Notes for PX444 - The Distant Universe

Andrew Levan & Elizabeth Stanway

Updated: November 28, 2017

These notes are a guideline to what is included in the PX444 course. They are based on a version first given by myself in 2015 and a modified version delivered by Dr Stanway in 2016. They are not exhaustive; the true content of the course is reflected by the actual lecture, which may differ slightly (or not so slightly) from the material that is presented here. You should plan to take notes in lectures and also be prepared to apply the principles discussed here in new contexts. Also note that the majority of figures are given during the lectures, and are not explicitly included here.

Please be aware that these notes may contain uncorrected typos as they have been prepared primarily for the use of the lecturer during lectures. Please let me (Andrew Levan) know of any significant (i.e. meaning-changing) errors.

PX444 - The Distant Universe

Lecturer: Prof Andrew Levan (A.J.Levan@warwick.ac.uk)

Weighting: 7.5 CATS

Commitment: 15 lectures

Assessment: 1.5 hour summer exam (answer 2 questions out of 3)

Website: http://go.warwick.ac.uk/px444 (Moodle)

Recommended texts:

Loeb & Furlanetto, *The First Galaxies in the Universe*, Princeton, 2013.
Padmanabhan, *Theoretical Astrophysics, vol III: Galaxies and Cosmology*, Cambridge, 2002.
Both books are available online via the university library. Further reading is also required - see overleaf.

Aims:

To study what is known about the Universe to the limits of current observations and beyond, to identify gaps in current knowledge and to look at future prospects.

Objectives:

By the end of the module, students should:

- have a qualitative understanding of the evolution of the early universe, from the Big Bang to the development of large scale structure
- understand the physics underlying theories of galaxy and structure formation, and how observations inform those theories
- be aware of the range and significance of observational evidence for stars, galaxies and structure in the distant Universe.

Synopsis:

Recent observations are beginning to reach back into the Cosmic Dawn - the era when the first stars and galaxies formed. The physical conditions at the time of their formation set the properties of these objects, and their evolution in turn sets the properties of the stars, galaxies and planets that follow. This research-driven module investigates the formation of structure in the early Universe, starting from the Cosmic Microwave Background and moving through the first generations of stars, and onto the large scale structures that we see today.

The module discusses the theory behind the formation of the first stars and galaxies from primordial density perturbations, the build-up of mass through hierarchical structure formation and the importance of feedback in shaping galaxies. It also highlights current observations being conducted to directly observe distant structures, which formed when the Universe was less than 5% of its current age. It outlines the insight that arises from them, and discusses how new observations with missions set for launch in the next few years might answer the remaining, central questions in the field.

Required Reading

Since this lecture course is motivated and informed by current research, a small amount of additional reading required. There are numerous papers that report results that are described in the course. Some of these you should be familiar with (e.g. familiarity may be expected when the course is assessed). Others may help you appreciate the contents of the course more. The core papers are:

- Sobral et al. 2015 ApJ 808 139 "Evidence for Pop III-like Stellar Populations in the Most Luminous Lyman-α Emitters at the Epoch of Reionization"
- Mortlock et al. 2011 Nature 474 616 "A luminous quasar at a redshift of $z = 7.085$"
- Alexander & Natarajan, 2014 Science 345 1330 "Rapid growth of seed black holes in the early universe by supra-exponential accretion"
- Bouwens et al. 2011 Nature 469 504 "A candidate redshift $z \sim 10$ galaxy and rapid changes in that population at an age of 500 Myr"
- Tanvir et al. 2009 Nature 461 1254 "A γ-ray burst at a redshift of $z \sim 8.2$"
- Worseck et al. 2011 ApJ 733 L24 "The End of Helium Reionization at $z \sim 2.7$"

Other papers you might find useful are listed below. These are still valuable, but are typically longer or more theoretical/computational in nature:

Bunker et al. 2004 MNRAS 355 374 "The star formation rate of the Universe at $z \sim 6$ from the Hubble Ultra-Deep Field"

Duncan & Conselice 2015 MNRAS 451 2030 "Powering reionization: assessing the galaxy ionizing photon budget at $z <$10"

Springel et al. 2005 Nature 435 629 "Simulations of the formation, evolution and clustering of galaxies and quasars"

Komatsu et al. 2009 ApJS 180 330 "Five-Year Wilkinson Microwave Anisotropy Probe Observations: Cosmological Interpretation"

Stacy et al. 2010 MNRAS 403 45 "The first stars: formation of binaries and small multiple systems"

Reading Advice

You have access to all these papers online through the NASA ADS system and the University of Warwick library. The library runs a proxy server and provides a tool you can add to your browser's favourites or toolbar. Clicking on this lets you authorise as a Warwick user when loading a paper. Search for "Try Warwick Library Proxy".

The abbreviation "et al" means "and others" (i.e. that there are too many authors to list them all in a sentence). It's usually used if there are more than three authors on the paper. The common form of reference for a paper is First Author et al (date), e.g. "Stanway et al (2011)"

Do not expect to understand every word, or all the technical detail.

Start by looking at the introduction, abstract and then conclusions before tackling the detail. Looking carefully at the figures and reading their captions is often useful. Most people take in more information if they revisit these papers more than once (e.g. when they first come up in the course, and again at the end of the course or while revising).

You may find it helpful to write a brief summary of the key points in your own words.

Points to Consider:

(i) What observations/calculations/models have been undertaken?
(ii) What question did they set out to answer?
(iii) What did the authors actually find?
(iv) Did this agree or disagree with predictions (or other work), or was it entirely unexpected?
(v) Is the interpretation unambiguous, or do the authors note alternatives?
(vi) Do the authors suggest future work, or note difficulties with their data?

Leading Astronomy Journals:

A&A	Astronomy and Astrophysics
AJ	The Astronomical Journal
ApJ	The Astrophysical Journal
ApJS	The Astrophysical Journal Supplement Series
MNRAS	Monthly Notices of the Royal Astronomical Society

also Nature, Science (both general journals with important astronomy content)

Contents

Topic 1

Introduction and Cosmological Background

Objectives for this lecture

- Introduction to the course and cosmological contexts
- Revise and introduce key terminology and cosmological concepts

Further Reading

- PX389 Cosmology lecture notes
- Liddle, "An Introduction to Modern Cosmology"
- Padmanabhan Ch 3

Expected length: 1 lecture

1.1 Overview of the course

The aim of this course is to understand the physical processes at play in the distant Universe, which due to the finite speed of light equates to the early Universe. We will spend most of the course working within a few billion years of the big bang, and almost all of the course in the first half of the history of the Universe.

We are interested in both the observations that can be made of this early epoch, such as the CMB and direct observations of distant galaxies, and the theories that exist to explain these. As a result, lectures will vary between largely observational in nature and largely theoretical.

While this is a standalone course (aside from the assumption of knowledge of the Physics core), PX444 bridges the gap between the very early Universe (discussed in PX389 Cosmology) and the massive galaxies we see around us today (seen in PX397 Galaxies), spanning the vast majority of the history of the Universe.As such, it builds on material that was presented in both Galaxies and Cosmology, and in this first lecture, we will review that material. If you have done those courses this will be revision if you have not it will be an introduction to the necessary material.

1.2 The expanding Universe

1.2.1 Hubble's law and Universal expansion

We live in an expanding Universe, in which at large scales (bigger than gravitationally bound groups of galaxies) all objects are moving away from each other. Relatively locally this is a linear relationship

$$\boxed{v = H_0 D}$$

where v is the recession velocity, D the distance and H_0 the Hubble Constant. At large distances, or earlier cosmic times, we see a deviation from this linear relation suggesting that the value of the proportionality relation depends on when it is observed. In other words, there exists a parameter $H(t)$ which changes with cosmic time and its value now (at time t_0), $H(t_0) = H_0$. This is known as H, the Hubble Parameter, and can vary, leading to the breaking of the linear relationship. The current best value of $H_0 = 67.8 \pm 0.9\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ and was determined by a combination of data from the Planck satellite with other constraints.

1.2.2 Cosmological redshift and the scale factor

Since objects further away from us are receding faster they have a measurable redshift.

$$\frac{\lambda_{obs}}{\lambda_{rest}} = 1 + z \tag{1.1}$$

However, unlike more familiar Doppler shift measurements, this redshift is created from the expansion of space-time itself. Light emitted at larger distances has taken time to reach us (travelling at light speed). As light travels through an expanding universe it is itself stretched to longer wavelengths – the further/longer it has travelled, the more it is stretched. When we look at higher redshifts we are seeing the Universe when it was *smaller* and *younger*. This means there is a link between the redshift and the *scale factor* (which we will denote here as a); higher redshifts give smaller scale factors. Indeed,

$$\boxed{a = \frac{1}{1+z}}$$

.

So, for example, at redshift 1, the Universe was half its current size. We will typically be working at $z > 5$, where the Universe was $1/6$th its current size, but therefore > 30 times denser.

We will use both redshift and the scale factor as proxies for cosmic time, where t_0 implies $z = 0$ and $a = 1$.

1.2.3 Components of the Universe

The Universe consists of many different components, including baryonic matter, dark matter, radiation and dark energy (often also referred to as the Cosmological Constant, which is formally a special case of dark energy). Because these are different materials, their relative densities behave differently as the Universe expands. For matter (either baryonic or dark) $\epsilon_{\mathrm{matter}} \propto 1/a^3$(as you would expect)[1].

[1]We are here expressing an energy density $\epsilon = \rho c^2$, rather than a mass density

For radiation, the photon density decreases as $1/a^3$ but as there is also redshifting the energy of each photon also scales with a, so the energy density, $\epsilon_{\rm radiation} \propto 1/a^4$. The cosmological constant is weird, as it has constant energy per unit volume, i.e. $\epsilon_\Lambda \propto a^0$, so as the Universe expands, the total dark energy contribution gets bigger. Because each of these components has a different equation of state linking their density to their pressure ($P = w\epsilon$, with $w = 0, 1/3, -1$ for matter, radiation and cosmological constant) the processes of star and galaxy formation and evolution will look different according to the regime in which component dominates the total energy density of the Universe.

While different components of the Universe dominate at different times, almost all of this course will be working in an era when the Universe was matter dominated.

1.2.4 Distances

There are many different distances in Cosmology; these occur because the expanding Universe and the effects of general relativity mean there is no one simple distance that is applicable in all scenarios. The important distances that we may consider in this course are

The *proper distance* (D_P) – The proper distance refers to the true (i.e. as measured by a ruler) distance between two points, specified at some point in time. It is familiar, but often not very useful since it changes over cosmic time.

The *co-moving distance* (D_M) – The co-moving distance is far more frequently used than the proper distance, since it takes into account the motion of the Universe. A co-moving distance is invariant under the expansion or contraction of the Universe and is normally scaled to the current time.

The *luminosity distance* (D_L) – This is the distance that an object appears to have because of its luminosity. i.e. if we knew the luminosity of an object then we could calculate its luminosity distance from its measured flux F since $F = L/4\pi D_L^2$. These luminosity distances are *not* equivalent to proper distance because of time dilation in the source, in which source photons are emitted less frequently by a factor of $(1 + z)$. Hence $D_L = D_p(1 + z)$

The *angular diameter distance* (D_A) - The angular diameter distance is relevant if we want to know the physical size of objects on the sky. Normally you would expect to use something like $s = r\theta$ to calculate the size of an object based on the (small) angle it subtends on the sky and its physical distance. However, this doesn't work for distant galaxies. $D_A = D_L/(1 + z)^2 = D_M/(1 + z)/$ An interesting result of this is that it reaches a maximum around $z = 1$ and then declines, so objects get bigger on the sky as they get further away.

1.2.5 Equations of Cosmological Expansion

There are various equations that are used to describe the expansion of the Universe, these include the Friedmann equation

$$\boxed{\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2}\epsilon - \frac{\kappa c^2}{R_0^2 a^2} + \frac{\Lambda}{3} = H^2}$$

The fluid equation

$$\boxed{\dot{\epsilon} + 3\frac{\dot{a}}{a}(\epsilon + P) = 0}$$

and the acceleration equation

$$\boxed{\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(\epsilon + 3P) + \frac{\Lambda}{3}}$$

We will look at the results from these as necessary as we go through the course, but the essence of much of cosmology is to obtain a good solution to the Friedmann equation which determines the growth of the scale factor with cosmic time.

It is worth taking a short while here to look at the relations between the Hubble parameter and other terms in a little more detail. From the Friedmann equation:

$$H = \frac{\dot{a}}{a} = \frac{1}{a}\frac{da}{dt}$$

but $a = (1+z)^{-1}$ so $da = -(1+z)^{-2}\,dz$, and

$$H = \frac{1}{1+z}\frac{dz}{dt}.$$

In a matter-dominated Universe, $a \propto t^{2/3}$. At the epochs we are considering, the Universe is almost-always matter dominated.

The evolution of the Hubble parameter with redshift is determined by solving the cosmological equations but actually has a relatively simple form:

$$H(z) = H_0\sqrt{\Omega_m(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda}$$

1.2.6 Density

In the case of a flat Universe (and in the absence of the Cosmological Constant) the κ term in the Friedman equation vanishes. In this case, there is a critical density directly related to the Hubble parameter,

$$\rho_{crit} = \frac{3H^2}{8\pi G} \tag{1.2}$$

Commonly this critical density is re-written as

$$\Omega = \frac{\rho_{universe}}{\rho_{crit}} \tag{1.3}$$

Glossary of common cosmological terms

This is a brief summary of the common cosmological terms. Other terms are used frequently but have the same meaning as in standard physics (e.g. P=Pressure) and so aren't included here. Be wary of symbols/letters that are reused in different contexts.

- a - the scale factor, a measurement of the size of the Universe, normally $a_0 = 1$, the subscript zero applied to any parameter refers to the current time.

- z - the cosmological redshift.
- H - the Hubble parameter $= \dot{a}/a$, a measurement of the expansion (or contraction of the Universe).
- h - is often used as a scaled Hubble parameter to account for the uncertainty in the current measurements of H_0 e.g. $H_0 = h \times 100$ km s^{-1} Mpc^{-1} or similar and $h = 0.678$ in the current best cosmological model.
- k or κ - normally denotes a measurement of the curvature (but occasionally can be the Boltzmann constant).
- Λ - The cosmological constant, $w = -1$ acts to accelerate the expansion of the Universe.
- w - the equation of state parameter, linked to the pressure via $P = w\epsilon$, $w = 0$ for matter $1/3$ for radiation and -1 for the cosmological constant.
- ϵ - the energy density, ρc^2.
- Ω - The density parameter, defined as $\Omega = \epsilon/\epsilon_c$ where ϵ_c is the critical density.

Commonly used constants:

Mass of a proton	m_p	$= 1.67 \times 10^{-27}$ kg
Mass of an electron	m_e	$= 9.11 \times 10^{-31}$ kg
Charge of an electron	e	$= 1.60 \times 10^{-19}$ C
Thompson cross-section	σ_T	$= 6.65 \times 10^{-29}$ m^2
Electron Volt	eV	$= 1.60 \times 10^{-19}$ J
Newton's Gravitational Constant	G	$= 6.67 \times 10^{-11}$ N m^2 kg^{-2}
Planck's constant	h	$= 6.63 \times 10^{-34}$ J s^{-1}
Speed of light	c	$= 3.00 \times 10^{8}$ m s^{-1}
Luminosity of the Sun	$L_\odot$	$= 3.84 \times 10^{26}$ J s^{-1}
Mass of the Sun	$M_\odot$	$= 1.99 \times 10^{30}$ kg
Parsec	1 pc	$= 3.09 \times 10^{16}$ m
erg (non-SI unit)	erg	$= 1.00 \times 10^{-7}$ J

Planck 2015 ΛCDM Cosmology:

Critical density	ρ_c	$= 6.63 \times 10^{-34}$ m^2 kg s^{-1}	
Hubble Constant	H_0	$= 100\,h$ km s^{-1} Mpc^{-1}	where $h = 0.678 \pm 0.009$
Matter Density	Ω_m	$= 0.308 \pm 0.012$	
Curvature Density	$\lvert\Omega_k\rvert$	< 0.005	where $\Omega_k = 1 - (\Omega_m + \Omega_\Lambda)$
Optical depth to Reionization	τ	$= 0.066 \pm 0.016$	
Reionization redshift	z_{re}	$= 8.8^{+1.7}_{-1.4}$	

Topic 2

The Cosmic Microwave Background

Objectives for this lecture

- To understand the origin of the Cosmic Microwave Background (CMB).
- To understand how its primary anisotropies arise.
- To identify some of the cosmological constraints that can be derived from it.

Further Reading

- Liddle, "An Introduction to Modern Cosmology", Ch 10
- Mo, van den Bosch & White, "Galaxy Formation and Evolution", Ch 3.5
- Padmanabhan Ch 6

Expected length: 2 lectures

2.1 CMB basics

We are used to viewing the Cosmic Microwave Background as being the uniformly distributed radiation that arises due to photons released at the time of recombination when the overall temperature of the Universe fell below about 3000 K and protons and electrons combine to form hydrogen atoms

$$\boxed{e + p \rightarrow H + \gamma}$$

.

This is a slight oversimplification. The photon to baryon ratio (the total number of photons relative to the number of protons in the Universe) is likely set very early in the Universe via matter – antimatter annihilation, and is $\sim 10^9 : 1$. This is the true 'afterglow of the Big Bang' and means that the photons produced from the actual combination process (where the ratio is 1:1) are of little importance to CMB itself. A more precise description of the origin of the CMB is that once the Universe combines (recombination is a misnomer because it was never combined previously) it become optically thin, and then the pre-existing photons can free stream to us rather than being intercepted by charged particles.

However, the true physics underlying the formation of the CMB are slightly more complex still. This is what give it the true power as a cosmological tool.

2.2 Formation of the CMB

Prior to the formation of the CMB, we can assume that the ratios of protons, electrons and hydrogen atoms obeyed some form of equilibrium. We will consider particularly the number of H-atoms since this is the most common form of baryonic matter in the Universe.

We can track the ionisation state of the Universe as a function of temperature using the Saha equation (if you haven't seen this before you can just assume it), which tracks the relative number of ions in a given state as a function of temperature. It can get quite complex, but for hydrogen, and for the case where the only two states under concern are "ionised" or "not-ionised" it is rather more straightforward. The ratio of ionised hydrogen (protons n_p) to neutral hydrogen (n_H) is given by.

$$\frac{n_p}{n_H} = \frac{1}{n_e}\left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} \exp\left(-13.6\mathrm{eV}/kT\right)$$

where n_e is the electron number density and T the temperature. We have a total baryon number density $n_b = n_H + n_p$, and since the Universe is globally neutral (1 electron for 1 proton) we can write

$$n_e = \frac{n_p}{n_p + n_H} n_b$$

so that the Saha equation becomes

$$\frac{n_p^2}{n_H(n_p + n_H)} = \frac{n_p^2}{(n_b - n_p)n_b} = \frac{1}{n_b}\left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} \exp\left(-13.6\mathrm{eV}/kT\right)$$

If we define the *ionisation parameter* as $X_e = n_e/n_b = n_p/n_b$, we can write this as.

$$\frac{X_e^2}{1 - X_e} = \frac{1}{n_b}\left(\frac{2\pi m_e kT}{h^2}\right)^{3/2} \exp\left(-13.6\mathrm{eV}/kT\right)$$

There are several important results hidden in this. The first is that recombination is dependent on the temperature of the plasma. Since this cannot change instantly, nor can the ionization state of the Universe. As a result, recombination is not, as is often described, instantaneous. Instead it takes place in a narrow (but none-the-less important) redshift or scale factor range. Since the temperature of radiation in the universe (i.e. its photon energy or frequency) is proportional directly to the scale factor, this should explain the spread of temperatures seen in the CMB (at least to some degree).

It is common to assume that recombination 'begins' when $X_e \approx 0.1$.

The second point is that the time at which recombination occurs is dependent on the number density of baryons. As we've seen, this scales with a^{-3}. Hence the redshift at which the CMB was emitted provides some immediate constraint on the total density of baryonic matter in the Universe, Ω_b (although this isn't actually the best constraint that the CMB can provide on this). It turns out (see Mo, van den Bosch & White, p147) that we can write the redshift of recombination as

$$(1 + z_{rec}) \approx 1367[1 - 0.024 \ln\left(\Omega_{b,0}\, h^2\right)]^{-1}$$

As recombination proceeds, the number of electrons decreases with the ionisation fraction, and the photons undergo decoupling. The interaction rate of photons and electrons is given by

$$\Gamma = n_e \sigma_t c = \frac{c}{\lambda}$$

where n_e is the electron number density, λ the mean free path and σ_t is the familiar Thompson cross section. The only variable of concern here is the falling electron density, due to both the expansion of the Universe, and recombination. For a matter dominated Universe $a(t) \propto t^{2/3}$ and $H = \dot{a}/a$. Solving the Friedmann equation in this case gives $H \propto a^{-3/2}$. We can calculate the time of decoupling by assuming Thomson scattering continues until $\Gamma > H^{-1}$ (i.e. a scattering event becomes unlikely in a Hubble time). At this point, photons undergo a final scattering, and then free-stream to observers. This decoupling redshift is sharply peaked at $z = 1090$ with a width $\Delta z \approx 0.29$ (based on Planck 2015 results, consistent with the $z = 1067 \pm 80$ derived using older WMAP constraints).

2.3 Observing the CMB

When we observe the sky at microwave frequencies. The CMB doesn't simply "pop" out to be observed, indeed there are multiple issues that must be considered. The first is that the CMB contains a significant dipole caused by the motion of the Earth around the sun, and this must first be removed. The are also residual effects which result in an anisotropic CMB because of the motion of the sun around the centre of the galaxy, and of our galaxy relative to the local group. Indeed, for precision cosmology, one can define a new zero velocity as being relative to the CMB.

Once all of these velocities have been removed, there is still a significant Galactic foreground (i.e. emission from within the Milky Way) to remove. Dust in the Galaxy emits as a low temperature (tens of Kelvin) black body, and this significantly contaminates the microwave regions of the spectrum. However, because the dust temperature is markedly higher than the CMB temperature, and the dust has significant spatial structure (e.g. along the galactic plane). It is possible to remove it by multi-frequency observations of the sky.

Once this is done, the CMB begins to emerge from the data, and structures can be seen on many different scales. A standard way to look at power on different scales is to take the power spectrum. This is familiar in a one dimensional sense e.g. if looking at a time series of photon arrival times from a periodic source, the power spectrum will provide a peak at the frequency of the source. The CMB is actually a projection of a spherical surface, and so a simple power spectrum cannot be applied, instead one uses *spherical harmonics*, and defines a multipole number (like a dipole $l = 2$ or quadrupole $l = 4$). We then see how the power is distributed amongst different multipoles. Rather than a temporal frequency, a multipole is an angular frequency. Small l is large scales, and large l corresponds to increasingly small scales projected onto the sky.

2.4 Structures in the CMB

The presence of structures in the CMB is actually at first sight rather surprising. The temperature of radiation in the Universe scales as $T \propto 1/a$ as it expands, and this is true both before and after decoupling. For radiation emitted at a specific time and in every direction, the temperature of the radiation (and hence the signal) might be expected to the the same everywhere. In other words, a naive expectation would be that the CMB is entirely uniform. To first sight it is - the variations from region to region on the sky are less than 1 part in 10^5... but they are clearly present. Why then are there apparent structures in the CMB?

2.4.1 The Sachs-Wolfe Effect

The actual cause of the apparent variations in the CMB is not strictly because different regions have different temperatures, but because different regions have different densities. A photon, emitted from an overdense region will experience two effects. First there is some gravitational redshift (loss in energy) applied as the photon climbs out of the potential well of the overdense region. Secondly, the mass in the region acts to slow down time (according to general relativity).

The gravitational redshift experienced by the photon (assuming that we are at large radii) is given by

$$z_{\rm gr} = \frac{1}{2}\frac{r_s}{R_e} = \frac{GM}{c^2 R_e}$$

where r_s is the Schwarschild radius and R_e the radius from which the photon is emitted. In the case of our over densities, the Schwarschild radius is small, and the radius of emission large, and so the approximation is valid. Since the energy of the photon is changing, so must its temperature so $T \propto z_{\rm gr}$ and we can easily replace the redshift with a temperature (or change in temperature), and also note that the gravitational potential $V = GM/R$. The change in temperature due to climbing out of the well (T_1) is then

$$\frac{\delta T_1}{T} = \frac{\delta V}{c^2}$$

The second effect is also relativistic, but can be understood simply. Proper time moves more slowly inside the potential well than outside, hence the cooling of the gas is also slower, and so it reaches the temperature for decoupling *later* than the rest of the Universe. Again this is linear, so we can write the time delay as

$$\frac{\delta t}{t} = -\frac{\delta V}{c^2}$$

We now reach 3000K (i.e. decoupling) later, when the Universe had a scale factor $a + \delta a$ and since $a \propto t^{2/3}$ (from the Friedmann equation for a matter-dominated Universe) we have

$$\frac{\delta a}{a} = \frac{2}{3}\frac{\delta t}{t} = \frac{-2}{3}\frac{\delta V}{c^2}$$

Now, if we consider the redshift from the point of photon emission $a = a_{cmb} + \delta a$, until $a = 1$ (the present day), there is less redshift and the temperature that we observe today (T_2) will be larger, so

$$\frac{\delta T_2}{T} = -\frac{\delta z}{z} = \frac{\delta a}{a} = -\frac{2}{3}\frac{\delta V}{c^2}$$

So, what we totally observe is the sum of these two effects, namely.

$$\frac{\delta_T}{T} = \frac{\delta T_1}{T} + \frac{\delta T_2}{T} = \frac{1}{3}\frac{\delta V}{c^2}$$

The combination of these is called the *Sachs-Wolfe effect*, and it is what produces the large-scale power in the CMB maps (dominating at scales larger than about 10 degrees). Note that it means that regions with high gravitational potential (matter overdensities) are redshifted (i.e. cooler) relative to the rest of the CMB.

2.4.2 Baryon Acoustic Oscillations

We believe that the initial conditions of the universe, as set through inflation put down structure at all scales (this is defined by a spectral index, which may deposit slightly different levels of power at different scales, but allows for structure at all of them).

Perturbations on small scales (i.e. caused by small matter overdensities at the surface of last scattering), are not dynamically stable. They will begin to collapse under their own gravity, and thus the material within them will develop a larger pressure than their surroundings. Acting under this pressure, the overdensity will begin to expand into its surroundings, until pressure equalises, and then 'overshoot' since the material will still be expanding. After this, the material within the region is rarefied (i.e. sparse) and external pressure will act to compress it, restarting the cycle. The result is pressure-supported oscillations.

These will occur if the dynamical time is longer than the pressure time, then pressure will respond to dynamical changes. The dynamical time is basically the freefall time ($t_{dyn} \sim 1/\sqrt{G\rho}$), and the pressure time the size of the region divided by the timescale ($t_{pre} = R/c_s$).

Formally in perturbation theory one can obtain wavelike solutions. This is well beyond the scope of this course, but full derivations can be found in the course texts.

These structures will oscillate from a minimum to a maximum, and the largest possible structures will have experienced only a single oscillation (i.e. they will just have compressed) at the time of CMB formation. Since the maximum size of this region is given by the pressure timescale $= R/c_s$ (i.e. the sound horizon at the surface of last scattering) this can be used as a standard ruler:

$$r_d = \int_0^{t_{rec}} \frac{\mathrm{c_s}\, dt}{a(t)} = \int_{z_{rec}}^{\infty} \frac{\mathrm{c_s}\, \mathrm{dz}}{H(z)},$$

where t_{rec} is the age of the Universe at recombination. We can approximate $c_s \sim c/\sqrt{3}$ but we still need to integrate H(z). In a matter dominated universe, $H(z) \propto \sqrt{\Omega_m (1+z)^3}$, so $r_d \propto (\Omega_m z_{rec})^{-0.5}$.

In fact, this is only partially true, it can be used as a standard ruler that is as good as our knowledge of the sound speed and its evolution, but as the sound speed depends on Ω_b we still need some more constraints to do things really well.

These largest structures (with wavenumber $k = 1$) have just had time for a single compression in the age of the Universe at CMB creation, that means that structures with half this size (wavenumber $k = 2$) will have just had time to compress and rarify, and one with $k = 3$ will have had time to compress/rarify and compress again. That means that structures with sizes of $R_{max}/2, 3, 4$ etc will be preferentially visible in the CMB power spectrum, because structures at all other sizes will be washed out (some expanding some contracting etc). As well as the first peak that is so useful for cosmology we can see additional peaks in the power spectrum - the *Baryon Acoustic Oscillations* or BAOs.

The strength of these peaks is a powerful way in which we can determine Ω_b, since baryons effectively weight down the oscillations (a bit like adding weights to a spring, it will fall to a deeper minimum). This means that the even numbered peaks (those which are due to the expansion of regions) are damped relative to the odd number peaks. The larger the value of Ω_b the greater the damping.

As one moves to gradually larger wavenumber peaks we begin to look at structures which may be older, and could have started life in the radiation dominated regime. Therefore they provide a route of measuring Ω_r.

BAOs did not vanish after the CMB was emitted. The matter in the Universe is still undergoing expansion and compression due to pressure-dominated oscillations, and it's only very recently that the dominance of the cosmological constant has significantly affected the basic physics.

As a result, we expect to see regions of compression and/or rarefaction in the matter distribution of galaxies, as well as in the primordial hydrogen of the Big Bang. Again these will appear as peaks in the power spectrum, this time in galaxy separation rather than temperature anisotropies. Essentially they are imprinted of rings of known size on the cosmic web. Since these evolve in a known way with cosmic time, they provide standard rulers at any given redshift and so provide a method to directly measure the evolution in the Hubble parameter, H. This can then be compared to the evolution expected from the standard cosmological model. Identifying BAOs in the spatial distribution of galaxies is challenging - it requires very large-area surveys of galaxies for which the redshift of every source is knwn, which is both time consuming and extremely expensive. In order to probe cosmologically interesting scales and time evolution, they must also be *deep* surveys, probing galaxies at much larger distances than well known local surveys such as the SDSS. However substantial progress has been made recently by dedicated projects such as the Deep Energy Survey (DES), which is looking at how the peaks shift between different redshifts and how that compares with the expectations from existing cosmological models.

2.4.3 Polarization

The CMB is also polarised as it results from the Thomson scattering of electrons. The polarisation can also contain structure, and may correlate (or not) with the temperature structures seen in the CMB. The use of polarisation is important in extending the reach of the CMB to additional cosmological questions such as gravitational waves from inflation (which we won't cover in this course) and reionization (which we will cover later).

Topic 3

The evolution of density perturbations

Objectives for this lecture

- To appreciate in broad terms how perturbations in the matter distribution at early times evolve.
- To put this in the context of the early stages of galaxy formation.
- To understand how fluid motions enable the creation and modelling of large scale structure.

Further Reading

- Loeb & Furlanetto, Ch 2, 3.7
- Padmanabhan, Ch 5, 7

Expected length: 2 lectures

We know from observations of the CMB that there exist small over densities in the Universe at early times. These were first thought to be surprisingly small ($\delta T/T \sim 10^{-5}$), compared with what would be expected for a matter dominated Universe of 10^{-3}. Indeed, these small perturbations are in themselves good evidence for the presence of dark matter, that started collapsing earlier, while the normal matter was entrained with the photons that ultimately form the CMB.

There are different approaches that one can use to determine how these perturbations will evolve with time. Formally, this is a fluid dynamics problem, and we can track the evolution of the cosmological fluid (i.e. matter, radiation, dark energy) under gravity and in the expanding Universe. This is necessary if we want to see the details of structure formation. However, this is a relatively complex approach, dependent on GR and the underlying cosmology, and we can learn quite a lot from a rather more straightforward view.

3.1 Overdensities as small closed Universes

A feature of gravity is that it is linear, and so we can ignore all regions outside the region of interest (if it is uniform, see Birkhoff's theorem). In essence this is the same thing as subtracting off a background and looking at what is left. So, if we have a Universe which is globally of the critical density, but with a perturbation in which the density is higher, we can treat the overdense region as an island Universe in which $\Omega > \Omega_c$.

If we now take the Friedmann equation,

$$\boxed{\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2}\epsilon - \frac{\kappa c^2}{R_0^2 a^2} = H^2}$$

We can substitute in for a matter-dominated Universe $\epsilon = \epsilon_0/a^3$ (still allowing for any curvature), and multiply by a^2 to get

$$\dot{a}^2 = \frac{8\pi G}{3c^2}\frac{\epsilon_0}{a} - \frac{\kappa c^2}{R_0^2}$$

This can be solved analytically, although it is tricky (and beyond the scope of this course). The solutions are parametric and are of the form

$$a(\theta) = \frac{4\pi G\epsilon_0 R_0^2}{3c^4}(1-\cos\theta)$$

and

$$t(\theta) = \frac{4\pi G\epsilon_0 R_0^3}{3c^5}(\theta - \sin\theta)$$

We can confirm that these are solutions of the equation by substituting them back in, noting that

$$\frac{da}{dt} = \frac{da}{d\theta}\frac{d\theta}{dt} = \frac{\frac{da}{d\theta}}{\frac{dt}{d\theta}} \tag{3.1}$$

so

$$\frac{da}{d\theta} = \frac{4\pi G\epsilon_0 R_0^2}{3c^4}\sin\theta$$

and

$$\frac{dt}{d\theta} = \frac{4\pi G\epsilon_0 R_0^3}{3c^5}(1-\cos\theta)$$

So we can now write $\dot{a}^2$ as

$$\dot{a}^2 = \frac{c^2}{R_0^2}\frac{\sin^2\theta}{(1-\cos\theta)^2} = \frac{c^2}{R_0^2}\frac{(1+\cos\theta)}{(1-\cos\theta)}$$

where the latter follows from $\sin^2\theta = 1-\cos^2\theta = (1+\cos\theta)(1-\cos\theta)$

Having done this we can now turn our attention to the RHS of the Friedmann equation. If we substitute in for $a(\theta)$ and remember that in a positively curved Universe $k=1$ we can see that,

$$\frac{8\pi G\epsilon_0}{3c^2 a} - \frac{c^2}{R_0^2} = \frac{c^2}{R_0^2}\left(\frac{2}{1-\cos\theta} - 1\right) = \frac{c^2}{R_0^2}\frac{(1+\cos\theta)}{(1-\cos\theta)}$$

In this case the RHS does indeed equal the LHS and so this is a valid solution of the Friedmann equation.

3.2 The size of structures

In a simple model of the Universe where only gravity governs behaviour, the system re-collapses back to a singularity. This can't happen in practice if we want to make galaxies - all we'd end up with would be black holes. What actually happens is that due to friction, pressure and other interactions, the virial theorem comes into play and redistributes gravitational potential into kinetic energy.

Systems viralize (i.e. settle into a system in which $2K + U = 0$) as they collapse. In effect the random motions of the stars, combined with pressure waves, act to keep the system from falling into a singularity. Note, that we hypothesise that even in the case of dark matter (assuming it is some form of particle, even if that particle is black holes) there is a process called *collisionless relaxation* in which the system will reach a finite size. This enables stars and galaxies to form.

So, if we want to estimate the size of a structure, we need to know the point at which the small closed Universe (i.e. perturbation) we are considering will turn around under gravity and starts collapsing. We can do this by calculating the maximum of a.

since

$$a(\theta) = \frac{4\pi G\epsilon_0 R_0^2}{3c^4}(1 - \cos\theta)$$

then

$$\frac{da(\theta)}{dt} = \frac{4\pi G\epsilon_0 R_0^2}{3c^4}(sin\theta)$$

and so the maximum value of a is just

$$a_{max} = \frac{8\pi G\epsilon_0 R_0^2}{3c^4}$$

Now if we return to the Virial theorem (PE + 2 KE =0), the total energy at turnaround is equal to the total energy at virialization

$$\frac{-GM^2}{a_{max}} = -\frac{GM^2}{R_{virial}} + KE$$

And since we also know that at virialization

$$\frac{GM^2}{R_{virial}} + 2KE = 0$$

We can solve to obtain that when the system has settled, $R_{virial} = 1/2a_{max}$.

3.3 Defining overdensities

What is interesting to look at here is just how dense these regions are at the point at which they start to collapse. We can calculate the value of t when a is at a maximum for these regions

$$t_{a,max} = \frac{4\pi^2 G\epsilon_0 R_0^3}{3c^5}$$

At this point the rest of the Universe has gone on expanding as a flat, matter dominated Universe while the overdense region collapses (decouples from the Hubble flow). The Friedmann equation in this case becomes

$$\dot{a_{\rm universe}} = \sqrt{\frac{8\pi G\epsilon_0}{3c^2}\frac{1}{a_{\rm universe}}}$$

which upon integration is

$$\int a_{\rm universe}^{1/2} da_{\rm universe} = \sqrt{\frac{8\pi G\epsilon_0}{3c^2}}\int dt$$

or

$$a_{\rm universe} = \frac{3}{2}\left(\frac{8\pi G\epsilon_0}{3c^2}\right)^{1/3} t^{2/3} \tag{3.2}$$

If we now substitute in for $t_{a,max}$ then we will obtain that the two scale factors are different by a factor of $(3\pi/4)^{2/3}$. The densities are then different by this factor cubed assuming this perturbation was a small one (i.e. $9\pi^2/16 = 5.5$).

If we assume that the object virializes at $t = 2 \times t_{a,max}$ (a reasonable approximation), then its radius will be $a_{max}/2$. At this point the Universe will be a factor of $2^{2/3} = 1.58$ bigger, and its density will have decreased by a factor of 4 ($= (2^{2/3})^3$). Hence the density contrast at the point at which the object virialises will be a factor of ~ 180 greater than the mean of the Universe. This tallies with an oft used definition of a dark matter halo, which is that it is a region that is 200 times greater than the mean.

3.4 Non-spherical collapse

This is an oversimplification of what is really a fluids process. In particular, not all over densities will be spherical, and so their evolution won't follow a simple pattern. They will first collapse along their smallest axis, to form pancake shape features (these features in practice will also have a finite width because of heating etc). It will then collapse along its next axis to form a filamentary structure. These filaments and voids are exactly what is seen both in structure formation calculations and in observations and form what is known as the *cosmic web*.

As a result, following a spherical over density, and demonstrating that it will behave as a mini-closed Universe is quite insightful, but not complete. In practice the over densities may not be homogeneous (e.g. constant density in space), they may not be spherical, and they will certainly contain material that behaves differently as it collapses (e.g. gas and dark matter). There are many processes that need to be input in order to understand star and galaxy formation.

Most of this work today is done via simulation. These can either be N-body simulations, in which gravity is the only important constraint, or, as is increasingly the case they may be fluid simulations.

3.5 Numerical simulations

A computer simulation breaks a segment of the Universe down into a number of particles or grid points. These may be test particles responding to, for instance, a gravitational or electromagnetic field without modifying it, or active particles which affect their surroundings. The properties of these (their location, mass, temperature etc) are recorded at a series of time-steps, between which the system is allowed to evolve according to the physical laws being modelled. The use of the simulation will depend on

1. the mass resolution, i.e. the amount of mass assigned to each particle. These may represent entire galaxies, individual stars or smaller mass elements.
2. the time resolution, i.e. the interval between time steps at which positions and other properties are calculated.
3. the grid size i.e. the volumes over which properties are averaged when fields and interactions are calculated.

Increasing any of these makes the simulation more precise and usually more useful, but each carries overheads in terms of computational operations and time.

Many simulations also employ so-called 'sub-grid' physics, in which a semi-analytic calculation of physical processes (e.g. formation of stars, energy dumped back into the intergalactic medium by accretion, turbulence) is performed between steps and applied to each element before the next simulation step.

Initial conditions

The initial conditions for cosmological simulations can be obtained from the power-spectrum of the CMB (i.e. what power to put in on which scales). Most simulations will start with an isotropic and homogeneous Universe (i.e. mass particles located on a uniform spatial grid) with periodic boundary conditions that make the region essentially infinite. We then superimpose some perturbations (under- and overdense regions) that have a power-spectrum consistent with the CMB. In practice, we probably want to start our simulations even earlier than the surface of last scattering.

We start with a density contrast

$$\delta(x,t) = \frac{\rho(x,t) - \bar{\rho}(t)}{\bar{\rho(t)}}$$

and look at the wave-like solutions that we get for their distribution

$$\delta(x) = \sum \delta(k) \exp\left(-ik.x\right)$$

Or working in k-space to get a Fourier transform

$$P(k) = \left\langle |\delta k|^2 \right\rangle$$

Inflation actually predicts this spectral index $P_i(k) = Ak^n$, with $n = 1$. So this is a good place to start.

3.6 N-body simulations

N-body simulations used to be the standard means of undertaking cosmological structure simulations and are still widely used. These just need to look at the primordial density fluctuations and some assumptions about the temperature of the dark matter (i.e. the velocity/temperature), along with a standard cosmological model. They could then be left to run with gravitational interactions determining the position and velocity of mass particles at successive timesteps.

$$\text{Force} = m_n \hat{a} = \sum_{i \neq n}^{N} \frac{G\, m_n\, m_i}{r_{ni}^2} \hat{r}_{ni}^2$$

A simple calculation of the gravitational force on each particle due to the locations of every other particle at the previous time step scales as N^2, where N is the number of massive particles. It also requires small time steps such that acceleration is constant in each step. This is the so-called Brute force method.

A softening length may be required to prevent particles from colliding (i.e. r_{ni} going to zero, and the force going to infinity).

The computation time can be improved upon with clever algorithms, for example tree algorithms (like the Barnes-Hut method) which neglect the influence of particles too far away from the test particle to have a significant effect. An alternative is to calculate the mean gravitational field on a mesh, instead of per particle. These meshes can be adaptive (i.e. become more finely gridded where there is a higher density of particles).

A well known example is the Millennium Run, run in 2005 to probe the early stages of galaxy formation. This had N$=2160^3 \sim 10$ billion particles. Each particle represented a billion solar masses of dark matter and these were placed in a cube 500 Mpc/h on each side. It followed the evolution and merging of dark matter haloes. A series of larger or more detailed simulations (e.g. MXXL) have followed from the same team.

N-body simulations are powerful tools for lots of things. Including:

- Abundance of haloes
- Spatial distribution of haloes
- Dark matter density distribution on small and large scales
- Merger histories
- Timescales for halo/galaxy formation
- Predictions for gravitational lensing

However, N-body simulations do not couple the baryons to the dark matter, and so while they are important in answering several fundamental questions they cannot answer all of them. To do this we turn to hydrodynamic simulations.

3.7 Simulating Cosmological fluids

An alternative approach is to treat matter/energy density in the Universe as fluid and consider the movement of elements within it. In the case of a cosmological fluid the equations we need to solve

are the continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho v) = 0,$$

and the Euler equation which describes the velocity field,

$$\frac{\partial v}{\partial t} + (v.\nabla)v = -\frac{\nabla p}{\rho} - \nabla\Phi$$

where Φ is the gravitational potential, which obeys the Poisson equation

$$\nabla^2\Phi = 4\pi G\rho.$$

We can easily check the validity of this approach in a simple case, by looking at a Newtonian expanding Universe. In this scenario the density is a function of time, but not of position.

Let us imagine that we have the Hubble law

$$v(r,t) = \frac{\dot{a}}{a}r = H(t)r$$

The continuity equation then becomes

$$0 = \frac{\partial \rho}{\partial t} + \rho\nabla.v + (v.\nabla)\rho = \frac{\partial \rho}{\partial t} + \rho\nabla.v$$

since $\nabla.\rho = 0$. We can also calculate

$$\nabla.v = \nabla(H(t)r) = 3H(t)$$

so we obtain

$$0 = \frac{\partial \rho}{\partial t} + 3H(t)\rho = 0$$

This just gives us a solution in which the density varies as $1/a^3$ (i.e. $\dot{\rho}/\rho = -3\dot{a}/a$), as we'd expect.

While this is a simple case, we want to solve the equations more generally, then we need to see what happens when we add a perturbation

$$\begin{aligned} \rho(r,t) &= \rho_0(t) + \delta\rho(r,t) \\ v(r,t) &= v_0(r,t) + \delta v(r,t) \\ \Phi(r,t) &= \Phi_0(r,t) + \delta\Phi(r,t) \end{aligned} \tag{3.3}$$

This is analytically tractable... but long winded. You can take a look at some of the textbooks for detailed derivations, if you're interested). The ultimate outcome is that we get wave-like solutions to a series of perturbation equations. These can either oscillate, or exponentially decay or grow.

The decaying mode are not interesting for structure formation, since they peter out very quickly; the growing ones are. These can be followed analytically for some time, solving the fluid equations for each element, but eventually the behaviour becomes non-linear and other effects (e.g. particle-particle interactions) come into play.

At this point, it becomes necessary to work with computational simulations rather than pursuing analytic solutions.

3.8 Hydrodynamic simulations

These typically use smoothed particle hydrodynamics (a discretisation of the fluid equations onto scales which probe local structure) to more fully map the behaviour of the baryonic content of the Universe. Because more interactions and processes are included this is much more computationally demanding and, in the past, has been used to model small scale processes (e.g. the collapse of a molecular cloud into stars). Recent advancements in both computing power and the use of clever adaptive gridding (to follow the matter) enables these to finally predict large scale structure.

An example is the Illustris simulation, which included parameters for gas cooling, photo-ionization, star formation and the interstellar medium, stellar evolution, stellar feedback and black hole feedback. The main simulation had 1820^3 hydrodynamic cells, in a box 106.5 Mpc on a side.

In particular, in comparison to N-body dark matter halo models, fluid approaches can derive:

- The morphologies of the galaxies
- The fate of gas (the IGM, the ISM etc).
- Turbulance
- Star formation and feedback
- Radiative transfer

There is also the addition of smoothed particle *magneto*dynamics, that enables magnetic fields to be followed.

However, these simulations must still be tuned to consider the details that are needed. For example, if you want to follow early star formation you want to resolve objects with mass of a solar mass or so (ideally much less), this immediately rules out simulations of large cosmological volumes.

Topic 4

The first stars

Objectives for this lecture

- To understand the concept of "Population III" stars.
- To appreciate the general principles that govern the formation of these first stars.
- To explore their evolution and ultimate death.

Further Reading

- Loeb & Furlanetto, Ch 3, 5
- Padmanabhan Ch 7
- Sobral et al (2015) - *required reading*.

Expected length: 2 lectures

4.1 Stellar Populations

The formation of the first stars marks the point at which the Universe has firmly deviated from the simple, analytical prescriptions that we have been discussing for the last few lectures. Beyond this point a myriad of physical processes begin to shape the Universe from the quantum scale (e.g. nuclear fusion in stars) upwards. Because observations are very challenging, if not impossible, most of our insights into this process are based on extrapolations and computer modelling. A complete numerical simulation of the Universe containing all of this physics from first principles is unimaginable at any point in the foreseeable future, and so we must study many aspects individually and then attempt to form a coherent picture. For example, in detailed cosmological simulations many stellar or gas physics problems are simply parameterised based on modelling of observations at more detailed scales.

We often talk about the first stars as Population III stars, this is because we believe there are broadly 3 generations of stars typed in terms of their metallicity (i.e. the mass fraction not in hydrogen or helium). These are Pop I (stars like the sun), Pop II (old stars, such as those in the halo, or in globular clusters) which are metal-poor compared to the Sun, and Pop III, the first generation of stars. To date only Pop I and Pop II stars have been observed (although there are some who would question this). This distinction is based on the metallicity and age of the stars observed and does *not* imply

that there are two (and only two) supernovae involved in the creation of a pop I star. Because each generation of stars forms at successively lower metallicities, they tend to have formed earlier in the history of the Universe. By definition Pop III stars are entirely formed of hydrogen and small amounts of helium.

4.2 Cooling

As the gas in a dark matter halo falls into the gravitational potential, it does so at roughly the circular speed at the viral radius (lets call this v_c), the gas then interacts and shock heating creates a hot gas with the *virial temperature*, which we can obtain by assuming the gas follows an ideal gas law and so $KE \sim kT$. Assuming the mean molecular mass of the gas is μm_p, and moves at v_c we obtain.

$$T_{vir} = \frac{\mu m_p v_C^2}{2k} = 1.04 \times 10^4 \left(\frac{\mu}{0.6}\right) \left(\frac{M}{10^8 M_\odot}\right)^{2/3} \left(\frac{1+z}{10}\right) \mathrm{K}.$$

where we have used the collapse of structures discussed in lecture 3 (section 3.2) to determine the dependence of virial radius and hence circular velocity on mass, and scaled these by useful values to compare with physical systems.

A gas cloud becomes unstable, and hence collapses, when pressure support from the gas is insufficient to balance gravity. This occurs due to the *Jeans instability*. Since the pressure depends on temperature, the system now needs to cool until the Jeans mass (i.e. critical mass for collapse) is smaller than individual stars. The Jeans mass arises from the equating of the pressure timescale $t_p \sim R/c_s$, with the dynamical (or freefall) timescale $t_{dyn} \sim 1/\sqrt{G\rho}$, to obtain a Jeans Length. With all the relevant constants this is

$$\lambda_J = \left(\frac{\pi c_s^2}{G\epsilon}\right)^{1/2} = \sqrt{\pi} c_s t_{dyn}. \tag{4.1}$$

The Jeans mass is simply $4/3\pi\rho\lambda_J^3$. Note that there is sometimes a factor of 2 in the definition of λ_J (i.e. diameter vs radius).

The collapse is predominantly therefore determined by a decrease in the sound speed, and the sound speed in an ideal gas is simply

$$c_s = \sqrt{\frac{\gamma k T}{m_m}}$$

where γ is the adiabatic index and m_m the mass of a molecule. Hence, cooling of the gas decreases the sound speed and impacts λ_J, triggering collapse.

Cooling occurs through the conversion of the kinetic energy of the gas into electromagnetic radiation, normally by collisional excitation of an atom, followed by its de-excitation (e.g. by dropping in energy levels and emitting photons). For strong transitions this process isn't very efficient, because once the photon is emitted it is simply absorbed again, and so takes a long time to emerge from the region, carry energy away and hence cool the gas cloud. The strongest cooling lines are therefore so-called forbidden line transitions. These are transitions where the transition appears to break some quantum mechanical selection rule to first order, but is actually allowed at a low rate when the full scenarios are considered (e.g. collisions may excite atoms into the usually-unpopulated upper state). Because these lines aren't easily excited, the photons are more likely to escape the region and allow it to cool. A good example of a forbidden line is the [OII] doublet (singly ionized oxygen) at 3727Å, which is the result of transitions from $2D^{3/2} \rightarrow 4S^{3/2}$ and $2D^{5/2} \rightarrow 4S^{3/2}$ quantum levels.

However, almost all of these forbidden line transitions are excited in multi-electron atoms such as oxygen, nitrogen etc. In the early Universe they are not available, since after the big bang, early nucleosynthesis created a Universe that was ∼75% H and 25% He with only traces of other elements. Because of this, it is hydrogen cooling that is needed. This can't be through strong H-lines that release photons that will simply be reabsorbed, but instead will proceed via the much weaker transitions in molecular hydrogen. It is hard to form H_2, but it can be done via the reactions

$$H + e \rightarrow H^- + \gamma$$

and

$$H^- + H \rightarrow H_2 + e$$

One does need free electrons to do this, but there appear to be sufficient of them in early star forming regions for this reaction to proceed.

It should be noted that UV-light (e.g. from the first stars) can easily break down any H_2 formed and so a molecular gas reservoir isn't an easy thing to maintain. In denser regions, self-shielding can protect some of the H_2 and becomes important.

When the gas passes some characteristic density, collisions become frequent enough to effectively redistribute energy in a local thermodynamic equilibrium. At densities above this, the cloud radiates as a black body, which is far less efficient than line emission. As a result, the gas cloud cannot lose energy effectively, and collapse stalls (or *loiters*). For a hydrogen cloud, this occurs at around 200 K.

4.3 Formation

Once the regions cool sufficiently under H_2 (or atomic H) cooling to form a core we reach a density which is typically insufficient for nuclear fusion to start, but does create a gravitationally bound, pressure supported object. The crucial difference between this object and a local protostar lies in the temperature. While local cooling is very efficient and allows temperatures of a few tens of Kelvin, the typical temperature in a pop III region will be several hundred Kelvin. This results in a high sound speed, and hence a higher Jeans mass. Indeed, we can write M_J as

$$M_J \approx 700 \left(\frac{T}{200\,K} \right)^{3/2} \left(\frac{n_H}{10^4 \mathrm{cm}^{-3}} \right)^{-1/2} M_\odot$$

So for the same density, the Jeans mass at 10K and 200K are different by a factor of ~ 90.

A single protostar will now begin to grow through accretion of infalling material. We can understand what happens in the two cases. Since the pressure is higher in the higher temperature case, more material must be added to cause a collapse and ignition, and so we would predict that the pop III stars would be more massive than the later generations.

We can paramaterize the accretion process as

$$\dot{m_*} = \phi_* \frac{m_*}{t_{ff}}$$

where t_{ff} is the free-fall time, ϕ_* is some dimensionless constant that depends on the medium and m_* the mass of the protostar and is roughly equal to the Jeans mass

$$m_* \approx M_J \approx \frac{c_s^3}{\sqrt{G^3 \rho}}$$

Hence (from c_s above), $\dot{m}_* \propto T^{3/2}$. If we calculate the parameter ϕ_* and hence the accretion rates, they will likely be in the range $10^{-2} - 10^{-3}$ $M_\odot$ yr^{-1}. For the lifetimes of the first massive stars (a few million years), this implies that they may grow to a maximum mass for a Pop III star of around 1000 solar masses. This is almost an order of magnitude larger than the largest 'normal' Pop I and II stars we see around us.

However, we don't know the extent to which these masses might actually be attained. In particular there are likely constraints applied by a combination of feedback and fragmentation.

4.3.1 Feedback

As the material accretes onto a protostellar core it will release some of its material/energy density as accretion luminosity. This accretion luminosity is simply given by

$$L_{acc} = \frac{Gm_*\dot{m}_*}{R},$$

i.e. the material gives up its initial gravitational potential energy as it is accreted. This comes out as light. which interacts with the infalling matter and can arrest its fall. If we use force balance we can write the inward force on a proton and balance this against the outgoing force on incoming protons (assuming a fully ionised medium).

$$\frac{Gm_*m_p}{r^2} = \frac{L_E}{4\pi r^2} c\sigma_T$$

to give us the Eddington luminosity (familiar to those who have done Astrophysics, Galaxies or High Energy Astrophysics). Above this luminosity accretion stops occuring as the material is blown outwards by radiation pressure. This is an important constraint, but it probably only matters relatively late in the formation of the stars since that is when the luminosity becomes higher, and when the inflating gas is closer to the new star (i.e. the gravitational potential released is highest). Hence while this may modify the masses of the stars (and certainly does in the local Universe) it is not likely to be the dominant effect in Pop III stars.

4.3.2 Fragmentation

A bigger concern arises from fragmentation of the material. In local star formation fragmentation of a molecular gas cloud into smaller units, each of which collapses into a star, is ubiquitous, and shapes the initial mass function (IMF) that we observe. In its usual form, this initial mass function is given by a two-piece power law in which $dN/dm \sim 0$ for 0.08 $M_\odot$ (the fusion limit) to 0.5 $M_\odot$ and then declines with $dN/dm \sim m^{-2.35}$ for higher masses. This means that high mass stars are extremely rare.

The extent to which high mass stars in the Pop III regime are therefore limited by fragmentation of gas clouds remains uncertain, and requires simulations to assess. These simulations have been gradually gaining resolution over recent years, and improved resolution appears to yield increased fragmentation, it may therefore yet be that the mass function of Pop III stars appears more like the mass function of Pops I and II.

Current leading hydrodynamic models (e.g. Hirano & Bromm 2016, Stacy et al 2016) suggest that even with feedback and fragmentation taken into account, the mass function determined for 'fragments' (i.e. potential protostars) in simulations of primordial gas clouds, is likely still 'top heavy' (i.e. richer in

massive stars than the local mass function). Where the local IMF has a slope of -2.35, slopes derived from Pop III simulations are still ∼-1 or shallower (some are flat). The precise slope, and the upper and mass limits are still very much open questions. For example simulations of 1000 $M_\odot$ clusters can yield 100 $M_\odot$ stars, while scaling relations in nearby clusters mean that, because of stochastic sampling, $M_{max} \sim 0.39 M_{cluster}^{2/3}$.

4.3.3 Summary of formation

So to summarize. For the first generation of stars we have the following stages

1. Gas accretion onto DM halo, shocking to virial temperature.
2. Creation of molecular hydrogen.
3. Radiative cooling from H_2, reaching some loitering density at around 200K.
4. Gravitational collapse and star formation once Jeans mass reached.
5. Feedback and fragmentation impact final masses.

4.4 Pop III.2

An alternate cooling mechanism comes into play at temperatures above 10^4 K, i.e. if the halo is highly ionized before collapse begins. This means that there are likely to be two sub-populations of first generation stars often called pop III.1 and III.2. Both are formed from pristine gas, unpolluted by supernovae, but pop III.2 have had their environments altered by the radiation from pop III.1 stars which heat their halos.

The collapse of the pop III.2 stars is slightly different because of the heated (and therefore ionized) medium in which they form. At these high temperatures, a deuterium-hydrogen with a permanent dipole can form:

$$H_2 + D^+ \rightarrow HD + H^+.$$

This can't be made for pop III.1 stars because the virial temperature, while higher than for normal stars, is too low for the reaction. But it *can* occur once pop III.1 stars have heated the region. HD cooling is more efficient than H_2 and has a higher critical density. Hence, for Pop III.2 stars, the stages above are broadly followed, but the HD cools the core to ~ 100K, resulting in a lower Jeans mass and a stellar mass distribution potentially closer to that in Pop I/II.

4.5 Evolution

Once the first stars have formed they begin to evolve. Because of the very different chemical makeup their evolution is also quite different to that of local stars. Again this comes down to the presence of heavy elements in their atmospheres. In Pop I and II stars, these elements absorb outgoing radiation and experience radiative levitation in the atmospheres. The approximate process is as follows:

- Atoms (e.g. O,N etc) interact with photons of energy equal to energy levels within them.
- Since the flux from the star is radially outwards they obtain an outwards velocity relative to their initial one

- At some later time the photon is re-radiated in a random direction, so the atom normally attains some additional velocity
- Because of this additional velocity the particle interacts with photons of a slightly different wavelength
- This process repeats to create a radiatively driven wind. In extreme cases, this blows the outer layers off the star.

Because various elements of the stars form a plasma, and this wind is carried out along magnetic field lines it doesn't purely consist of metals, but can also carry a significant fraction of H out as well.

This loss of material has profound effects for the star. Firstly it reduces its mass throughout its main sequence life, causing it to behave as a much less massive star as it ages. For example, the Galactic cluster W31 contains a highly magnetic neutron star (called a magnetar), despite also hosting stars that may be as massive as 40 $M_\odot$ today. For a single moment of star formation general consensus says that this star should have collapsed to form a black hole and not neutron star.

The second change in the behaviour of the stars arises because the material in the wind also carries away angular momentum, since the field lines remain anchored to the star out to large radii. Hence metals dramatically slow the rotation of stars through their lives, so that most end as relatively slow rotators.

In Pop III stars neither of these processes is as strong. Hence, a massive population III star might well start out more massive and stay more massive for longer, and therefore are typically more luminous and hotter.

4.6 Observational Signatures

Such massive stars are supported by radiation pressure rather than internal convection, and so radiate at their Eddington luminosity. For a star with M$= 100\,M_\odot$, L$\sim 10^6\,L_\odot$.

The same star will have both a hot core (since the CNO cycle is inefficient) and a high effective surface temperature, $T \sim 10^5$ K. As a result it will be very blue (have a hard spectrum), and kick out photons sufficient to ionize not only Hydrogen (13.6eV) but also Helium (24.4, 54,6eV). This hard spectrum leads to distinct observational signatures, mostly due to the surrounding interstellar medium (the residual gas between stars):

1. Hydrogen in the ISM will be excited and emit strongly in narrow lines associated with Lyman-α, H-α, H-β etc transitions.
2. Since the stellar population is young Lyman-α, in particular, will have a very high equivalent width.
3. Helium will *also* be strongly excited, leading to He I and He II emission lines. These are uncommon in normal Pop I/II stellar populations.
4. There will be no strong lines of other elements (as might be expected if some other emission mechanism was responsible for exciting the Helium in a higher metallicity environment).
5. Scattering from dense nebular gas produces a continuum emission which is slightly redder than the very blue stellar emission. However, it will still be blue overall and this will be reflected in its photometric colours.

A handful of very distant galaxies have been proposed as Pop III candidates, notably an example at $z = 6.6$ known as CR7 (Sobral et al 2015). However alternate explanations have also been proposed for many of their features.

4.7 Pop I, II and III SNe

A Pop III star will end its life with much of its initial mass and rotation intact. This can have profound implications for its explosion mechanism.

In local stars supernovae form when the star runs out of nuclear fuel to burn. It does this when an iron core is formed, and no more energy can be extracted from further nucleosynthesis. Once the iron core reaches the Chandrasaker mass it can no longer support itself via degeneracy pressure and so collapses along with the rest of the star. The star then bounces off this core, electron capture ensures creating copious neutrinos and leaving a ~ 1.4 $M_\odot$ neutron star, and a rapidly expanding supernova remnant. Nucleosynthesis within this supernova remnant then powers the explosions, with the radiation trapped while the SNe is optically thick, and released as it gets bigger and becomes optically thin. These SNe are driven by radioactive decay and their brightness is determined either directly from the quantity of ^{56}Ni produced or by the interaction of the outgoing shock wave with the circumstellar medium.

It actually isn't really this simple, and there are multiple possible ways to do this (bounce, neutrino drive, acoustic) all of which have their advantages and disadvantages. None-the-less, these core collapse SNe describe most of the supernovae we observe in the Universe today.

In those SNe (excluding those driven by an interaction with the circumstellar medium) the peak luminosity is approximately proportional to the mass of nickel produced.

In the first generation stars the cores may be much more massive, and attain much higher temperatures. This means they may die in alternative routes. One such route is called a pair instability SNe. In this case the temperature inside the star becomes sufficiently hot for pair production from photons. If these are made they move more slowly than the photons and hence reduce the pressure in the star causing a collapse. This collapse can be complete (in which case we have a pair instability supernova) or it may be incomplete losing only some mass (a pulsation pair instability supernova).

The rotation of these stars can also be important. In particular, for rapidly rotating stars conservation of angular momentum means that the material in the star cannot directly rain down on the centre during core collapse. In particular if the specific angular momentum is greater than that necessary at the innermost stable orbit of the black hole then a centrifugally supported disc will form. This disc may be stable or unstable (more likely) but could feed the black hole at an extremely high rate (solar masses per second, note because it is very fast the Eddington limit doesn't apply). This is likely to be the mechanism that creates gamma-ray bursts.

GRBs are clearly found in the local Universe, as are a recently identified population of supernovae with luminosities 100 times greater than those of normal SNe. Both GRBs and SNe appear to favour environments that have a low metallicity. However, none of the environments appear to be pristine. Hence, it is likely that these processes operate on a scale, where they become more common at lower metallicity. Hence they may be even more common, or even more spectacular in the early Universe.

Topic 5

Observations of the first stars and galaxies

Objectives for this lecture

- To understand the difficulties inherent in finding distant galaxies
- To explore some of the methods used to identify and confirm them

Further Reading

- Loeb & Furlanetto, Ch 10
- Sobral et al (2015) - *required reading*.
- Bouwens et al (2011) - *required reading*.

Expected length: 1.5 lectures

5.1 The Challenge of Finding Distant Galaxies

We have been talking about distances to galaxies. Aside from for galaxies in the relatively local Universe, or those which host a standard candle (such as an SNIa), we are really talking about measuring their redshifts. For a given cosmological model there is a direct relationship between redshift and distance, and hence the age of the Universe when the light from a distant source was emitted.

So, if we want to know how far away a galaxy is, we need to measure its redshift. Ideally, we'd just take a spectrum of all the galaxies in the sky, measure their redshift from the wavelength of emission or absorption lines, and hence know the distance to everything. However, this isn't possible with current technology (nor with any technology that is likely to be available soon), since the galaxies are both too numerous, and often too faint. You might then consider that if we want to find the farthest galaxies we just focus on those that are faintest. However, this isn't easy either - both because these are the most challenging to observe, and because they are difficult to find.

The distribution of brightness of galaxies in the Universe is described by a luminosity function

$$\phi(L)dL = \left(\frac{\phi_*}{L_*}\right)\left(\frac{L}{L_*}\right)^{\alpha}\exp\left(-L/L_*\right)dL,$$

which is characterised by a faint end slope, α. This faint end slope is always negative, meaning that there are more faint galaxies than anything else, so if you picked a galaxy because it was faint then in general you would be much more likely to find a faint, (fairly) nearby galaxy, than a very distant one.

So, how might we find the most distant objects? – to do this we need a technique which allows much more multiplexing that spectroscopy. The obvious answer is imaging. After all, the Hubble ultra-deep fields (and a handful of other deep field observations now scattered across the sky) are already taken in multiple colours (i.e. measuring the light at a handful of distinct wavelengths). Each deep-field image is very small on the sky (probes a small volume) but potentially contains thousands of galaxies, so if we can find a good route to determining distances from those then that will help.

5.2 Photometric Redshifts

Fundamentally, an astronomical image obtains the average brightness of galaxies in a known wavelength range specified by the sensitivity range of a filter and detector. As imaging collects photons from a broad wavelength range, across a fairly large field of view (arcminutes), it is efficient and so can be done with more than one filter, building a catalogue in many different colours.

This is particularly useful if there are features in the spectrum that are strong enough to show up in broad band observations. Such features might include sudden drops in flux such as the Balmer break at 4000Å, which is caused by absorption in the atmosphere of stars (it is caused by the termination of the Balmer series $n = \infty \rightarrow 2$), for example. We might also see strong features due to dust absorption, or from strong emission lines, dropping boosting the flux in a particular filter relative to its neighbours.

Through imaging we map out the distribution of light with wavelength to create a spectral energy distribution. If we do this well enough then we don't need to have a spectrum (although one is always nice). The galaxy type and approximate redshift can be estimated from comparison to a set of (redshifted) templates.

In principle with a well sampled spectral energy distribution we can reasonably determine the *photometric redshift* of a galaxy.

5.3 The Lyman Break Technique

However, the biggest strength of photometry arises when we want to search for even more distant galaxies. Here we can make use of a special case of the photometric redshift technique, focussing on a feature that isn't intrinsic to the galaxies, but arises in the intergalactic medium, called the *Lyman break*. This is caused by clumps of neutral gas along the line of sight. Each of these will absorb photons at a wavelength corresponding to the Lyman-α (1216Å) transition of hydrogen, as seen at their own redshift (i.e. 1216Å$\times(1 + z_{\rm cloud})$. Even small amount of neutral hydrogen absorb all light with wavelengths shorter than the Lyman-limit (912Å rest), the so called EUV region, and so this will also disappear from the spectrum.

As we move to higher redshift, we have a longer path length, and so are more likely to pass through a H-cloud. The Universe is also becoming increasingly denser, and so the number density of such clouds might increase (in fact, it scales as $\sim (1 + z)^{2.3}$). Hence there is more and more absorption

as we move to high redshift. Each cloud creates a narrow absorption feature at the wavelength of Lyman-alpha at the relevant redshift 1216Å$\times(1+z)$. If there are many of them then rather than seeing individual lines we see a forest of narrow absorption line systems known as the *Lyman-alpha forest*.

Anywhere outside our own galaxy, the drop in flux at 912Å (rest) is total, causing a step in the spectral energy distribution.

By $z \sim 3$, the Lyman-alpha forest is dense enough to eliminate half of the flux between 912 and 1216Å (in the rest frame of the distant object), causing a second step.

By $z \sim 5$, the drop at 1216Å is $>$95% of the flux level longwards of the break.

As a result, we can use this feature to detect high-redshift galaxies, because they will exhibit strong breaks in their photometry at one or both of the Lyman break wavelengths (in the rest frame). We can obtain a sample of high redshift galaxies (or indeed anything else with a strong feature) by taking observations in filters that straddle this break and looking for objects with large colours. The blue filter is the so called veto filter, a high redshift galaxy should appear very faint (or non-existent) in this filter. It should then appear in the adjacent filter, straddling or longwards of the break, and be present in a third filter further redward. It is usual to require a blue colour longwards of the break to remove contamination from intrinsically red sources that might e.g. have heavy dust extinction. This is the route that we use to obtain simple redshift estimates of distant *Lyman break galaxies*, and identify samples for further study.

Note: because this selection technique works in the rest-frame ultraviolet, it selects *star-forming* galaxies. It also fails on very dusty galaxies, since these are redder than their stars would suggest.

Because the galaxies we're looking at are very faint and are intrinsically small, we need to minimise the effects of Earth's atmosphere (which tends to scatter and spread out light). As a result most of this work is done with the *Hubble Space Telescope* (*HST*) which is above the atmosphere. Because Hubble has a very small field of view, the 'deep fields' we use for this work are sometimes called *pencil-beam surveys* - they're looking at a long thin region of spacetime. Much of this work has been done in a handful of established deep fields (e.g. HUDF, GOODS, HDF etc). The new *James Webb Space Telescope* (*JWST*) will also observe deep fields, and is likely to extend our samples to higher redshifts (further into the infrared) but will not observe in the optical.

Examples of this method include the work of Bouwens et al (2011) who used three filters – Y, J and H – which centre around 1 μm. By identifying sources which appear in the H band, but drop out in Y and J, i.e. which has a break at about 1.4 μm, they found a candidate $z \sim 10$ galaxy, as well as a number of candidates detected in J but not Y, which may be at $z \sim 8$.

Sobral et al (2015) also used a varient of this technique to find their PopIII galaxy candidate CR7 (and other candidates). This source is a strong Lyman-α emitter (i.e. has a very powerful emission line at 1216Å rest) and so was identified as having an excess of flux in a narrowband filter (which measures the line) compared to an adjacent broadband filter (which measures the continuum). Most Lyman-alpha emitters (LAEs) are also Lyman-break galaxies (LBGs), although they are often very faint.

5.4 Confirming high-z galaxies

It is important to remember that these photometrically-selected sources are initially *candidate* high redshift galaxies. They have colours consistent with high redshifts, but there is no guarantee that they are actually at high z, short of taking a spectrum. Contamination of samples is a serious issue - only rare low redshift objects have these extreme colours, but the real high redshift galaxies are rarer still.

Most contaminants can be identified through taking more photometric images at longer wavelengths, but this is difficult and in many cases, the objects are too faint to be entirely certain.

Spectroscopic confirmation of these galaxies is hard, but is possible. Essentially they are so faint that even in a broad filter we barely detect them. This means that even with large telescopes we cannot measure their properties in absorption. Fortunately, young, star forming galaxies often have strong emission lines, and these emission lines are detectable. In a few cases beyond $z \sim 7$, deep observations (tens of hours of 8m telescope time) have yielded detections of Ly-alpha in emission. Because targets are rare, usually only one target can be observed at a time. This makes this observing very expensive. In 2012, the Keck telescope calculated the cost of a night of observing (8 hours) as \$54k. Other large telescopes have similar costs. So spending 10 hours on a single target means spending almost seventy thousand dollars, *per galaxy*.

The situation is, in fact, worse than this. Ten hours might suffice for a bright Lyman alpha emitter like CR7. For a fainter, more typical, galaxy at z 6-7 we might need 20 hours or more. To add to the problem, only about 20% of galaxies at $z \sim 5$ and perhaps 10% or less at $z \sim 7$ actually have a strong emission line that escapes the galaxy and reaches Earth. So realistically, we're looking at spending \$1 M on telescope time for every confirmed $z \sim 7$ LBG.

Even then, we have only secured a single emission line, so it is difficult to be sure that it is Ly-α, rather than other emission lines of lower redshift objects. However, the solution here comes from the asymmetry of the Ly-α line. The blue wing is scattered by outflows and winds, as well as being eaten away by the Lyman-alpha forest. The red wing is broadened by back-scattered light, and so ultimately Ly-alpha has a strong asymmetric profile. If we have sufficient signal to noise we can see this profile, and so be confident of our redshift. In *very* rare cases, mostly at $z \sim 5-6$, a second line can be detected (e.g. C III] 1909Å, O III] 1665Å or He II 1640Å) which locks down the redshift. Note that any of these lines requires a very young, and probably lowish metallicity stellar population.

However, it may be that not all galaxies have Ly-alpha in emission, or that in some cases all of this Ly-alpha is absorbed before it can reach us, and so spectroscopically-confirmed samples will always provide only a lower limit on the true properties of the high$-z$ galaxy population.

Clearly the biggest problem with high-z galaxies is their faintness. This means that while they provide us with a large number of *possible* high-z galaxies there will always be complex corrections for galaxies that cannot be seen (the faint end)and for contributions from contaminating sources. A bright source population (i.e. very luminous) would therefore be very valuable. One possibility for finding such sources would be to make use of gravitational lensing (which was a target of the Hubble Frontier Fields programme). Unfortunately, this phenomenon is rare and seldom seen (the volumes strongly lensed at high redshift are tiny).

Topic 6

Cosmic Probes

Objectives for this lecture

- To understand the use of GRBs and their host galaxies as cosmic probes
- To understand the use of Quasars as cosmic probes
- To appreciate the problem of building the highest redshift quasars

Further Reading

- Loeb & Furlanetto, Ch 7, 10
- Mortlock et al (2011) - *required reading*.
- Alexander & Natarajan (2013) - *required reading*.
- Tanvir et al (2009) - *required reading*.

Expected length: 1.5 lectures

6.1 Quasars

6.1.1 The nature and detection of quasars

Quasars (or QSOs) are accreting supermassive black holes which lie at the centre of galaxies, and are part of the family of Active Galactic Nuclei (AGN). They occur when we view towards (though not exactly down) the polar axis of an accreting black hole. They have the advantage of being extremely luminous, and so can be seen, and are even bright beyond $z \sim 6$.

Quasars also have a luminosity function that is reasonably described by a Schechter function (which we use for galaxies):

$$\phi(L)dL = \left(\frac{\phi_*}{L_*}\right)\left(\frac{L}{L_*}\right)^{\alpha} \exp\left(-L/L_*\right)dL,$$

However, increasingly the use of double power laws has become more common and is used by, for example the Sload Digital Sky Survey (SDSS) analyses of quasar populations:

$$\phi(L)dL = \frac{\phi_*}{(L/L_*)^{\alpha+1} + (L/L_*)^{\beta+1}}dL.$$

Either way, the crucial normalisation factor ϕ_* is lower in the case of quasars by a factor of $10^4 - 10^6$. In other words, if a single HST field contains a handful of galaxies at $z \sim 7$ then we need to observe hundreds of thousands times that area in order to find distant quasars. This is hard work. It cannot be done from space, but we must come back to the ground. Indeed, finding these distant quasars was one of the main reasons for the construction of the Sloan Digital Sky Survey, and has also been attempted by other wide area surveys, such as those using the CFHT and VISTA telescopes.

The basic principle of colour selection is to use the Lyman break to identify quasars, but it is particularly challenging because of the sheer number of other sources at the magnitudes we're interested in. In galaxy Lyman break surveys, we can get rid of virtually all the bright objects as potential high-z systems, but a bright quasar may have 20th magnitude even at $z \sim 7$. While local observers consider this faint, for extragalactic astronomy, this is a very bright magnitude. As a result, selection against contaminants is harder, and often requires morphological information (selecting against extended galaxies with the same colours) and additional wavebands (to identify the very flat spectrum typical of quasars). Large numbers of quasar candidates are also more likely to be targeted by spectroscopic surveys, since their relatively bright magnitudes make detection and identification more straightforward.

None-the-less many high-z quasars have now been found, and these provide a valuable route to measuring the high-z Universe.

6.1.2 The problem of mass build-up

The mass of a quasar can be estimated by assuming that it radiates at the Eddington limit.

$$L_{Edd} = \frac{4\pi G M m_p c}{\sigma_T} = \epsilon \dot{m} c^2$$

Where ϵ is an efficiency factor. For the most distant quasars this suggests that the mass is $> 10^9$ solar masses. This is particularly challenging, because when we look at how we might make supermassive black holes the normal assumption is that they grow by accretion. The seed black holes which begin this process are created in stellar core collapse. If this is the case then as the mass accretion rate is proportional to the mass, the growth rate is exponential.

$$M_{BH} = M_0 \exp{(t/t_{\rm e-fold})}.$$

where $t_{\rm e-fold}$ is a constant growth time-scale fixed by the above equation. Unfortunately, the time needed to build up very massive black holes in the early Universe is $>$ age of the Universe.

Let's take ULAS J1120+0641 (Mortlock et al, 2011) as an example. This source has an observed luminosity of $5.3^{+3.9}_{-1.8} \times 10^{13}\,\mathrm{L}_\odot$. This is the Eddington luminosity for a black hole of mass $2 \times 10^9\,\mathrm{M}_\odot$ (using the standard values for $\mathrm{L}_\odot$ and $\mathrm{M}_\odot$ and the equation above). Assuming efficiency $\epsilon = 0.1$ (reasonable for accretion from a thin disk), the e-folding timescale $t_{\rm e-fold}$=0.04 Gyr.

If we start from a seed black hole with M$\sim 10\,\mathrm{M}_\odot$ (reasonable for normal stellar evolution), then the time to reach $10^9\,\mathrm{M}_\odot$ is given by

$$t = 0.04\,\mathrm{Gyr} \times \log_e\left(\frac{\mathrm{M_{BH}}}{\mathrm{M_0}}\right) = 0.76\,\mathrm{Gyr}.$$

But this source was observed at $z = 7.085$, which corresponds to a time just 0.77 Gyr after the Big Bang. In other words, this source has to have been accreting, non-stop, at maximum possible efficiency, ever since the Universe was created. This is impossible for two reasons: (i) since the stars needed to seed it didn't even form until $z \sim 20$, when the universe was already $\sim$0.2 Gyr old; and (ii) since we know that quasars at lower redshift have a short duty cycle, i.e. that they accrete rapidly for a short period ($\sim 10^5$ yr) and then 'switch off' until the nearby gas cloud rebuilds. In other words, the presence of this quasar, and others like it, implies that we need to find other mechanisms to explain their high masses at such early times.

6.1.3 Possible solutions

Primordial black holes: Perhaps there are primordial seed black holes that form at very early times (e.g. well before the first stars form), maybe because of extreme density fluctuations induced by inflation or other processes. If black holes can form early in the Universe then we might well be able to grow them to the necessary masses, although these would need to be massive black holes in the first place.

One subset of hypothesised primordial seed black holes is the 'direct collapse' scenario. In this, some of the densest regions at high redshift may have collapsed so rapidly that they do not ignite hydrogen burning until their cores already lie within their own event horizons. This could occur fairly late - at a similar time to the first star formation. Note that a DCBH explanation has been proposed for CR7, but remains far from conclusive.

Pop III seeds: As we've seen, Population III stars may be much larger than their Pop I/II counterparts. If so, and if they survive a supernova without total disruption, they may form much larger black holes - of 100 s or even 1000 s of solar masses. While this is rather unlikely (most sources that massive would likely destroy themselves in a PISN, while also blowing out any nearby material it might then accrete to grow further) it would neatly bypass the very slow initial phases of the build-up process.

Super-Eddington accretion: A third possibility is that we undergo super-Eddington accretion at some point. The Eddington limit is used to describe an equilibrium situation, and so if the accretion isn't under equilibrium then it doesn't apply. One option might be to merge black holes in haloes. This has some appear, since haloes will merge under the presence of gravity, and eventually gravitational wave emission will cause the black holes at the centres to merge. However, this isn't so appealing as a means of obtaining very massive black holes early in the Universe because the merger process is long, and only roughly doubles the mass of the black hole.

A more appealing possibility is that the black hole accretes from the dense interstellar medium in which it is found. The Jeans mass of a collapsing cloud early in the Universe might be 10^5 solar masses, while it may form only one hundred solar mass star and a single 10 solar mass black hole. However, once formed, the black hole will accrete from the interstellar medium at the Bondi-Hoyle rate

$$\dot{M} = \frac{4\pi\rho G^2 M^2}{c_s^3}.$$

In this case $\dot{M} \propto M^2$ and so the growth can be very fast. It can overcome the Eddington limit because the black hole moves through the ISM so its radiation isn't impacting the same gas that is falling in. However, the challenge is in getting the ISM density sufficiently high to obtain the full accretion on a sensible timescale, and over a prolonged period.

6.1.4 Cosmic Lighthouses

The primary uses of quasars as cosmic probes is as cosmic lighthouses, illuminating a beam through their surroundings. This means they can be used to probe their nearby proximity zone (which will affect the shape of the Lyman-alpha line), and the ionization state in the nearby intergalactic medium. Indeed quasars provided our first strong evidence for the epoch of reionization (see next topic).

Since their intrinsic spectra are so bright, they also allow us to probe for absorption systems along the line of sight, and hence the cold neutral component of the cosmic web. These manifest both as a Lyman-alpha forest, and as a phenomenon known as *quasar absorption line systems* – the metal element equivalent. The 'barcode'-like pattern of metal absorption lines is overprinted on the quasar spectrum. Several patterns, each with a different redshift, can be printed onto the same spectrum. By probing atomic species such as C IV and Mg II, they allow the history of cosmic chemical enrichment to be probed.

One advantage they hold over GRBs (as we will see) is that their fluxes tend to be (relatively) stable over time, and so a rapid response is not required after some trigger, and observations can be built up over an extended period. A disadvatage is that they likely occur in regions very overdense with respect to the cosmic mean density at their redshift (i.e. making a big quasar requires a big galaxy). As such they are poor traces of the bulk of material, and star formation, in the distant universe - which lies in much less extreme regions. On the other hand, they are excellent tracers of the highest density regions in the distant Universe.

6.2 Gamma Ray Bursts as Cosmic Probes

6.2.1 The nature and detection of GRBs

Gamma Ray Bursts are, by the simplest definition, just that - an astrophysical transient defined by the detection of gamma ray photons where none were seen before. This defines two of their properties - their peak flux, and the duration within which 90% of their photons were detected, t_{90}. Initially detected by satellites designed to identify nuclear bomb tests (which also release gamma rays), they were identified as first astrophysical and then extragalactic based on their luminosity, frequency and distribution on the sky.

A classification into two sub-populations was introduced when the sample size was large enough. GRBs divide into short bursts ($t_{90} < 2\,$s, generally hard/blue in their spectra, often at $z < 1$) and long bursts ($t_{90} > 2\,$s, soft spectra, typically originating at higher redshifts). These two populations are believed to have different progenitors.

Short GRBs are believed to arise from the collision and merger of two neutron stars. While this is of great interest in the context of gravitational waves, it is not the focus of this course.

As we mentioned in the context of Pop III stellar deaths, long GRBs are believed to arise from the core-collapse of a massive, rotating star which launches a relativistically beamed jet along our line of sight. Their association with type 1c (hydrogen-free) supernovae suggests the progenitor is typically an evolved star which has lost its hydrogen envelope.

Studies of the host galaxies of long GRBs suggest that they are strongly associated with star formation, and that they are more likely to occur at sub-Solar metallicities. As such, they should be good tracers of star formation in the distant Universe, regardless of the size of galaxy (or dark matter halo) hosting the forming stars.

The identification of these sources is typically undertaken from space,and has been transformed by the

Swift space observatory. The gamma-ray sensitive Burst Alert Telescope (BAT) can see 1.4 steradians (out of $4\,\pi$ on the sky) but is poor at localising bursts. When a burst is detected, the telescope slews to centre the estimated location in the X-Ray Telescope (XRT) which has a much smaller field of view, but a much better angular resolution. This, ideally, permits an optical counterpart and host galaxy to be identified either in the on-board ultraviolet-optical telescope (UVOT) or by from ground-based observatories. Since the event (and its afterglow) is transient, a rapid response is necessary from all concerned.

6.2.2 The most distant GRBs

GRBs emit an isotropic-equivalent energy (i.e. the energy we would estimate, assuming that the emission was equal in all directions rather than beamed), as high as $10^{53}\,\mathrm{ergs\,s^{-1}}$. This contrasts with $10^{44}\,\mathrm{ergs\,s^{-1}}$ for a luminous quasar. As a result, they can be seen to very large distances. In theory, the brightest bursts identified to date could be seen out to $z \sim 10 - 20$ (if they occured there and if Swift or another observatory happened to be looking in the right direction at the time).

Since not every burst has an associated (or at least detected) optical afterglow, figuring out that's where it came from might be considerably harder!

The most distant GRBs identified to date have included one at $z = 9.4$ (GRB 090429B), one at $z = 8.2$ (GRB 090423) and a handful more at $z \sim 5 - 6$. The redshifts of most of these sources was determined from their afterglows (i.e. the initial optical transient), since their host galaxies are typically exceptionally faint. Indeed even long integrations with the Hubble Space Telescope (currently the most sensitive instrument for this work) has failed to identify credible hosts in most cases.

In recent years, GRBs have overtaken quasars and galaxies as the most distant sources with secure redshifts (although there are now credible galaxy candidates at distances greater still).

6.2.3 GRBs as Cosmic Probes

The use of GRBs as probes is similar to that of QSOs. Their very bright afterglows give a line of sight against which absorption line systems reveal the presence of structure. This begins within the host galaxy itself, allowing the hydrogen column density and chemical composition of the host galaxy to be determined.

Their redshift distribution likely traces the cosmic star formation history.

The host galaxies of distant GRBs are typically low in mass ($\sim 10^8\,\mathrm{M_\odot}$). This means that they're very common - indeed you'd expect many in a single Hubble Telescope field. Unfortunately, as already mentioned, they are also exceptonally faint and so are not typically detected in surveys. GRBs therefore reveal the low luminosity star formation that likely dominates the formation of stars in the distant Universe, and give otherwise inaccessible information about those regions. This is of particular importance in probing the last great phase change in the history of the Universe: reionization.

Topic 7

Reionization

Objectives for this lecture

- Understand the physical processes which drove the reionization of the Universe.
- Appreciate how we might constrain and observe the process of reionization

Further Reading

- Loeb & Furlanetto, Ch 9, 11, 12
- Padmanabhan Ch 9
- Mortlock et al (2011) - *required reading*.
- Worseck et al (2011) - *required reading*.

Expected length: 2 lectures

7.1 Background

After the creation of the CMB the Universe the baryonic component of the Universe consisted mostly of neutral hydrogen. This hydrogen flowed with the other baryons and dark matter as it was entrained into dark matter haloes and subsequently formed the first stars, as was discussed earlier in the course. We have looked at searches for some of these early objects, but we now want to know what they can tell us about processes and key events in the history of the Universe, and we will start with reionization.

The *Epoch of Reionization* (EoR) refers to a period during which the universe made its last major state transition from neutral to ionised. Although neutral hydrogen clouds remain (for example they are the cause of the Lyman-alpha forest), the vast majority of hydrogen in the Universe is now ionised. This ionisation must have happened through some source of UV-photons early in the history of the Universe, each of this will have ionized its surroundings, before those ionized regions overlap. We will now examine this process, and consider how we may map it.

7.2 Strömgren Spheres

If a source emitting ionising radiation is placed within a neutral hydrogen cloud then it will emit ionising photons at some rate $\dot{N}$. These will ionise the H-atoms in the immediate vicinity, but they will then attempt to recombine, at some other rate. If these processes are in equilibrium then

$$\frac{n_{HI}}{t_{ion}} = \frac{n_{HII}}{t_{rec}} \tag{7.1}$$

where n_{HI} and n_{HII} refer to the number density of neutral and ionised H, and t_{ion} and t_{rec} are the ionization and recombination timescales. For the IGM anytime after reionization occured, the gas is highly ionised, so $n_{HII} \sim n_H$ (i.e. almost the total H density) and so

$$\frac{n_{HI}}{n_H} = \frac{t_{ion}}{t_{rec}} \tag{7.2}$$

The ionisation time depends on the source of ionising photons, but the recombination time is straightforward

$$t_{rec} = \frac{1}{n_e \alpha_r} \tag{7.3}$$

where $\alpha_r \sim 4 \times 10^{-13} (T/10^4\,K)^{-0.7}\,\mathrm{cm}^3\ \mathrm{s}^{-1}$ is a recombination co-efficient and n_e the electron density.

Now lets consider a region occupied by a single ionising source, emitting $\dot{N}$ photons per second, and consider the change in the electron density at some distance r from the source

$$\frac{\partial n_e}{\partial t} + \nabla.(n_e \overrightarrow{v}) = \nabla.\frac{\dot{N}}{4\pi r^2}\overrightarrow{r} - \alpha_r n_e n_p \tag{7.4}$$

Here the first term is the change of electron density with time, the second accounts for the motions of electrons due to peculiar velocities (i.e. in and out of a region), we will set this equal to zero. The third term is just the number of ionising photons flowing through the shell at radius r and the final term is the number of recombinations. In an equilibrium system the left hand side will be zero (i.e. recombination rate is equal to the ionisation rate and the number density of electrons is constant). Then:

$$\nabla.\frac{\dot{N}}{4\pi r^2}\overrightarrow{r} = \alpha_r n_e n_p \tag{7.5}$$

Now if we integrate this equation over some volume, taking use of the divergence theorem then we get

$$\int \frac{\dot{N}}{4\pi r^2} dS = \frac{4\pi}{3} r_s^3 n_e n_p \alpha_r \tag{7.6}$$

where we have used the divergence theorem to change to an integral over a surface (i.e. $\int \nabla.F dV = \int F dS$). We can then re-arrange this to get the formula for the radius of a Stromgren sphere.

$$r_s = \left(\frac{3\dot{N}}{4\pi\alpha_r n_e n_p}\right)^{1/3} = \left(\frac{3\dot{N}}{4\pi\alpha_r n_H^2}\right)^{1/3} \tag{7.7}$$

Since $n_e = n_p = n_H$ in a highly ionised system.

In principle then each ultraviolet-luminous (i.e. young, massive) star will ionise a bubble (or *Strömgren Sphere*) around itself that has this radius. However it isn't that simple. The ionised region will have the same number density of protons as the rest of the interstellar medium, but a higher number density of electrons. It will also, because it is ionised, be much hotter than its surroundings. This means its pressure will be far higher than the pressure around it, and it will expand.

Balancing the internal and external pressure of the sphere requires

$$2n_{H,in}T_{in} = n_{H,out}T_{out} \tag{7.8}$$

or

$$n_{H,in} = n_{H,out}\frac{T_{out}}{2T_{in}} \tag{7.9}$$

If we substitute this in above we get a new expression for the radius of the Strömgren sphere in equilibrium (i.e. after it expands).

$$r_s = \left(\frac{3\dot{N}}{\pi\alpha_r n_{H,out}^2}\right)^{1/3}\left(\frac{T_{in}}{T_{out}}\right)^{2/3} \tag{7.10}$$

Since the temperature inside the Stromgren sphere may be 10^4 and outside perhaps $< 100K$ (i.e. the temperature of the cooling gas), the sphere growth due to pressure can be large (e.g. a factor of 20-50). Hence, when one has many ionizing sources turning on and each creating Strömgren spheres they will tend to grow and merge together.

7.3 The expanding Universe

The above would capture much of the important physics if all of this was taking place against a static background. However, it is not. Particularly at high redshift, in the epoch where we believe reionization occurred, the Universe is expanding and so the $\partial n_e/\partial t$ term is not going to be always zero.

We must therefore solve this equation in an expanding Universe. Remember:

$$\frac{\partial n_e}{\partial t} = \nabla.\frac{\dot{N}}{4\pi r^2}\overrightarrow{r} - \alpha_r n_e n_p$$

where we are still ignoring peculiar velocities and dealing with a homogeneous isotropic universe.

Now we can try to integrate this over a volume as before.

$$\frac{\partial}{\partial t}\int n_e dV = \int \nabla.\frac{\dot{N}}{4\pi r^2}\overrightarrow{r}dV - \int \alpha_r n_e n_p dV \tag{7.11}$$

Again we can convert the volume integrals on the right to a surface integral, and treat α_r as a constant if T is constant. We will also use volume average quantities e.g.

$$\langle x \rangle = \frac{1}{V} \int x dV \tag{7.12}$$

Now lets re-write the integral equation above with these quantities:

$$\langle n_e \rangle \frac{dV}{dt} + V \frac{dn_e}{dt} = \dot{N} - \alpha_r \langle n_e n_p \rangle V \tag{7.13}$$

Now we can switch to an expanding Universe. Since we're considering particle densities, these will scale with the Universe as $n_e \propto 1/a^3$ and

$$\frac{1}{\langle n_e \rangle} \frac{d \langle n_e \rangle}{dt} = \frac{d \ln \langle n_e \rangle}{dt} = -3 \frac{\dot{a}}{a} \tag{7.14}$$

This then gives us

$$\frac{dV}{dt} - 3VH(t) = \frac{\dot{N}}{\langle n_e \rangle} - \alpha_r \frac{\langle n_e n_p \rangle}{\langle n_e \rangle} \tag{7.15}$$

Now we consider the recombination timescale for a fully ionised gas ($n_e = n_p$).

$$t_{rec} = \frac{1}{\alpha_r \langle n_e \rangle} \tag{7.16}$$

and we can define an effective recombination timescale (=1/recombination rate)

$$\bar{t}_{rec} = \frac{\langle n_e \rangle}{\alpha_r \langle n_e n_p \rangle} = \frac{1}{C} t_{rec} \tag{7.17}$$

where we have defined a *clumping factor* C

$$C = \frac{\langle n_e n_p \rangle}{\langle n_e \rangle^2} \tag{7.18}$$

We have taken a while to get to this, but it is very important. If the gas distribution in the Universe is clumpy (and it is) then in the dense, clumped regions recombinations will be a lot more efficient because the scale with the square of the density. If we now re-write equation 7.15 using this effective recombination timescale then we get

$$\frac{dV}{dt} = \frac{\dot{N}}{\langle n_H \rangle} - V \left[\frac{1}{\bar{t}_{rec}} - 3H(t) \right]$$

The useful thing about this is that we have separated how the volume of the ionised region will change with time into components that depend on the source (first term on left hand side), and those that do not. We can only ignore the expansion of the Universe in regions that are likely to be very dense, and hence have short recombination times. i.e. when

$$\frac{1}{\bar{t_{rec}}} >> 3H(t) \tag{7.19}$$

It turns out we statisfy this scenario when

$$C\left(\frac{1+z}{10}\right)^{3/2}\left(\frac{n_H}{\bar{n_H}}\right) >> 2.7 \tag{7.20}$$

(For the moment you will just have to take my word for this!)

Our ionization equation is not necessarily trivial to solve in all scenarios, but if we look at a scenario where photons are copious (i.e. many photons per hydrogen atom) then recombinations don't matter. We can solve just the first term

$$\frac{dV}{dt} = \frac{\dot{N}}{\langle n_H \rangle} - \frac{V}{\bar{t}_{rec}}$$

to get

$$V(t) = \frac{\dot{N}\, t_{rec}}{C\, \langle N_H \rangle}\left[1 - \exp\left(-t/\frac{t_{rec}}{C}\right)\right] \tag{7.21}$$

So for $t << t_{rec}/C$ we can take a Taylor expansion of e^{-x} and have

$$V(t) = \frac{\dot{N}}{\langle n_H \rangle} t \tag{7.22}$$

which is valid at early times, when the rate of production of ionising photons is much faster than recombinations.

For later times, when $t >> t_{rec}/C$, the exponential term tends to zero and

$$V(t) = \frac{\dot{N}}{\langle n_H \rangle}\frac{t_{rec}}{C} \tag{7.23}$$

7.4 Filling factors and recombinations

Since we are worried about the Universe as a whole, we are not just interested in the Strömgren spheres blown by individual sources, but in how they start to fill the Universe.

It is common to define a volume-averaged cosmic filling factor Q which is simply

$$Q = \frac{V}{V_{tot}} \tag{7.24}$$

where V is the total volume of ionized regions and V_{tot} of the Universe as a whole. We can substitute this back into equation 7.15 to get

$$\frac{dQ}{dt} = \frac{\dot{N}}{V_{tot}\, \langle n_H \rangle} - \frac{Q}{t_{rec}} \tag{7.25}$$

If we look at early times, again where $t << t_{rec}/C$ we then require that

$$\frac{\dot{N}}{V_{tot}} t = \langle n_H \rangle \tag{7.26}$$

Since the left hand side of this equation is also a density (this time a photon density), we see that each hydrogen atom needs just one photon to ionize it - in other words, at these early times recombination is unimportant.

If we now look at later times $t >> t_{rec}/C$ then we find that

$$Q(t) = \frac{\dot{N}}{V_{tot} \langle n_H \rangle} \frac{t_{rec}}{C} \tag{7.27}$$

and

$$\frac{\dot{N}}{V_{tot}} = \frac{\langle n_H \rangle}{t_{rec}/c} \tag{7.28}$$

Since this depends on recombinations, calculating the filling factor at this point requires us to balance the recombination in the clumpy medium. This may actually be quite hard at high redshift.

So we have a rough description of how the Universe may be reionized:

Individual sources will switch on and blow their own Strömgren spheres. These will gradually overlap until most/all of the Universe is reionized. In over-dense (clumpy) regions it is possible that some hydrogen escapes reionization, or that it subsequently recombines. This provides the gas clouds which give rise to the Lyman-alpha forest. However, for most of the low density IGM, the expansion of the Universe takes over. This reduces the density (and temperature) sufficiently that the Universe never recombines and only the initial ionization is important.

As a result, while the photon production rate in the current Universe would be far too low to rapidly reionize it (were it to be neutral initially), this doesn't matter, because the recombination timescale in the IGM is so long (bear in mind that the critical density is only about 1 proton per cubic metre, and only 4% of the Universe consists of baryonic matter anyway). Many fewer photons are required to maintain the ionization state at low redshift than create it at in the distant Universe.

7.5 Observational Constraints on Reionization

We now turn our attention to how we might actually map the process of reionization. This is important: we would like to know how quickly it happened and when. Was it a one way process, or did the Universe reionize, then recombine then ionise again? All of these questions require some means of detecting and mapping the process. There are several possibilities here, with different pros and cons.

7.5.1 The Gunn-Peterson trough

We have already seen how the Lyman-alpha forest impacts the appearance of high redshift galaxies, but let's now consider what might happen as not only the individual absorbers, but the Universe at large becomes neutral. In particular, lets imagine that some light (F) passes through a region of length ΔL and hydrogen column density n_H. The change in its intensity will be given by

$$\Delta F = -F n_h \sigma \Delta L \tag{7.29}$$

where σ is the cross section for absorption by hydrogen. If we take the increments as infinitiseimal we can deal with this as a differential.

$$dF = -F n_h \sigma dL \tag{7.30}$$

which can be integrated to provide the relation:

$$F_{obs} = F_0 \exp\left(-\int n_H \sigma dL\right) = F_0 \exp\left(-\tau\right). \tag{7.31}$$

The integral here in the exponential is just the optical depth, τ (density $\times$ cross section $\times$ length).

The cross section here is non-trivial since it depends predominantly on the wavelength (or frequency) of the incoming light, in the rest-frame of the hydrogen that it interacts with. Typically we can write

$$\sigma(\nu) = \frac{\pi e^2}{m_e c} f \phi(\nu) \tag{7.32}$$

where f is the oscillator strength and $\phi(nu)$ is what is known as the line profile function, it essentially impacts the probability of exciting a line as a function of frequency. Since we will only excite lines if we hit a particular transition we can write this as

$$\sigma(\nu) = \sigma_0 \nu_0 \phi(\nu - \nu_0)$$

or

$$\sigma(\nu) = \sigma_0 \nu_0 \delta_D(\nu - \nu_0)$$

where the line profile is effectively a delta function (true for fairly low column densities).

In principle we care about all of the lines in hydrogen, but as most interstellar or intergalactic H lies in the ground state, the most common feature to worry about will be the Lyman-alpha transition at 1216Å in the rest frame, or in an expanding universe $\nu = \nu_{obs}(1+z)$ (note this is the opposite of wavelength $\lambda_{obs}\ (1+z) = \lambda$, since redshift stretches wavelength and hence compresses frequency). In this case

$$\frac{d\nu}{dz} = \nu_{obs} \tag{7.33}$$

So starting with the optical depth we had before

$$\tau = \int_0^1 n_H \sigma(\nu) dL \tag{7.34}$$

we can rewrite it as

$$\tau(\nu_{obs}) = \int_{\nu_{obs}}^{\nu_{obs}(1+z)} n_H(z) \sigma(\nu) \left|\frac{dL}{dz}\right| \frac{d\nu}{\nu_{obs}} \tag{7.35}$$

If we now utilise the delta function approximation for the cross section we can obtain.

$$\tau(\nu_{obs}) = \frac{\sigma_0 \nu_0}{\nu_{obs}} n_H(z) \left| \frac{dL}{dz} \right| \tag{7.36}$$

at frequencies where the absorption occurs. This is useful because it tells us that the level of absorption contains information about the density of neutral hydrogen along the line of sight to the source of the light. To put this into a more cosmological footing we can write:

$$dL = cdt = c\frac{dt}{da}\frac{da}{dz}dz = -\frac{c}{H(z)}\frac{dz}{(1+z)} \tag{7.37}$$

since $a = 1/(1+z)$. Now we can re-write the optical depth again as

$$\tau(\nu_{obs}) = \sigma_0 \left(\frac{c}{H_0}\right)\left(\frac{H_0}{H(z)}\right) n_H(z) \tag{7.38}$$

Now what we are actually interested in is not just the H column density (n_H), but the neutral hydrogen column density (n_{HI}). So lets look at that

$$n_{HI} = \frac{n_{HI}}{n_H} n_H = \frac{n_{HI}}{n_H}\frac{n_H}{\bar{n_H}}\bar{n_H} \tag{7.39}$$

The advantage of this is that the mean H density is quite well known. It is 75% of the baryons. We will call this fraction X, and rewrite n_H in terms of this and the cosmic baryon density

$$\bar{n_H} = \frac{\bar{\rho_B} X}{m_p} = \frac{X}{m_p}\Omega_B \rho_{crit}(1+z)^3 \tag{7.40}$$

So, if we put all this together with some numbers we find that.

$$\bar{n_H} = 1.2 \times 10^{-5}\mathrm{cm}^{-3}\left(\frac{X}{0.75}\right)\left(\frac{\Omega_B}{0.023}\right)\left(\frac{1+z}{4}\right)^3 \tag{7.41}$$

or in terms of optical depth.

$$\tau(\nu_{obs}) = 5.2 \times 10^5 \frac{n_{HI}}{n_H}\frac{n_H}{\bar{n_H}}\left(\frac{H_0}{H(z)}\right)\left(\frac{X}{0.75}\right)\left(\frac{\Omega_B}{0.023}\right)\left(\frac{1+z}{4}\right)^3 \tag{7.42}$$

At high-z, where

$$H(z)^2 \approx H_0^2 \Omega_m (1+z)^3 \tag{7.43}$$

This becomes

$$\tau(\nu_{obs}) \approx 1.7 \times 10^5 \frac{n_{HI}}{n_H}\frac{n_H}{\bar{n_H}}\left(\frac{\Omega_m}{0.147}\right)^{-1/2}\left(\frac{X}{0.75}\right)\left(\frac{\Omega_B}{0.023}\right)\left(\frac{1+z}{4}\right)^{3/2} \tag{7.44}$$

The important thing to note here is that the pre-factor in the optical depth is huge. This means that as soon as there is almost any neutral hydrogen at a given redshift, the optical depth for hydrogen

absorption is very large. For individual gas clouds, this is the cause of the Lyman$-\alpha$ forest, but if the IGM as a whole is neutral then what we actually get is a trough – total absorption at wavelengths corresponding to the redshift where the IGM is. All of the light is blocked.

Observations have now found evidence for this trough, called the Gunn-Peterson trough in the spectra of quasars at $z > 6.4$. It is very sensitive to the end of reionization, since only a small neutral fraction is needed to cause it.

7.5.2 Galaxy observations

We have treated the absorption line profile resulting from Lyman-α transitions as a delta function. In fact it is a more complex function, which depends on random velocities of atoms in the absorber and on the column density of neutral hydrogen - in other words the line has 'wings' which are far broader (span a larger wavelength range).

In fact, we find that,

$$\tau(\Delta\lambda) = \frac{\tau_0 R_\alpha}{\pi}\left(\frac{\Delta\lambda}{\lambda}\right)^{-1}$$

where $\Delta\lambda$ is the offset from the line central wavelength λ, τ_0 is the Gunn-Peterson optical depth and R_α is a constant describing the decay of the Lyman-alpha line.

If neutral hydrogen exists close to the source, the effect of this damping wing can be seen extending longwards (to the red) of the source's Lyman-α rest frame wavelength. The effect of this can be seen in a number of contexts:

1. A drop in the number of Lyman break galaxies with Lyman-alpha emission lines at $z > 7$ suggests that some of the Lyman-alpha emission is being absorbed by a red damping wing.
2. The number density of narrowband-selected Lyman-alpha emitters (LAEs) drops off sharply between $z \sim 5.8$ and $z \sim 6.6$.
3. The red damping wing can be observed (and modelled) directly against individual spectra of high redshift quasars (e.g. Mortlock et al 2011) and GRBs (e.g. Tanvir et al 2009, Totani et al 2014).

Again, this effect is sensitive to relatively small neutral fractions (up to a few tenths in the case of Lyman-alpha emission), and so is primarily sensitive to the end stages of reionization.

7.5.3 CMB observations

An alternative route to map reionization occurs when we use observations of polarisation signals in the CMB. The CMB free streams to us because at the point of its creation the Universe transitions from being opaque to transparent. The reason for this transition is the absence of free electrons. This happens at $z \sim 1000$, so the density of free electrons at this point is high. If reionization happens at $z \sim 10$ then the Universe is around a million times less dense. However, these free electrons are still a source of scattering for the CMB photons. They will damp the spectrum, providing a source of power at large radii (beyond the first peak) and importantly they will also be polarised. So for this additional source of absorption, and as above, we have

$$\tau(z) = 0.063\Omega_B \int_0^z \frac{H_0}{H(z)}(1+z)^2\left(\frac{\rho}{\bar{\rho}}\right) f_e(z)dz \tag{7.45}$$

where ρ here is the density of the intergalactic medium and f_e the electron fraction. This can be measured from by fits to both the temperature and polarization power spectra of the CMB.

First we should note that this cross section is very sensitive to the electron fraction, rather than the neutral fraction. In other words while the Gunn-Peterson trough probes the last bit of reionization, the CMB scattering probes the first bit. Again in appropriate units, and for instantaneous reionization at $z = 10$ we have.

$$\tau(z_{ri}) \approx 0.08 \left(\frac{1 + z_{ri}}{1 + 10} \right)^{3/2} \tag{7.46}$$

So about 10% of photons are scattered after the reionization process if it occured at $z = 10$.

In fact this *optical depth to reionization* from the CMB has been subject to some uncertainty since it depends on the very large scales and polarization, both of which are difficult to measure accurately. The latest results from Planck (2015) suggests that $\tau = 0.066 \pm 0.016$, slightly different from the $\tau = 0.088 \pm 0.015$ reported by WMAP.

Assuming that this corresponds to a fairly rapid transition with a width $\Delta z_{\rm re} = 0.5$, the optical depth suggests a reionization redshift $z_{\rm re} = 8.8^{+1.3}_{-1.2}$.

7.5.4 21 cm emission

Another route we have is to use the hydrogen itself as means of measurement. This would be possible if we could map 21cm emission. This arises in neutral hydrogen when the electron and proton flip from parallel to anti-parallel spin alignments. It is used locally to map rotation curves in galaxies, but can equally map neutral hydrogen in the intergalactic medium. The optical depth here is given by

$$\tau = 0.3 T_S^{-1} \frac{n_{HI}}{n_H} \left(\frac{\rho}{\bar{\rho}} \right) \left(\frac{1 + z}{1 + 9} \right)^{3/2} \tag{7.47}$$

The spin temperature, T_S, and in particular its relation to the CMB temperature $T_{CMB} = 2.73 \times (1 + z)$ K, is important because it determines if one should look for the 21cm line as an emission feature or an absorption feature against the CMB (i.e. is energy needed to flip the electron, or is it likely already pushed into a high state by CMB photons and likely to release its energy as photons). For most reionization models, it should appear as an emission feature.

We have yet to use 21cm lines to map out reionization. Technically it is highly challenging. The Earth's ionosphere blocks most of the signal at metre wavelengths (i.e. redshifted 21cm), and the foregrounds are astonishingly strong, but the method offers particular promise. This is because we can quite literally map out reionization with it. Because the line transition has a fixed wavelength we can determine the redshift of detected signal, and also of the transition where 21cm emission becomes absent when the region is reionized (no electrons). More importantly it might be possible to work out the topology - i.e. did reionization occur through a few large bubbles that overlapped or many small ones. This is important to understanding the sources of reionization. The 21cm emission also maps the sky at all locations, rather than just those with stars/quasars etc, which might not be typical of the IGM as a whole.

Several current and future telescopes, notably the Square Kilometer Array (SKA, due on-sky from about 2020), aim to perform this experiment. Initially they'll focus on determining a power spectrum (like that of the CMB) at each redshift, before maps become possible.

7.5.5 Current best observational constraints

Current best constraints suggest that hydrogen reionized only once, in a fairly smooth process. The CMB constraints imply that the bulk of the process occured at $z \sim 9$, assuming a rapid progress, but are largely sensitive to the start of reionization. The Lyman-alpha forest and G-P trough results tell us that the process did not complete until $z \sim 6$, but are only sensitive to the final stages of the process. Constraints from LAE densities and the red damping wing of Lyman-alpha as seen in GRB and Quasar spectra lie in between these limits.

Thus we infer a rapid evolution from mostly-neutral to mostly-ionized through a process of overlapping ionized regions, between $z = 6$ and $z = 8$. Given the redshift, the energy requirements and the density of high redshift galaxies, the high energy photons kicked out by young stars in star-forming galaxies are believed to be the main sources of reionization. Hydrogen 21cm line mapping will be required to confirm this picture.

7.6 Helium Reionization

Finally, we can develop insights by considering parallels to a very similar process that occured at lower redshifts where it is easier to observe.

While hydrogen is the most common element in the Universe, helium is the second most abundant. Helium has two electrons. Ionizing the first requires a potential energy of 26.6 eV, while ionizing the second requires 54.4 eV (as opposed to the 13.6 eV needed by hydrogen). Since this corresponds to a higher temperature than needed for hydrogen, the first helium atoms formed before the first hydrogen atoms at recombination.

However that means it also requires more energy input to reionize it, and this was not possible in the early Universe. In fact, as studies of the Helium Lyman-α forest (i.e. absorption by neutral helium in its ground state) show, the process of first helium reionization (He I $\rightarrow$ He II) did not occur until $z \sim 2.7$, at which point a rapid change in the Helium optical depth is observed (Worseck et al 2011).

Both the energy requirements for Helium reionization and its redshift suggest that this process probably wasn't driven to early star formation (as is believed to be the case for Hydrogen reionization), but instead due to the much harder spectrum produced by quasars. Nonethless, studies of its topology and progression provide insights into how we interpret the first EoR at $z \sim 9$.

Topic 8

The history of star formation

Objectives for this lecture

- To understand how and when the majority of stars in the Universe were formed
- To understand how they enriched the chemical composition of the Universe

Further Reading

- Loeb & Furlanetto, Ch 8
- Madau & Dickinson (2012)
- Bouwens et al (2011 - *required reading*.

Expected length: 2 lectures

8.1 Balancing the Ionizing Photon budget

The reionization history of hydrogen tells us that the first sources of UV-light switched on and reionized the Universe somewhere between 400-900 million years after the big bang. In order to reionize the Universe it was necessary to create *at least* one photon per baryon... many more if recombinations are considered.

Since we can now identify UV luminous sources at these distances we can therefore ascertain if it is likely they can reproduce the necessary ionising photon budget.

8.1.1 The ionizing photon requirement

One side of the ionising photon budget is relatively easy to constrain, we know the baryon density, and so can simply set the number of photons required to be equal to this to first order. If reionization is essentially instantaneous (as often assumed in CMB analyses) then there is little time for recombinations. In practice, the filling factor is given by

$$\dot{Q}_{HII} = \frac{\dot{N}_{ion}}{\langle n_H \rangle} - \frac{Q_{HII}}{\langle t_{rec} \rangle} \tag{8.1}$$

(e.g. Duncan & Conselice 2015). The last term is based on recombination, and so depends on complex issues such the clumpiness. The first term just refers to the production of ionising photons.

We can estimate the photon output of sources in the distant Universe from observations as

$$\dot{N} = f_{esc}\eta_{ion}\rho_{UV} \tag{8.2}$$

Here ρ_{UV} is a UV luminosity at some point redward of Lyman-alpha where direct observations can be made. The factor η_{ion} then refers to the fraction of that light which is produced as ionising photons (e.g. if we assume a single black body that can be calculated directly, but not for an amalgamation of stars unless we assume some stellar population model). The factor, f_{esc} refers to the *escape fraction* of Lyman-continuum photons (shortwards of 912Å), and is the most uncertain.

Ideally to estimate the escape fraction you would observe a galaxy at wavelengths shortward of Lyman-alpha, but the Gunn-Peterson trough makes this essentially impossible at high-z. In practice the escape fraction might be reduced by dust, scattering off free electrons etc. In the local Universe it is apparent that f_{esc} is indeed small, perhaps only around 0.1-10%. The early Universe differs from the local Universe significantly: there is far less gas and dust because of the absence of heavy elements, but the Universe is also much denser, and the first stars are embedded in the densest regions. It is far from clear that Lyman-continuum photons can readily escape the galaxy-level environments in which it is produced and then go on to reionize the Universe. Together with the clumpiness this represents a key unknowns in the availability of ionising photons, and the rates of recombination.

Since the sources responsible for reionization are believed to be star forming galaxies, and their ultra-violet continuum luminosity scales in proportion to the number of young stars, it is often convenient to express the photon density required for reionization as a star formation rate:

$$\dot{\rho}_{SFR} = 0.012\mathrm{M}_{\odot}\mathrm{yr}^{-1}\mathrm{Mpc}^{-3}\ \left(\frac{(1+z)}{8}\right)^3\left(\frac{C/3}{f_{esc}/0.2}\right)\left(\frac{T}{10^4\mathrm{K}}\right)\left(\frac{10^{53.3}}{\dot{N}}\right) \tag{8.3}$$

8.1.2 Ionizing photon production

If we want to work out how do do this we need to know how many galaxies are emitting ultraviolet light per unit volume, and how bright each one is. This distribution is characterised by the galaxy luminosity function. This is usually parameterised as a *Schecter function*:

$$\phi(L)dL = \left(\frac{\phi_*}{L_*}\right)\left(\frac{L}{L_*}\right)^{\alpha}\exp\left(-L/L_*\right)dL,$$

Here ϕ_* is the characteristic number density of a population of galaxies, and L_* is its characteristic luminosity. α is a constant which describes the *faint-end slope*, i.e. the distribution tends to a power law with slope α at low luminosities.

This is reasonably well constrained by deep field observations across a broad range of redshifts, although substantial corrections always have to be made for incompleteness and contamination of photometric galaxy samples. It is has become clear that the number density of galaxies declines sharply with redshift, particularly at $z > 6$, and that the faint end slope is steep, $\alpha \approx -2$.

Because this is determined using UV light we can actually convert from a luminosity to a star formation rate for a given luminosity of galaxy. e.g.

$$SFR(UV)(M_{\odot}yr^{-1}) = 1.4 \times 10^{-28}L_{\nu}(ergs\,s^{-1}\,Hz^{-1}) \tag{8.4}$$

where the calibration is determined from local star forming regions (Kennicutt 1998). Typically an L_* galaxy will have a star formation rate of a few solar masses per year.

By integrating the luminosity function, we can work out the total number of galaxies

$$N = \int \phi(L) dL = \int \left(\frac{\phi_*}{L_*}\right) \left(\frac{L}{L_*}\right)^{\alpha} \exp\left(-L/L_*\right) dL, \tag{8.5}$$

and if we multiply each galaxy by its luminosity we can obtain the total luminosity (per Mpc^{-3})

$$S(L) = \int L\phi(L) dL = \int (\phi_*) \left(\frac{L}{L_*}\right)^{\alpha+1} \exp\left(-L/L_*\right) dL, \tag{8.6}$$

There are two critical things to determine before we can evaluate this equation. The first is the limits of the integral, the second is the value of the faint end slope. In the local Universe the faint end slope is fairly shallow $\alpha \sim -1.5$ shallower. This means that defining a total integral between $0 \rightarrow \infty$ is plausible. This is an awkward integral, and is often expressed in terms of the gamma-function

$$\int_0^{\infty} \phi(L) dL = \phi\Gamma(\alpha + 1) \tag{8.7}$$

or

$$\int_0^{\infty} L\phi(L) dL = \phi L_*\Gamma(\alpha + 2) \tag{8.8}$$

where the gamma function is $\Gamma(n) = (n-1)!$.

If the lower limit isn't zero, then the integral would be $\phi L_*\Gamma(\alpha + 2, L_0/L_*)$, which is the incomplete gamma-function, with a rather more complex form.

$$\Gamma(n, x) = (n-1)!\, e^{-x} \sum_{k=0}^{k=n-1} \frac{x^k}{k!} \tag{8.9}$$

This is an awkward integral to work with, and since galaxies can't have a luminosity of zero it's one we're going to have to consider. Fortunately we are often only considered with the fainter end of the integral, below the exponential cut-off. In which case we can omit the bright end (which just contributes a constant to the luminosity density) and just integrate the power-law component. e.g.

$$S(L) = \int L\phi(L) dL = \int_{\mathrm{Lmin}}^{\mathrm{Lmax}} (\phi_*) \left(\frac{L}{L_*}\right)^{\alpha+1} dL + A, \tag{8.10}$$

where A is the integral of the bright end, which is a relatively small contribution and doesn't change.

However, we do need to know the limits here. We shouldn't go all the way to zero, because this would imply a contribution from stars that were in fact, smaller than stars. Indeed, it is common to set a limit around 10^{-5} M$_\odot$ yr^{-1}, since these regions are probably sufficiently small that a single supernova would expel all the gas, and so stop further star formation. In fact, we already know we can't integrate all the way to zero for steep faint end slopes (as indicated at high-z) simply because they diverge, giving infinite luminosity density, and as the night sky is dark this is clearly unphysical.

8.1.3 Balancing the budget

So, if we integrate the luminosity density function and then convert the integral into a star formation rate, we will be able to ascertain whether the galaxies that we observe can reionize the Universe.

The first thing we should do is consider the galaxies that we can directly see. When we observe the high redshift Universe we are only looking at the tip of the luminosity function (not much below L* at the highest redshift). Therefore it is clear that at these redshifts there is insufficient photon density, *in the galaxies we observe* to reionize the Universe. Therefore, if galaxies are responsible then it must be the faint galaxies that do it. For a sensible faint end limit, this implies a faint end slope at least as steep as -2. The latest Planck results have helped here, bringing the EoR down to redshifts where star forming galaxies are more abundant.

In fact, current research suggests that star forming galaxies can match current reionization constrants if:

1. the faint end is steep;
2. we integrate down to faint luminosities;
3. we have a large number of ionizing photons per star formed;
4. we have a high ionizing escape fraction.

While one or more of these conditions may not be met, the best current constraints at least put them within the realms of possibility.

If it proves that galaxies are not responsible then we must look for other possibilities, such decaying particles, mini accreting black holes etc. One thing that is (at least apparently) clear, is that because the number density of bright quasars declines so steeply before we enter the epoch of reionization, the quasars are not sufficient to reionize the Universe.

8.2 Star formation density history

So, we have looked at how the first structures in the Universe formed, how they are found, and how the star forming galaxies within them impact the Universe around them. The final part of the course is to put this into context, and look at how these first objects evolve through to the Universe that we see today.

We have already looked at the volume-averaged star formation at one redshift – the epoch of reionization. We can now consider how it changes over cosmic time. If we track luminosity functions (and particularly ultraviolet luminosity functions) across the Universe, we can map when the stars in the Universe formed, and from the shape of the UV luminosity function can also infer in which types of galaxies the stars were forming.

The shape of the star formation density history is quite striking (see recent review by Madau & Dickinson 2014). This diagram, known as the *Star Formation Rate Density (SFRD)* diagram or the *Madau-Lilly* plot, is the result of decades of hard work by hundreds of astronomers. Working backwards from the current day to earlier times, the volume-averaged star formation rate shows a strong rise by at least an order of magnitude from the preset day to $z \sim 1$. It is fairly flat from there out to $z \sim 5$ and then appears to drop rapidly. This drop is problematic for reionization (i.e. there are fewer stars forming at early times where we need them most), but also suggests that star formation took a while

to get going. It did not start immediately after the big bang, or even immediately after the first stars were capable of forming.

The physical processes which govern this shape are, of course, complex but we can develop a broad picture of them. In the distant Universe, $z > 5$, we're seeing the star formation rate tracing galaxy formation through collapse of density perturbations. As the Universe ages, more and more of the primordial structures reach sufficient density to support star formation, while the biggest structures are feeding on hydrogen gas from the interstellar medium and so able to keep forming stars at a relatively high rate. At the peak of cosmic star formation, around $z \sim 1 - 5$, we are seeing a number of competing effects. The most massive galaxies are now depleting their gas supplies (i.e. running out of fuel) and can no longer effectively accrete it from the sparse intergalactic medium. At the same time many of them will be undergoing mergers (see next lecture), where one or more galaxies fall into the same potential well and gradually lose their individual identities. This tends to ram together what gas remains in shocks and trigger an intense star burst that both pushes up the star formation density and blows out any residual gas through radiation pressure. To add to the excitement, most galaxies will by now have a sizeable supermassive black hole at their nucleus. The disruption of gas orbits caused by galaxy interactions tends to send material in towards the centre of the galaxies and trigger AGN activity - which again blows out any remaining material. All of these feedback mechanisms mean that at the same time as they form stars rapidly, massive galaxies are ruling out any substantial future star formation. Below $z \sim 1$, the most massive galaxies are quenched, while smaller galaxies are likely to be in regions which are already depleted in gas density, and which gradually run out of fuel. The accelerating expansion of the Universe means that accretion of new material from the intergalactic medium becomes more difficult, and we have seen a gradual decline in newly formed stars towards the present day, and they form further apart – both phenomena which lead to a decline in the star formation rate density.

The star formation rate density diagram in relatively simple to directly infer from the UV luminosity density (i.e. looking for galaxies in the rest-frame infrared and adding up the total light in a given volume), but we must bear in mind the fact this observed density is only a lower limit. We can correct this for the faint end (below observational limits), but importantly we must also correct it for dust extinction (which removes a large number of UV photons). The dust extinction terms are somewhat well calibrated a low redshift, but uncertain in the distant Universe. It may be that at high-z there is little dust (as might be expected if it hasn't had time to form), or it may be that we simply can't observe many dusty galaxies to infer the dust.

When we look at the luminosity functions we also see that there is a significant change in their shapes. L* becomes brighter, and the faint slope becomes shallower as we move to lower redshift. This means that brighter, L* galaxies come to dominate the UV luminosity at $z \sim 1 - 3$, where sub-L* galaxies dominated at early times. The peak of the star formation rate history indicates that a significant fraction of the total star formation was taking place in very massive systems at intermediate redshifts. These will evolve in the massive galaxies that we observe today. However a similarly large fraction of *their* stars formed earlier, in many small progenitors, supporting the heirarchical merger scenario.

Interestingly, as we move towards the current epoch, much of the star formation again appears to move away from the most massive galaxies (e.g. elliptical galaxies) that have little star formation. Instead most local star formation is in small galaxies – a phenomenon known as *downsizing*. Indeed, it is these dwarf galaxies that have the highest specific star formation rate (star formation rate/ stellar mass of galaxy) in the local Universe. This is probably because we are running out of gas from which to form stars. In fact, given the ratio of gas to stars in galaxies we see today, it is likely that 95% of all the stars that will ever form have already formed.

The integral of the star formation density history gives the history of stellar mass density - i.e. how much of the current mass in stars was in place by a given redshift (modulo uncertainties on how many of the stars have returned their mass to the ISM). This suggests that at least half of the current stellar

mass was in place by $z \sim 1.5$.

8.3 The history of chemical enrichment

As stars formed, lived and died in the distant Universe, they processed pristine neutral gas into an intergalactic medium that is progressively more and more enriched by heavier elements – metals.

This process is going to depend on a number of factors we have already encountered: the density of matter in stars ρ_*, the density of gas not in stars ρ_g and the volume-averaged star formation rate ψ. It will also depend on two factors specific to this process: the net metal yield, y, which gives the quantity of metals produced by supernovae per solar mass of star formation, and the return factor, R, which governs what fraction of this is rapidly returned to the IGM as opposed to locked within long lived stars and stellar remnants. Finally we need a way of keeping track of the metal content of the IGM. This is done using Z - the *metallicity mass fraction*, or just *metallicity*.

These parameters are linked by a series of differential equations:

$$\frac{d\,\rho_*}{d\,t} = (1-R)\psi, \tag{8.11}$$

$$\frac{d\,\rho_g}{d\,t} = -\frac{d\,\rho_*}{d\,t}, \tag{8.12}$$

and

$$\rho_g \frac{d\,\mathrm{Z}}{d\,t} = y(1-R)\psi \tag{8.13}$$

The first equation states that the change in stellar mass must account for the mass returned to the interstellar medium in any time interval, while the second states that the only reason for gas density to change is if the gas has been formed into stars (strictly only true for a closed-box model of a galactic halo, but we will assume that here). The third equation calculates the total mass of metals formed in a time interval (the gas density multiplied by the changing fraction of that density in metals, $\rho_g d\mathrm{Z}$), and states that this will be a fixed fraction of the change in stellar mass (i.e. the yield from SNe of stars in a given mass range).

If $\Delta\rho_*$ stars form , then a mass of $y\Delta\rho_*$ metals is formed. But there are also new stars being formed which soak up a mass $\mathrm{Z}\rho_*$ of metals at the current metallicity, Z. So the change in the total metal mass is the difference of these two terms:

$$\Delta(\mathrm{Z}\rho_g) = (y - \mathrm{Z})\Delta\rho_*$$

Substituting for $\Delta\rho_*$ the equations above,

$$d(\mathrm{Z}\rho_g) = (y - \mathrm{Z})(\psi(t)\,(1-R))dt$$

The metallicity mass fraction starts effectively from zero at t=0, so the left hand side is just $\mathrm{Z}(z)\rho_g(z)$, i.e the values at time of observation, so,

$$\mathrm{Z}(z)\rho_g(z) = \int_{t=0}^{t(z)} (y - \mathrm{Z(t)})(\psi(t)\,(1-R))dt$$

$$= (1-R)\int_{t=0}^{t(z)} (y - \mathrm{Z(t)})\psi(t)dt$$

$$= (1-R)\left[y\int_{t=0}^{t(z)} \psi(t)dt - \int_{t=0}^{t(z)} \mathrm{Z(t)}\psi(t)dt\right]$$

We can simplify this somewhat. The time integral of the star formation rate is just the total stellar mass formed by z, i.e. $\rho_*(z)_{\mathrm{formed}}$. The time integral of the star formation rate weighted by the metallicity, Z, gives the mean mass in metals of all stars formed to date $\rho_*(z)_{\mathrm{formed}}\langle \mathrm{Z}(z)\rangle$. So,

$$\mathrm{Z}(z)\rho_g(z) = (1-R)\rho_*(z)_{\mathrm{formed}}(y - \langle \mathrm{Z}(z)\rangle),$$

and since (1-R) is the fraction of mass retained in stars,

$$(1-R)\rho_*(z)_{\mathrm{formed}} = \rho_*(z).$$

Which means that the volume-averaged metallicity mass fraction, Z(z) at redshift z is given by:

$$\mathrm{Z}(z) = \frac{\rho_*(z)(y - \langle \mathrm{Z}_*(z)\rangle)}{\rho_g(z)},$$

where $\langle \mathrm{Z}_*(z)\rangle$ is the mean metallicity mass fraction of stars born prior to redshift z.

This is useful because if we know the density of stars and gas at a given redshift, we can constrain the other parameters.

If we can measure the metallicity directly, for example through quasar absorption line studies, we can determine $\langle \mathrm{Z}_*(z)\rangle$ and thus consider constraints on the required star formation history. This tells us how metal content in the universe evolves over time. All our uncertainty regarding the previous star formation history is absorbed into the term $\langle \mathrm{Z}_*(z)\rangle$ - i.e. the mean metallicity fraction in stars born to date. Of course, a full solution requires solving the paired differential equations, which is usually done numerically.

It's worth mentioning however that we've glossed over some pretty big unknowns. The metal yield of supernovae and the return fraction of metals into the interstellar medium are both difficult to measure directly and generally fixed from theory or simulations. For normal stellar populations a typical supernova yield $y = 2\%$, while $R \approx 0.4$, but both are dependent on the mass function and timescale considered. We've assumed that neither of these have changed significantly over time. In fact, there was likely a very different yield and return fraction from Pop III stars, suggesting that we ought to split this chemical evolution into two phases: before and after the Pop III/II transition. We also need to bear in mind timescales for the recycling term - only stars in a certain mass range will return their excess mass to the interstellar medium in any given timescale (high mass stars for short timescales, both high and low mass in long timescales).

Finally, and perhaps most crucially for galaxy evolution, we need to consider the fraction of the metals that not only escape being trapped in stars but also escape from the potential well of their galaxies and into the intergalactic medium. This is parameterised by the *mass loading factor*, η, which is simply the mass carried by winds per unit star formation. It is going to depend on two factors: the rate of supernovae and other energy releases to drive interstellar material out of the galaxy as winds, and the fraction of those winds comprised of heavy metals.

We can obtain the mass in stellar winds from a conservation of momentum argument:

$$\dot{M}_{\rm wind} = \frac{dp/dt}{v_\infty} \sim \dot{M}_* \left(\frac{v_{\rm wind}}{v_\infty} \right)$$

where dp/dt is the rate of transfer of momentum from supernovae and starbursts (which is proportional in turn to the star formation rate $\dot{M}_*$ and v_∞ is the escape velocity of the galactic halo. $v_{\rm wind}$ is the velocity of the wind driven per unit star formation and is typically around $300\,{\rm km\,s^{-1}}$ (which is what happens if you inject about 1 supernova-worth of energy per $100\,{\rm M_\odot}$ of stars formed).

As a result, small halos with escape velocities $v_\infty < v_{\rm wind}$ are the most efficient enrichers of the interstellar medium, and most of the enrichment occured at high redshifts (when haloes were typically smaller).

In the distant Universe, this dependence on galaxy (or strictly halo) mass can be folded into the return factor R through the use of luminosity functions and their evolution, so R is then redshift and luminosity dependent.

Topic 9

From the distant Universe, to the Universe today

Objectives for this lecture

- To appreciate how the initial conditions in the Universe give rise to the visible galaxy scale structures we observe today.

Further Reading

- Loeb & Furlanetto, Ch 8
- Madau & Dickinson (2012)
- Bouwens et al (2011 - *required reading*.

Expected length: 1 lecture

We have explored the properties of haloes, stars and galaxies in the distant Universe in an epoch known as *Cosmic Dawn*, when the first sources began to illuminate their surroundings. We have also had hints of *Cosmic Noon*, the peak of the star formation history, which also coincides with the peak of quasar activity and helium reionization. Now we need to explore how these early periods give rise to the Universe we see around us today. This is not a trivial question to answer, since as we saw earlier in the course, it the product of a large number of competing physical processes that must be folded into models. Nonetheless, we have many observation constraints.

9.1 Building galaxies

Galaxies are thought to be formed through hierarchical structure formation, in which galaxies grow predominantly through mergers and interactions with other galaxies, rather than through accretion of gas from the IGM. The first galaxies will form in initial dark matter haloes with relatively low masses. These formed stars, and the luminosity function was weighted strongly to the low mass end. However, if you integrate the luminosity function today it is clear that the "average" luminosity is L_*, that is the luminosity density is dominated by large galaxies, not by small ones.

The solution to this is simply that these small galaxies don't exist in the same numbers today that they did in the early Universe, because they have merged with other galaxies in either major or minor mergers. The boundary between these ategories varies somewhat from study to study but an interaction typically requires a mass ratio between the two components of $\sim 3:1$ or less to be considered a major merger. On the whole we expect minor mergers to be far more common in the local Universe, but major mergers to be more common at early times.

This is explicitly predicted in numerical models (e.g. the Millenium Run or Illustris simulations we mentioned earlier), but can also be tested observationally by looking directly at galaxies in the process of merging. i.e. taking deep images and looking at the merger fraction. Minor mergers are hard to identify at high redshift since the faint, accreting galaxy may not be visible, but major mergers will cause significant disruption to both galaxies and result in galaxies that have a *train-wreck* morphology. That is to say they will be significantly disturbed.

We can quantify this.Typically we define mergers based on their asymmetry – that is to say that we take a galaxy, rotate it by 180 degrees and subtract the new image from the original. Formally, using one of the most common definitions, the asymmetry is given by

$$A_\phi = \frac{\sum |I_0 - I_\phi|}{2 \sum |I_0|},$$

where I_ϕ is the flux in a given pixel, and $\phi = 180^\circ$ is the angle through which the image has been rotated (Conselice et al 2000). For a symmetric galaxy any residual flux is likely to be small and randomly distributed so A_{180} is small. By constrast, a merger will have significant asymmetry and a higher value of A_{180}.

We can also look at other parameters including central concentration, skewness and any visual sign of mergers or tidal tails. A lot of this sort of work is still best done (initially at least) by eye, and some big surveys have recruited armies of *citizen science* volunteers – Galaxy Zoo is the archetypal example of this.

Galaxy mergers are often the trigger for massive starbursts, since they cause shocks and increase the density of gas - a so-called *wet merger*. They may also bring in a new gas supply to a galaxy which has hitherto lost their own. In some cases, such as in giant ellipticals, the merger is *dry* – neither galaxy has fuel for star formation, and the primary effect is to randomise the orbits and so destroy any disk-like structure in the infalling galaxies.

9.2 Building Supermassive Black Holes (SMBHs)

As we have seen, black holes can grow by accretion from their hosts in the form of stars or gas which fall into the event horizon. This may be the dominant channel for the growth of supermassive black holes, but it is not the only one.

Galaxy mergers are also responsible for the growth of supermassive black holes, and the creation of active galactic nuclei (e.g. quasars). Each galaxy in a merging pair will possess its own supermassive black hole, which has settled to the centre of the halo potential well. While the gravitational forces between the galaxies, combined with pressure associated with dust and gas, will result in the merger of the galaxy pair, it is not so straightforward to merge the black holes that were in their centres. The friction created by interstellar gas may not be enough, and so we also need dynamical friction from the stars in the field.

As a black hole moves through a region stars are drawn towards it. However, during this time it continues to move, and the stars end up drawn into its wake, following behind it. This creates an effective net gravitational force that slows the black hole down. As it slows down in falls inward

through the joint potential (lower velocity = lower orbit) until it gets close to the other supermassive black hole that is either at the centre of the potential (minor merger), or has also fallen down (major merger). Here however we encounter a problem, because the supply of stars within the final few parsecs may not be sufficient for dynamical friction to be effective (the orbit period is now short to be significantly affected). Therefore some further means is needed to get the holes closer together. This is likely to come from some form of gas or discs, although the precise solution is unclear.

None-the-less these influences create much closer black hole binaries. Once a black hole is sufficiently close gravitational radiation will take over - energy is being lost not just by friction or in radiation, but also in gravitational waves. Once this becomes dominant, the time to merger is given by

$$t_{Merger} = \frac{5c^5}{256G^3} \frac{a^4}{M_1 M_2 (M_1 + M_2)} \tag{9.1}$$

For a supermassive black hole pair $M_1 = M_2 = 10^6$ $M_\odot$, we can get merger times less than the Hubble time once the separation is down to $1/1000$th of a parsec. Hence eventually gravitational radiation will result in the merger. The final merger itself will still happen at fairly large radius, because the Schwarschild radius of each black hole is large. This means there will be a low frequency gravitational wave transient at this point.

9.3 Building Galaxies like the Milky Way

Through a combination of observations like these and modelling the physical processes that develop, we can start to build a picture of how galaxies like the Milky Way came to form:

They started life as miniscule perturbations in the density field of the Universe. We see the imprints of these on the CMB and can trace their early oscillations as they collapsed under gravity to form dark matter halos. These collected baryons in their gravitational wells, which collapsed in turn and then cooled until they could ignite the first stars which lived and died in a spectacular, short lived burst known as *First Light* or *Cosmic Dawn*. Around these first dense regions, more familiar stars and galaxies formed, as knots of material in a cosmic web which was gradually reionized by early star formation in the galaxies themselves. The same star formation also enriched the Universe, driving metal-rich gas out of small galaxies in supernova-driven winds.

The star formation rate swelled over time, driven in part by mergers which collected the baryons from a variety of initial halos into one massive halo. Around the peak of star formation, at *Cosmic Noon*, a galaxy like the Milky Way experienced several major mergers, and most will also have undergone quasar activity which suppressed future star formation. After this, the star formation density decreases, but the mass of the galaxy continues to be swelled by ongoing minor mergers, each of which might also bring in fuel for new star formation. Not having undergone a major merger in some time, a massive spiral like the Milky Way has had time for the stars and gas to settle into an angular-momentum supported disk, but this isn't the end of the story.

In a few billion years, our next major merger will occur, introducing us to an uncomfortably close relationship with the Andromeda galaxy. By that time, future students might look back to now as the Distant Universe. If so, they'll have a rather remarkable view.

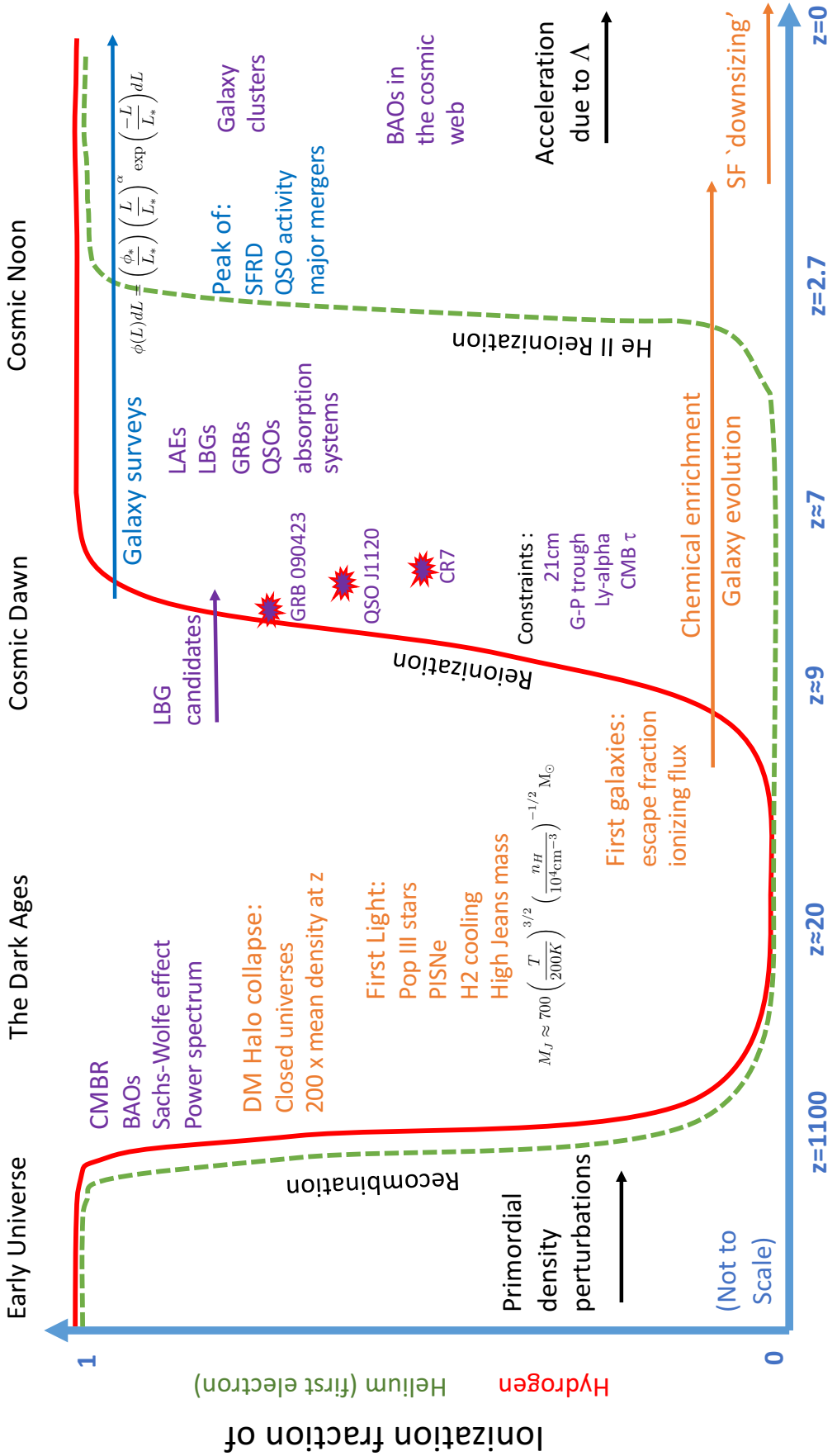

Figure 9.1: The Distant Universe in a single page.

9.4 Course conclusions

We are now at the end of the course. We have seen how the initial conditions of the Universe were set, and revealed how we measure them through observations of the CMB. We have described how the initial seed perturbations collapse as closed Universes to create dark matter haloes, and how gas within these haloes coalesces into the first stars. We have outlined how these first stars impact the Universe at large through a process of reionization, and how observational searches might look for them. We have also discussed three distinct routes (galaxies, quasars and GRBs) for observationally probing into the early Universe. Finally, we have looked in broad terms at how these early galaxies have evolved both in terms of star formation rate and in physical properties. From this you hopefully now have a reasonable idea of what physical processes were important in the early Universe, how they shaped the first structures to form, and how this in turn impacted the Universe that we see around us today.

Appendix A

Distant Universe Example Questions

Presented here are some example questions to get you thinking about aspects of the course, and to give you practice in (i) analysing and condensing the qualitative information or underlying physical principles, and (ii) working with the quantitative relations and information used in the course. Where brief notes are requested, you should aim to condense the key points to about half a page (or half a page per topic for questions with multiple parts/options). Note that standard values for most of the relevant constants are given on page 5.

Q1. (a) Sketch the angular power spectrum of the temperature fluctuations in the Cosmic Microwave Background (CMB) with labelled axes and appropriate scales.

(b) How does the first peak provide an estimate of the total density of the Universe?

(c) Explain how an estimate of the matter density can be obtained from additional peaks.

Q2. The sound horizon at the CMB surface of last scattering is given by

$$r_d = \int_0^{t_{rec}} \frac{\mathrm{c_s}\, dt}{a(t)},$$

where a is the scale factor, c_s is the sound speed and t_{rec} is the age of the Universe at recombination.

(a) Show that this can also be written as

$$r_d = \int_{z_{rec}}^{\infty} \frac{\mathrm{c_s\, dz}}{H(z)}$$

(b) Assuming the sound speed is a constant, show that $r_d \propto (\Omega_m z_{rec})^{-0.5}$, stating any assumptions you make.

(c) How is the sound horizon measured at different redshifts?

Q3. We can treat the evolution of primordial density perturbations as the expansion and collapse of a small, closed Universe evolving according to the Friedmann equation.

(a) Show that a parametric solution of the form,

$$a(\theta) = \frac{4\pi G \epsilon_0 R_0^2}{3c^4}(1 - \cos\theta) \qquad\qquad t(\theta) = \frac{4\pi G \epsilon_0 R_0^3}{3c^5}(\theta - \sin\theta),$$

satisfies the Friedmann equation in this case.
[Hint: begin by calculating $da/d\theta$ and $dt/d\theta$.]

(b) Calculate the maximum size reached by a perturbation with energy density η_0.

(c) Explain why the Universe if filled by a cosmic web rather than black holes.

Q4. Outline the key advantages and disadvantages of N-body simulations as opposed to hydrodynamic simulations. Give examples of their uses.

Q5. A primordial overdensity has a virial mass of $3 \times 10^7\,\mathrm{M}_\odot$ at $z = 15$.

(a) Assuming infalling gas with a mean molecular weight $0.6\,m_p$, calculate its virial temperature.

(b) Why does this region not immediately collapse under gravity?

(c) How does the region cool?

(d) The overdensity collapses when it reaches a hydrogen density of $3 \times 10^4\,\mathrm{cm}^{-3}$ and a temperature of 500 K. What is the Jeans mass of the resulting star?

(e) How would this differ if the star formed in similar gas at 200 K?

Q6. Sobral et al (2015) identified a galaxy named *CR7* as a candidate Pop III starburst.

(a) What observational signatures do we expect from Pop III sources?

(b) Which signatures of CR7 supported this interpretation?

(c) Is the interpretation of CR7 unambiguous and universally accepted?

Q7. What is the stellar *initial mass function* and why is it relevant to discussions of Pop III stars and their deaths?

Q8. The Lyman break technique is widely used to identify distant galaxy candidates.

(a) Why is this technique particularly useful at high redshifts?

(b) We are interested in galaxies that may contribute to reionization at $z = 9$. Suggest the central wavelengths of three filters that may be used for Lyman break selection in this instance.

(c) How might we confirm that any selected galaxies are indeed at $z \sim 9$?

(d) Why does this technique become more challenging at the highest redshifts?

Q9. Outline the use of (i) GRBs and (ii) QSOs as probes of the distant Universe. Give example cases where possible.

Q10. The growth of supermassive black holes at high redshift is something of an unsolved problem.

(a) Explain the physical interpretation of the *Eddington Luminosity* and show that

$$L_{Edd} = \frac{4\pi G M m_p c}{\sigma_T} = \epsilon \dot{m} c^2$$

(b) Assume that a quasar with black hole mass $M = 6 \times 10^9\,\mathrm{M}_\odot$ has been accreting at the Eddington rate with an efficiency $\epsilon = 0.2$ and a duty cycle (i.e. active accretion time fraction) of 50%.

i. Given a seed mass of $10\,\mathrm{M}_\odot$, when did the black hole form?

ii. This quasar is observed when the universe was 800 Myr old. Does this present a contradiction?

iii. How might the rapid growth timescales of distant quasars be reconciled with these constraints?

Q11. An ultraviolet luminous source will ionize its surroundings. In an expanding Universe, the volume ionized is given by:

$$\frac{dV}{dt} = \frac{\dot{N}}{\langle n_H \rangle} - V\left[\frac{1}{\bar{t}_{rec}} - 3H(t)\right]$$

where

$$C = \frac{\langle n_e n_p \rangle}{\langle n_e \rangle^2}.$$

(a) Outline the steps required to reach the above result.

(b) How does it differ from the case of a non-expanding, static region?

(c) What is the parameter C called, and what is its relevance (i) in the distant Universe, (ii) at later times?

Q12. The cross section of a hydrogen absorption feature is given by:

$$\sigma(\nu) = \frac{\pi\, e^2}{m_e\, c}\, f\, \delta(\nu_{\rm obs} - \nu_0),$$

where f is an oscillator strength, $\nu_{\rm obs}$ is the observed frequency and ν_0 the rest frequency of the absorption feature.

(a) Show that the optical depth, $\tau(\nu_{\rm obs})$, depends on the observed frequency as,

$$\tau(\nu_{\rm obs}) = \sigma_0 \left(\frac{c}{H_0}\right)\left(\frac{H_0}{H(z)}\right) n_H(z).$$

[Hint: Remember that scale factor $a = (1+z)^{-1} = \nu_{\rm obs}/\nu_0$.]

(b) Is the use of $\delta(\nu_{\rm obs} - \nu_0)$ in the above equation valid in all cases? Support your argument with examples.

Q13. Describe the the constraints on reionization derived from at least **three** named astronomical observations, indicating these on a sketch.

Q14. Explain what is meant by the cosmic Star Formation Rate Density (SFRD) history. Your answer should consider the evolution of the relevant function, the physical interpretation of this evolution, how it is constrained, and any uncertainties affecting it.

Q15. When, where and how was the baryonic matter in the Universe enriched from primordial gas to higher metallicities?

Q16. Two massive spiral galaxies merge at a redshift $z = 2$.

(a) What would be the likely observational signatures of this merger?

(b) Which would merge more rapidly, the central black holes or the stellar distribution?

(c) Once in close proximity, the black holes will merge on a timescale governed by

$$\frac{dl}{dt} = \frac{-64}{5} \frac{G^3\, m_1\, m_2 (m_1 + m_2)}{c^5\, l^3},$$

where l is the orbital seperation and we have neglected any effect of eccentricity in the orbit.

i. Why does the orbit shrink?

ii. Show that

$$t_{Merger} = \frac{5c^5}{256G^3} \frac{l^4}{M_1 M_2 (M_1 + M_2)}$$

iii. What would be the merger time for two $10^7\,\mathrm{M}_\odot$ black holes, originally seperated by 0.05 pc?

iv. How might we observe the merger event?