstein and Hu, 1999] D. J. Eisenstein and W. Hu. Power Spectra for Cold Dark Matter and Its Variants. *ApJ*, 511:5–15, January 1999. 83

[El-Zant *et al.*, 2001] A. El-Zant, I. Shlosman, and Y. Hoffman. Dark Halos: The Flattening of the Density Cusp by Dynamical Friction. *ApJ*, 560:636–643, October 2001. 97

[Ewald, 1921] P. P. Ewald. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik*, 369:253–287, 1921. 87

[Ferrière, 2001] K. M. Ferrière. The interstellar environment of our galaxy. *Reviews of Modern Physics*, 73:1031–1066, October 2001. 40, 42

[Finkbeiner, 2003] D. P. Finkbeiner. A Full-Sky Hα Template for Microwave Foreground Prediction. *ApJS*, 146:407–415, June 2003. 41

[Flores and Primack, 1994] R. A. Flores and J. R. Primack. Observational and theoretical constraints on singular dark matter halos. *ApJ*, 427:L1–L4, May 1994. 106

[Flynn *et al.*, 1996] C. Flynn, A. Gould, and J. N. Bahcall. Hubble Deep Field Constraint on Baryonic Dark Matter. *ApJ*, 466:L55+, August 1996. 38, 40

[Freeman, 1970] K. C. Freeman. On the Disks of Spiral and so Galaxies. *ApJ*, 160:811–+, June 1970. 18

[Fukugita *et al.*, 1995] M. Fukugita, K. Shimasaku, and T. Ichikawa. Galaxy Colors in Various Photometric Band Systems. *PASP*, 107:945–+, October 1995. 6

[Garbari *et al.*, 2011] S. Garbari, J. I. Read, and G. Lake. Limits on the local dark matter density. *MNRAS*, 416:2318–2340, September 2011. 109

[Gatto *et al.*, 2013] A. Gatto, F. Fraternali, J. I. Read, F. Marinacci, H. Lux, and S. Walch. Unveiling the corona of the Milky Way via ram-pressure stripping of dwarf satellites. *MNRAS*, 433:2749–2763, August 2013. 107

[Giodini *et al.*, 2009] S. Giodini, D. Pierini, A. Finoguenov, G. W. Pratt, H. Boehringer, A. Leauthaud, L. Guzzo, H. Aussel, M. Bolzonella, P. Capak, M. Elvis, G. Hasinger, 14 other authors, and the COSMOS Collaboration. Stellar and Total Baryon Mass Fractions in Groups and Clusters Since Redshift 1. *ApJ*, 703:982–993, September 2009. 43

[Gnedin and Zhao, 2002] O. Y. Gnedin and H. Zhao. Maximum feedback and dark matter profiles of dwarf galaxies. *MNRAS*, 333:299–306, June 2002. 98

[Goerdt *et al.*, 2006] T. Goerdt, B. Moore, J. I. Read, J. Stadel, and M. Zemp. Does the Fornax dwarf spheroidal have a central cusp or core? *MNRAS*, 368:1073–1077, May 2006. 111

[Goerdt *et al.*, 2010] T. Goerdt, B. Moore, J. I. Read, and J. Stadel. Core Creation in Galaxies and Halos Via Sinking Massive Objects. *ApJ*, 725:1707–1716, December 2010. 97

[Governato *et al.*, 2010] F. Governato, C. Brook, L. Mayer, A. Brooks, G. Rhee, J. Wadsley, P. Jonsson, B. Willman, G. Stinson, T. Quinn, and P. Madau. Bulgeless dwarf galaxies and dark matter cores from supernova-driven outflows. *Nature*, 463:203–206, January 2010. 99

[Green and Wald, 2011] S. R. Green and R. M. Wald. New framework for analyzing the effects of small scale inhomogeneities in cosmology. *Phys. Rev. D*, 83(8):084020, April 2011. 100

[Hahn *et al.*, 2013] O. Hahn, T. Abel, and R. Kaehler. A new approach to simulating collisionless dark matter fluids. *MNRAS*, 434:1171–1191, September 2013. 92

[Harrison, 1989] E. Harrison. *Darkness at Night: a riddle of the Universe*. Harvard University Press, 1989. 57

[Harvey *et al.*, 2015] D. Harvey, R. Massey, T. Kitching, A. Taylor, and E. Tittley. The nongravitational interactions of dark matter in colliding galaxy clusters. *Science*, 347:1462–1465, March 2015. 104

[Hawking, 1971] S. Hawking. Gravitationally collapsed objects of very low mass. *MNRAS*, 152:75–+, 1971. 43

[Hernquist *et al.*, 1991] L. Hernquist, F. R. Bouchet, and Y. Suto. Application of the Ewald method to cosmological N-body simulations. *ApJS*, 75:231–240, February 1991. 87

[Hernquist, 1990] L. Hernquist. An analytical model for spherical galaxies and bulges. *ApJ*, 356:359–364, June 1990. 93

[Hills, 1980] J. G. Hills. The effect of mass loss on the dynamical evolution of a stellar system - Analytic approximations. *ApJ*, 235:986–991, February 1980. 97

[Hobbs *et al.*, 2016] A. Hobbs, J. I. Read, O. Agertz, F. Iannuzzi, and C. Power. NOVel Adaptive softening for collisionless N-body simulations: eliminating spurious haloes. *MNRAS*, 458:468–479, May 2016. 92

[Hogg, 1999] D. W. Hogg. Distance measures in cosmology. *ArXiv Astrophysics e-prints*, May 1999. 65

[Hu and Dodelson, 2002] W. Hu and S. Dodelson. Cosmic Microwave Background Anisotropies. *ARA&A*, 40:171–216, 2002. 73, 74

[Hubble, 1929] E. Hubble. A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. *Proceedings of the National Academy of Science*, 15:168–173, March 1929. 56

[Hubble, 1936] E. P. Hubble. *Realm of the Nebulae*. 1936. 18

[Ibata *et al.*, 2011] R. Ibata, A. Sollima, C. Nipoti, M. Bellazzini, S. C. Chapman, and E. Dalessandro. The Globular Cluster NGC 2419: A Crucible for Theories of Gravity. *ApJ*, 738:186–+, September 2011. 53

[Kalberla and Kerp, 2009] P. M. W. Kalberla and J. Kerp. The Hi Distribution of the Milky Way. *ARA&A*, 47:27–61, September 2009. 40

[Kalberla *et al.*, 1999] P. M. W. Kalberla, Y. A. Shchekinov, and R.-J. Dettmar. H_2 dark matter in the galactic halo from EGRET. *A&A*, 350:L9–L12, October 1999. 42

[Klessen, 1997] R. Klessen. GRAPESPH with fully periodic boundary conditions - Fragmentation of molecular clouds. *MNRAS*, 292:11, November 1997. 87

[Klypin *et al.*, 2002] A. Klypin, H. Zhao, and R. S. Somerville. ΛCDM-based Models for the Milky Way and M31. I. Dynamical Models. *ApJ*, 573:597–613, July 2002. 38

[Klypin *et al.*, 2011] A. A. Klypin, S. Trujillo-Gomez, and J. Primack. Dark Matter Halos in the Standard Cosmological Model: Results from the Bolshoi Simulation. *ApJ*, 740:102, October 2011.

[Kuzio de Naray and Kaufmann, 2011] R. Kuzio de Naray and T. Kaufmann. Recovering cores and cusps in dark matter haloes using mock velocity field observations. *MNRAS*, 414:3617–3626, July 2011. 105

[Lake, 1989] G. Lake. Testing modifications of gravity. *ApJ*, 345:L17–L19, October 1989. 53

[Lewis and Bridle, 2002] A. Lewis and S. Bridle. Cosmological parameters from CMB and other data: A Monte Carlo approach. *Phys. Rev. D*, 66(10):103511, November 2002. 76

[Limousin *et al.*, 2008] M. Limousin, J. Richard, J.-P. Kneib, H. Brink, R. Pelló, E. Jullo, H. Tu, J. Sommer-Larsen, E. Egami, M. J. Michałowski, R. Cabanac, and D. P. Stark. Strong lensing in Abell 1703: constraints on the slope of the inner dark matter distribution. *A&A*, 489:23–35, October 2008. 37

[Łokas, 2009] E. L. Łokas. The mass and velocity anisotropy of the Carina, Fornax, Sculptor and Sextans dwarf spheroidal galaxies. *MNRAS*, 394:L102–L106, March 2009. 110

[Lynden-Bell, 1969] D. Lynden-Bell. Galactic Nuclei as Collapsed Old Quasars. *Nature*, 223:690–694, August 1969. 101

[Macciò *et al.*, 2012] A. V. Macciò, S. Paduroiu, D. Anderhalden, A. Schneider, and B. Moore. Cores in warm dark matter haloes: a Catch 22 problem. *MNRAS*, 424:1105–1112, August 2012. 91

[Marasco and Fraternali, 2011] A. Marasco and F. Fraternali. Modelling the H I halo of the Milky Way. *A&A*, 525:A134+, January 2011. 40

[Mashchenko *et al.*, 2008] S. Mashchenko, J. Wadsley, and H. M. P. Couchman. Stellar Feedback in Dwarf Galaxy Formation. *Science*, 319:174–, January 2008. 99

[Mateo, 1998] M. L. Mateo. Dwarf Galaxies of the Local Group. *ARA&A*, 36:435–506, 1998. 98

[Merritt *et al.*, 2006] D. Merritt, A. W. Graham, B. Moore, J. Diemand, and B. Terzić. Empirical Models for Dark Matter Halos. I. Nonparametric Construction of Density Profiles and Comparison with Parametric Models. *AJ*, 132:2685–2700, December 2006. 93

[Michelson and Morley, 1887] A. Michelson and E. Morley. On the Relative Motion of the Earth and the Luminiferous Ether. *American Journal of Science*, 34:333–345, 1887. 20

[Michelson, 1881] A. Michelson. The Relative Motion of the Earth and the Luminiferous Ether. *American Journal of Science*, 22:120–129, 1881. 20

[Milgrom, 1983] M. Milgrom. A modification of the Newtonian dynamics as a possible alternative to the hidden mass hypothesis. *ApJ*, 270:365–370, July 1983. 48, 52

[Moffat, 2005] J. W. Moffat. Gravitational theory, galaxy rotation curves and cosmology without dark matter. *JCAP*, 5:3–+, May 2005. 48

[Moffat, 2006] J. W. Moffat. Scalar tensor vector gravity theory. *JCAP*, 3:4–+, March 2006. 48, 53

[Moore, 1994] B. Moore. Evidence against dissipation-less dark matter from observations of galaxy haloes. *Nature*, 370:629–631, August 1994. 106

[Mulder, 1983] W. A. Mulder. Dynamical friction on extended objects. *A&A*, 117:9–16, January 1983. 139

[Mutka and Mähönen, 2002] P. T. Mutka and P. H. Mähönen. Approximation of Light-Ray Deflection Angle and Gravitational Lenses in the Schwarzschild Metric. I. Derivation and Quasar Lens. *ApJ*, 576:107–112, September 2002. 35

[Natarajan and Zhao, 2008] P. Natarajan and H. Zhao. MOND plus classical neutrinos are not enough for cluster lensing. *MNRAS*, 389:250–256, September 2008. 53

[Navarro *et al.*, 1996a] J. F. Navarro, V. R. Eke, and C. S. Frenk. The cores of dwarf galaxy haloes. *MNRAS*, 283:L72–L78, December 1996. 97, 98

[Navarro *et al.*, 1996b] J. F. Navarro, C. S. Frenk, and S. D. M. White. The Structure of Cold Dark Matter Halos. *ApJ*, 462:563–+, May 1996. 93

[Noh and Hwang, 2006] H. Noh and J.-C. Hwang. Newtonian versus relativistic nonlinear cosmology. *General Relativity and Gravitation*, 38:703–710, May 2006. 99

[Nussbaumer and Bieri, 2011] H. Nussbaumer and L. Bieri. Who discovered the expanding universe? *ArXiv e-prints*, July 2011. 27, 56

[Oñorbe *et al.*, 2015] J. Oñorbe, M. Boylan-Kolchin, J. S. Bullock, P. F. Hopkins, D. Kerěs, C.-A. Faucher-Giguère, E. Quataert, and N. Murray. Forged in FIRE: cusps, cores, and baryons in low-mass dwarf galaxies. *ArXiv e-prints*, February 2015. 99

[Paczynski, 1986] B. Paczynski. Gravitational microlensing by the galactic halo. *ApJ*, 304:1–5, May 1986. 45

[Peacock, 1999] J. A. Peacock. *Cosmological physics*. Cosmological physics. Publisher: Cambridge, UK: Cambridge University Press, 1999. ISBN: 0521422701, 1999. 1, 57, 72

[Peebles, 1980] P. J. E. Peebles. *The large-scale structure of the universe*. Research supported by the National Science Foundation. Princeton, N.J., Princeton University Press, 1980. 435 p., 1980. 1, 71

[Penzias and Wilson, 1965] A. A. Penzias and R. W. Wilson. A Measurement of Excess Antenna Temperature at 4080 Mc/s. *ApJ*, 142:419–421, July 1965. 73

[Perlmutter *et al.*, 1999] S. Perlmutter, G. Aldering, G. Goldhaber, R. A. Knop, P. Nugent, P. G. Castro, S. Deustua, S. Fabbro, and The Supernova Cosmology Project. Measurements of Omega and Lambda from 42 High-Redshift Supernovae. *ApJ*, 517:565–586, June 1999. 63

[Pfenniger *et al.*, 1994] D. Pfenniger, F. Combes, and L. Martinet. Is dark matter in spiral galaxies cold gas? I. Observational constraints and dynamical clues about galaxy evolution. *A&A*, 285:79–93, May 1994. 42

[Phillips, 1999] A. C. Phillips. The Physics of Stars, 2nd Edition. *Physica Scripta Volume T*, July 1999. 10, 40, 97

[Planck Collaboration *et al.*, 2013] Planck Collaboration, P. A. R. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, J. Aumont, C. Baccigalupi, A. J. Banday, and et al. Planck 2013 results. XVI. Cosmological parameters. *ArXiv e-prints*, March 2013. 74, 76

[Pontzen and Governato, 2012] A. Pontzen and F. Governato. How supernova feedback turns dark matter cusps into cores. *MNRAS*, 421:3464–3471, April 2012. 99

[Pontzen and Governato, 2014] A. Pontzen and F. Governato. Cold dark matter heats up. *Nature*, 506:171–178, February 2014. 97

[Pontzen *et al.*, 2015] A. Pontzen, J. Read, R. Teyssier, F. Governato, A. Gualandris, N. Roth, and J. Devriendt. Milking the spherical cow: on aspherical dynamics in spherical coordinates. *ArXiv e-prints*, February 2015. 99

[Power *et al.*, 2003] C. Power, J. F. Navarro, A. Jenkins, C. S. Frenk, S. D. M. White, V. Springel, J. Stadel, and T. Quinn. The inner structure of ΛCDM haloes - I. A numerical convergence study. *MNRAS*, 338:14–34, January 2003. 84

[Press and Schechter, 1974] W. H. Press and P. Schechter. Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. *ApJ*, 187:425–438, February 1974. 83

[Press *et al.*, 1992] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C. The art of scientific computing.* Cambridge: University Press, —c1992, 2nd ed., 1992. 85, 118

[Read and Gilmore, 2005] J. I. Read and G. Gilmore. Mass loss from dwarf spheroidal galaxies: the origins of shallow dark matter cores and exponential surface brightness profiles. *MNRAS*, 356:107–124, January 2005. 97, 98, 99

[Read and Moore, 2005] J. I. Read and B. Moore. Tidal streams in a MOND potential: constraints from Sagittarius. *MNRAS*, 361:971–976, August 2005. 53

[Read and Trentham, 2005] J. I. Read and N. Trentham. The baryonic mass function of galaxies. *Royal Society of London Philosophical Transactions Series A*, 363:2693–+, 2005. 42

[Read *et al.*, 2006] J. I. Read, T. Goerdt, B. Moore, A. P. Pontzen, J. Stadel, and G. Lake. Dynamical friction in constant density cores: a failure of the Chandrasekhar formula. *MNRAS*, 373:1451–1460, December 2006. 138

[Read *et al.*, 2016a] J. I. Read, O. Agertz, and M. L. M. Collins. Dark matter cores all the way down. *MNRAS*, 459:2573–2590, July 2016. 99, 106, 107

[Read *et al.*, 2016b] J. I. Read, G. Iorio, O. Agertz, and F. Fraternali. The stellar mass-halo mass relation of isolated field dwarfs: a critical test of ΛCDM at the edge of galaxy formation. *ArXiv e-prints*, July 2016. 106, 107

[Read *et al.*, 2016c] J. I. Read, G. Iorio, O. Agertz, and F. Fraternali. Understanding the shape and diversity of dwarf galaxy rotation curves in LCDM. *ArXiv e-prints: 1601.05821*, September 2016. 105, 107

[Refsdal, 1964] S. Refsdal. The gravitational lens effect. *MNRAS*, 128:295–+, 1964. 31

[Refsdal, 1966] S. Refsdal. On the possibility of determining the distances and masses of stars from the gravitational lens effect. *MNRAS*, 134:315–+, 1966. 45

[Richer *et al.*, 2006] H. B. Richer, J. Anderson, J. Brewer, S. Davis, G. G. Fahlman, B. M. S. Hansen, J. Hurley, J. S. Kalirai, I. R. King, D. Reitzel, R. M. Rich, M. M. Shara, and P. B. Stetson. Probing the Faintest Stars in a Globular Star Cluster. *Science*, 313:936–940, August 2006. 40

[Rubin *et al.*, 1980] V. C. Rubin, W. K. J. Ford, and N. . Thonnard. Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 /R = 4kpc/ to UGC 2885 /R = 122 kpc/. *ApJ*, 238:471–487, June 1980. 18

[Saha and Read, 2009] P. Saha and J. I. Read. The Cluster Lens ACO 1703: Redshift Contrast and the Inner Profile. *ApJ*, 690:154–162, January 2009. 102

[Saha *et al.*, 2006] P. Saha, J. I. Read, and L. L. R. Williams. Two Strong-Lensing Clusters Confront Universal Dark Matter Profiles. *ApJ*, 652:L5–L8, November 2006. 103

[Sanders, 2003] R. H. Sanders. Clusters of galaxies with modified Newtonian dynamics. *MNRAS*, 342:901–908, July 2003. 53

[Schwarzschild, 1916] K. Schwarzschild. On the Gravitational Field of a Mass Point According to Einstein's Theory. *Abh. Konigl. Preuss. Akad. Wissenschaften Jahre 1906,92, Berlin,1907*, pages 189–196, 1916. 29

[Seljak and Zaldarriaga, 1996] U. Seljak and M. Zaldarriaga. A Line-of-Sight Integration Approach to Cosmic Microwave Background Anisotropies. *ApJ*, 469:437, October 1996. 71

[Seljak *et al.*, 2003] U. Seljak, N. Sugiyama, M. White, and M. Zaldarriaga. Comparison of cosmological Boltzmann codes: Are we ready for high precision cosmology? *Phys. Rev. D*, 68(8):083507, October 2003. 71

[Seljak *et al.*, 2006a] U. Seljak, A. Makarov, P. McDonald, and H. Trac. Can Sterile Neutrinos Be the Dark Matter? *Physical Review Letters*, 97(19):191303–+, November 2006. 101

[Seljak *et al.*, 2006b] U. Seljak, A. Slosar, and P. McDonald. Cosmological parameters from combining the Lyman-α forest with CMB, galaxy clustering and SN constraints. *Journal of Cosmology and Astro-Particle Physics*, 10:14–+, October 2006. 76

[Sheng, 1989] Q. Sheng. Solving linear partial differential equations by exponential splitting. *IMA J. Numerical Analysis*, 9(2):199–212, 1989. 89

[Shu, 1982] F. H. Shu. *The physical universe. an introduction to astronomy.* A Series of Books in Astronomy, Mill Valley, CA: University Science Books, 1982, 1982. 1

[Silk, 1968] J. Silk. Cosmic Black-Body Radiation and Galaxy Formation. *ApJ*, 151:459–+, February 1968. 72

[Simon *et al.*, 2003] J. D. Simon, A. D. Bolatto, A. Leroy, and L. Blitz. High-Resolution Measurements of the Dark Matter Halo of NGC 2976: Evidence for a Shallow Density Profile. *ApJ*, 596:957–981, October 2003. 11

[Skordis *et al.*, 2006] C. Skordis, D. F. Mota, P. G. Ferreira, and C. Bœhm. Large Scale Structure in Bekenstein's Theory of Relativistic Modified Newtonian Dynamics. *Physical Review Letters*, 96(1):011301, January 2006. 55, 77, 78

[Smoot *et al.*, 1992] G. F. Smoot, C. L. Bennett, A. Kogut, E. L. Wright, J. Aymon, N. W. Boggess, E. S. Cheng, G. de Amici, S. Gulkis, M. G. Hauser, G. Hinshaw, P. D. Jackson, M. Janssen, E. Kaita, T. Kelsall, P. Keegstra, and 12 other authors. Structure in the COBE differential microwave radiometer first-year maps. *ApJ*, 396:L1–L5, September 1992. 73

[Spergel *et al.*, 2007] D. N. Spergel, R. Bean, O. Doré, M. R. Nolta, C. L. Bennett, J. Dunkley, G. Hinshaw, N. Jarosik, E. Komatsu, L. Page, H. V. Peiris, L. Verde, M. Halpern, and nine other authors. Three-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Implications for Cosmology. *ApJS*, 170:377–408, June 2007. 76

[Springel *et al.*, 2001] V. Springel, N. Yoshida, and S. D. M. White. GADGET: a code for collisionless and gasdynamical cosmological simulations. *New Astronomy*, 6:79–117, April 2001. 87

[Stadel *et al.*, 2009] J. Stadel, D. Potter, B. Moore, J. Diemand, P. Madau, M. Zemp, M. Kuhlen, and V. Quilis. Quantifying the heart of darkness with GHALO - a multibillion particle simulation of a galactic halo. *MNRAS*, 398:L21–L25, September 2009. 93

[Stadel, 2001] J. G. Stadel. Cosmological N-body simulations and their analysis. *Ph.D. Thesis, Univ. Washington*, 2001. 86

[Strigari *et al.*, 2007] L. E. Strigari, M. Kaplinghat, and J. S. Bullock. Dark matter halos with cores from hierarchical structure formation. *Phys. Rev. D*, 75(6):061303, March 2007. 91

[Sullivan *et al.*, 2011] M. Sullivan, J. Guy, A. Conley, N. Regnault, P. Astier, C. Balland, S. Basa, R. G. Carlberg, D. Fouchez, D. Hardin, I. M. Hook, D. A. Howell, R. Pain, N. Palanque-Delabrouille, K. M. Perrett, C. J. Pritchet, J. Rich, V. Ruhlmann-Kleider, D. Balam, S. Baumont, R. S. Ellis, S. Fabbro, H. K. Fakhouri, N. Fourmanoit, S. González-Gaitán, M. L. Graham, M. J. Hudson, E. Hsiao, T. Kronborg, C. Lidman, A. M. Mourao, J. D. Neill, S. Perlmutter, P. Ripoche, N. Suzuki, and E. S. Walker. SNLS3: Constraints on Dark Energy Combining the Supernova Legacy Survey Three-year Data with Other Probes. *ApJ*, 737:102, August 2011. 76

[Sutherland and Dopita, 1993] R. S. Sutherland and M. A. Dopita. Cooling functions for low-density astrophysical plasmas. *ApJS*, 88:253–327, September 1993. 41

[Suzuki, 1991] M. Suzuki. General theory of fractal path integrals with applications to many-body theories and statistical physics. *J. Math. Phys.*, 32(2):400–407, 1991. 89

[Swaters *et al.*, 2010] R. A. Swaters, R. H. Sanders, and S. S. McGaugh. Testing Modified Newtonian Dynamics with Rotation Curves of Dwarf and Low Surface Brightness Galaxies. *ApJ*, 718:380–391, July 2010. 53

[Teyssier *et al.*, 2013] R. Teyssier, A. Pontzen, Y. Dubois, and J. I. Read. Cusp-core transformations in dwarf galaxies: observational predictions. *MNRAS*, 429:3068–3078, March 2013. 99

[Tisserand *et al.*, 2007] P. Tisserand, L. Le Guillou, C. Afonso, J. N. Albert, J. Andersen, R. Ansari, É. Aubourg, P. Bareyre, J. P. Beaulieu, X. Charlot, C. Coutures, R. Ferlet, and and The EROS-2 Collaboration. Limits on the Macho content of the Galactic Halo from the EROS-2 Survey of the Magellanic Clouds. *A&A*, 469:387–404, July 2007. 45

[Tremaine and Gunn, 1979] S. Tremaine and J. E. Gunn. Dynamical role of light neutral leptons in cosmology. *Physical Review Letters*, 42:407–410, February 1979. 91

[Tremaine and Weinberg, 1984] S. Tremaine and M. D. Weinberg. Dynamical friction in spherical systems. *MNRAS*, 209:729–757, August 1984. 138

[Udalski *et al.*, 1992] A. Udalski, M. Szymanski, J. Kaluzny, M. Kubiak, and M. Mateo. The Optical Gravitational Lensing Experiment. *AcA*, 42:253–284, 1992. 45

[van Albada *et al.*, 1985] T. S. van Albada, J. N. Bahcall, K. Begeman, and R. Sancisi. Distribution of dark matter in the spiral galaxy NGC 3198. *ApJ*, 295:305–313, August 1985. 18

[Viel *et al.*, 2008] M. Viel, G. D. Becker, J. S. Bolton, M. G. Haehnelt, M. Rauch, and W. L. W. Sargent. How Cold Is Cold Dark Matter? Small-Scales Constraints from the Flux Power Spectrum of the High-Redshift Lyman-α Forest. *Physical Review Letters*, 100(4):041304–+, February 2008. 101

[Volders, 1959] L. M. J. S. Volders. Neutral hydrogen in M 33 and M 101. *Bull. Astron. Inst. Netherlands*, 14:323, September 1959. 18

[Walker and Peñarrubia, 2011] M. G. Walker and J. Peñarrubia. A Method for Measuring (Slopes of) the Mass Profiles of Dwarf Spheroidal Galaxies. *ApJ*, 742:20, November 2011. 110, 111

[Walsh *et al.*, 1979] D. Walsh, R. F. Carswell, and R. J. Weymann. 0957 + 561 A, B - Twin quasistellar objects or gravitational lens. *Nature*, 279:381–384, May 1979. 31

[Wang and White, 2007] J. Wang and S. D. M. White. Discreteness effects in simulations of hot/warm dark matter. *MNRAS*, 380:93–103, September 2007. 92

[Weinberg, 2008] S. Weinberg. *Cosmology*. Oxford University Press, 2008. 1

[Will, 1993] C. M. Will. *Theory and Experiment in Gravitational Physics*. Cambridge University Press, March 1993. 20

[Wu *et al.*, 1999] K. K. S. Wu, O. Lahav, and M. J. Rees. The large-scale smoothness of the Universe. *Nature*, 397:225–230, January 1999. 30

[Wyrzykowski *et al.*, 2011] L. Wyrzykowski, J. Skowron, S. Kozłowski, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, I. Soszyński, O. Szewczyk, K. Ulaczyk, R. Poleski, and P. Tisserand. The OGLE view of microlensing towards the Magellanic Clouds - IV. OGLE-III SMC data and final conclusions on MACHOs. *MNRAS*, 416:2949–2961, October 2011. 45

[Yadav *et al.*, 2005] J. Yadav, S. Bharadwaj, B. Pandey, and T. R. Seshadri. Testing homogeneity on large scales in the Sloan Digital Sky Survey Data Release One. *MNRAS*, 364:601–606, December 2005. 30, 56

[Yoshida, 1993] H. Yoshida. Recent Progress in the Theory and Application of Symplectic Integrators. *Celestial Mechanics and Dynamical Astronomy*, 56:27–43, March 1993. 88, 89

[Young, 1980] P. Young. Numerical models of star clusters with a central black hole. I - Adiabatic models. *ApJ*, 242:1232–1237, December 1980. 96

[Zel'Dovich, 1970] Y. B. Zel'Dovich. Gravitational instability: An approximate theory for large density perturbations. *A&A*, 5:84–89, March 1970. 80

[Zentner, 2007] A. R. Zentner. The Excursion Set Theory of Halo Mass Functions, Halo Clustering, and Halo Growth. *International Journal of Modern Physics D*, 16:763–815, 2007. 83

[Zhao *et al.*, 2006] H. Zhao, D. J. Bacon, A. N. Taylor, and K. Horne. Testing Bekenstein's relativistic Modified Newtonian Dynamics with lensing data. *MNRAS*, 368:171–186, May 2006. 54

[Zwicky, 1933] F. Zwicky. Die Rotverschiebung von extragalaktischen Nebeln. *Helvetica Physica Acta*, 6:110–127, 1933. 17

[Zwicky, 1937] F. Zwicky. On the Masses of Nebulae and of Clusters of Nebulae. *ApJ*, 86:217–+, October 1937. 17, 18, 31