The cosmic history of the intergalactic medium

Cristiano Porciani

December 2010

Contents

Chapter 1

Introduction

1.1 What is this course about?

Galaxies are gravitationally bound collections of stars, gas, dust, cosmic rays, and, very likely, non-baryonic dark matter. Deep optical images show that they are the main building blocks of the universe (Figure 1.1). Galaxy surveys have indicated that the spatial distribution of galaxies displays a foamy cellular texture (Figure 1.1). The most outstanding feature is the presence of compact associations of galaxies containing up to a few thousand members and extending to Mpc scales (galaxy clusters and groups). The space in between massive clusters is bridged by a highly structured network of (almost) uni-dimensional arrays of galaxies (filaments) extending for tens of Mpc. The filaments are interweaved with vast regions of space nearly devoid of galaxies. These "cosmic voids" are approximately spherical and extend up to ~ 50 Mpc.

A question then arises spontaneously: is there gas (or baryonic material in general) in the space between galaxies? In fact, it seems unlikely that galaxy formation would have been 100% efficient locking all the baryons into galaxies. We also know that galaxy evolution can eject material out through galactic winds, tidal stripping or more violent galaxy encounters. Therefore we expect that there exists an intergalactic medium (IGM).

This course will review the current knowledge on the matter. In particular, it will address questions like:

- How many baryons are locked in stars and galaxies?
- How many of them reside in clusters, filaments, and voids?
- What is the state of the diffuse gas forming the IGM?
- What was the history of the IGM over cosmic time?
- What does the IGM tell us about cosmology and physics in general?

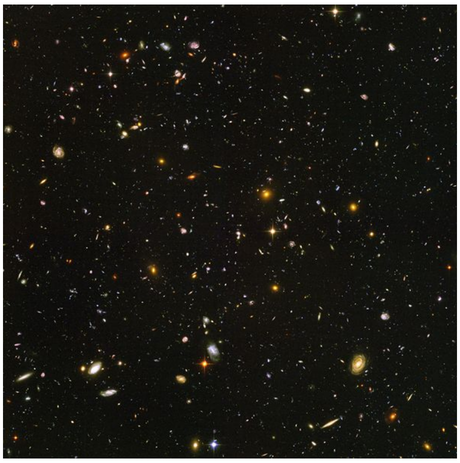


Figure 1.1: The Hubble "Ultra Deep Field" (UDF) has been obtained pointing to the same region of the sky during 400 orbits of the Hubble Space Telescope (HST) for a total exposure time of 1 million seconds (11.3 days). Nearly 10,000 galaxies are visible in this image. Some of them are very dim, we receive 1 photon per minute from them. Note that the area in the picture covers nearly 1/50 of the lunar surface. The whole sky contains 12.7 million more time area the the HUDF.

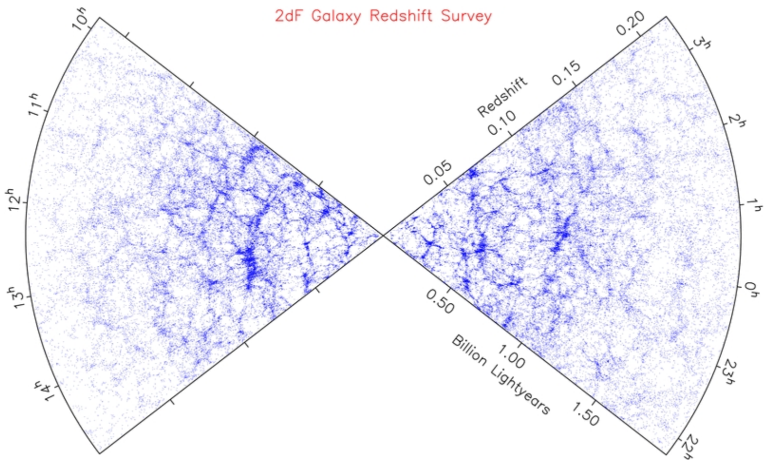


Figure 1.2: A map of the distribution of galaxies in a thin wedge on the sky from the 2-degree Field Galaxy Redshift Survey (from Colless et al. 2003). Each point marks the position of a galaxy with a measured distance. Note that, at small distances, where the sampling rate of this flux-limited survey is higher, large-scale structures are clearly visible while at larger distances they are hidden by shot noise.

This is a very young field of research, dating back just to the late 1960s. Despite it has matured fairly quickly, especially in the last decade, there still exists many exciting directions to explore both theoretically and experimentally. In the near future, it will certainly provide many research opportunities for young scientists.

1.2 Historical remarks: the birth of modern cosmology

This course will follow a pedagogical appoach. Ideas will be introduced in such a way to make them more easily understandable. Material will not be presented following the chronological order of discoveries. It is very instructive, however, to spend a little bit of time glimpsing the history of this field of research. Beyond helping us to become familiar with some of the concepts, this will give us the opportunity to appreciate the difficulties that the community had to face, and the alternative explanations that now have faded into oblivion. Also it will be an ideal way to thank all the scientists of the past (including all those who will not be mentioned here) who contributed moving the horizon of knowledge a little bit further. Remember that behind each name there was a human being first. This people were infants, grew up, ate, slept, studied, made errors, loved, hated, suffered, aged like anyone of us. To acknowledge this, I will try to show you slides with pictures of most of them.

Of course, for time reasons, our little excursus through history will be far from complete. Every now and then we will interrupt following the flow of history to summarize what the current understanding is regarding a specific topic. These summaries are highlighted by boxes surrounding the text.

1.2.1 The general theory of relativity

We start our brief historical excursion from 1915 when Albert Einstein (1879-1955) published his general theory of relativity. Postulating the equivalence principle (i.e. asserting the complete physical equivalence of gravitational acceleration and the inertial acceleration of a reference frame), he presented a metric theory of gravitation where space-time was warped by the presence of matter. The theory predicted new phenomena like gravitational redshift (light becomes redder receding from a massive body) and gravitational light deflection (light rays passing close to a massive body are bended towards the body).

The general theory of relativity has been experimentally and observationally tested to a high-degree of accuracy (in the weak-field limit) from mm scales to solar-system scales (nearly 16 orders of magnitude) and is still our favoured model of gravitation. Note that applying the general theory of relativity to the size of the visible universe still requires an extrapolation over 13 orders of magnitude.

1.2.2 The expansion of the universe

Einstein himself soon used his new theory to describe the entire universe. On February 15 1917 he published the paper "Cosmological considerations in the General Theory of Relativity" where he presented a model of the universe. This conventionally marks the birth of modern cosmology. Einstein assumed that the universe on large-scales looks the same from any point in it (i.e. it is homogeneous) and in any direction (i.e. it is isotropic). This now goes under the name of "cosmological principle" and has deep philosophical implications linked to the Copernican principle. In the same paper, Einstein also added a cosmologically repulsive term (the cosmological constant Λ, intended as a fundamental constant of nature) to the original field equations to keep the universe static under the action of gravity on matter. In the same year, Willem de Sitter (1872-1934) and Tullio Levi-Civita (1873-1941) independently found a curious solution to Einstein's equations. It represented an "empty" universe with no matter but including the cosmological term. Arthur Eddington (1882-1944) noted that, although the de Sitter solution was static (the metric did not contain any explicitly time-dependent term), any test particle in the space-time manifold would exhibit a radial motion. The solution also implies a cosmological redshift[1] that becomes increasingly apparent at large distances. De Sitter wrote: "Consequently the frequency of light-vibrations diminishes with increasing distance from the origin of co-ordinates. The lines in the spectra of very distant stars or nebulae must therefore be systematically displaced towards the red, giving rise to a spurious positive radial velocity." This feature was called the de Sitter effect. For small distances r, the redshift was expected to scale as $z \propto r^2$. Paul Ehrenfest (1880-1933) is said to have been the first to realize that the redshift was a pure effect of the curvature of space-time similar to the gravitational redshift predicted by Karl Schwarzschild (1873-1916) around a spherically-symmetric, non-rotating, non-charged mass in 1916 (the first exact solution to Einstein's field equations).

Alexander Friedman (1888-1925) and Georges Lemaître (1894-1966) pioneered the discussion of time-dependent cosmological models. In 1922 and 1927, respectively, they independently worked out a solution to general rela-

[1] The wavelength $\lambda_{\rm obs}$ at which an observer detects a given spectral features is shifted towards the red with respect to the corresponding wavelength at emission $\lambda_{\rm em}$ by the "cosmological redshift" $z = \lambda_{\rm obs}/\lambda_{\rm em} - 1$.

tivity field equations where the universe was expanding. Both of them used the original equations by Einstein without the cosmological term. In 1927 Lemaître also showed that his models admitted a linear redshift-distance relation $z \propto r$ (to first order in r).

Progress came also from more mathematical studies. In 1929 Tolman (1881-1948) proved that there are only 3 static solutions of Einstein's equation that also satisfy the cosmological principle. All of them were already known: the metric of special relativity and the two world models by Einstein and by de Sitter. In 1930 Eddington showed that Einstein's static solution is unstable against spatially homogeneous and isotropic perturbations that makes it either expand or contract. Howard Robertson (1903-1961) in 1935 and Arthur Walker (1909-2001) in 1936 independently proved that the metric used by Friedman and Lemaître is the only one on a Lorentzian manifold that is both homogeneous and isotropic. Note that this is a geometric result that holds for any metric theory as it does not rely specifically on the equations of general relativity.

Laborious activity was begun simultaneously on the observational side. Between 1913 and 1917, Vesto Slipher (1875-1969) managed to take optical spectra of a number of "spiral nebulae"[2] finding that the vast majority (21 out of 25) showed spectral lines shifted towards the red with respect to those produced by reference arc lamps in the spectrograph. He interpreted this red-shift as a Doppler effect ($z \sim v/c$ with v the recession velocity and c the speed of light): most nebulae were receding from the Earth. As a result of the disruption of communications during the First World War, de Sitter and Slipher did not know of each other's accomplishments. In 1923, Edwin Hubble (1889-1953) discovered variable Cepheid stars in the Andromeda spiral nebula. This provided convincing evidence that nebulae were "island universes": i.e. stellar systems similar to our Milky Way (what we nowadays call galaxies) at enormous distances. With the help of Milton Humason (1891-1972) he extended Slipher's catalog of velocities and, using the period-luminosity relation for Cepheid stars previously discovered by Henrietta Leavitt (1868-1921), he could estimate the distance of the galaxies. In 1929 he provided observational evidence for a linear relation between the recession velocity of a galaxy and its distance. Note that a similar analysis (looking for a quadratic relation, however) had already been fruitlessly attempted by Knut Lundmark (1889-1958) and by Ludwik Silberstein (1872-1948). Hubble was very cautious regarding the interpretation of his results and wanted to leave open the possibility for a quadratic relation

[2] At the time the origin of the nebulae was unclear. In 1920 the Great Debate took place where Harlow Shapley (1885-1972) and Heber Curtis (1872-1942) disputed over the size of the universe. Curtis argued that the universe is composed by many galaxies like our own which had been identified as spiral nebulae. Shapley argued that the spiral nebulae were nearby gas clouds and that the universe was made by one big galaxy, the Milky Way. But this is another story.

that would not have implied a time-dependent solution of Einstein's equations. In his 1929 paper he wrote: "The outstanding feature is the possibility that the velocity-distance relation may represent the de Sitter effect. . . . In this connection it may be emphasized that the linear relation found in the present discussion is a first approximation representing a restricted range in distance." However, in 1931 and 1934 Hubble and Humason enlarged their dataset showing beyond any doubt the linearity of the effect. In 1933 de Sitter then wrote "We know now, because of the observed expansion, that the actual universe must correspond to one of the nonstatic models. . . . The static models are, so to say, only of academic interest." In fact, one year before, he had published a paper together with Einstein where they proposed that the cosmological constant should be set equal to zero and worked out a world model with flat spatial sections at constant cosmic time that was filled with matter. As like as in the Friedmann-Lemaître models, time had a beginning. The peculiarity was that galaxies recede for ever but the cosmic expansion rate asymptotically coasts to zero as time advances to infinity. The model provides the separating case between recollapsing and infinitely expanding models.

Nevertheless, the concept of an expanding universe seemed so bizarre that people immediately started finding alternative explanations for the observed redshifts. Hubble himself was always very reluctant to believe that the redshifts represent a true expansion rather than being "caused by an unknown law of nature."

Today we have additional pieces of evidence in favour of the expansion of the universe. The main one concerns the cosmic microwave background and will be discussed in section 1.2.5. Another one comes from using the light curves of Type Ia supernovae as clocks. As originally proposed by Olin Wilson (1909-1994) in 1939, relativistic time dilation should stretch the light curves proportionally to a factor $1 + z$. This effect has now been observed and data give a spectacular confirmation of the theory: the light curves of distant supernovae are consistent with those of nearby ones whose time axis is dilated by a factor $1 + z$ (Leibundgut et al. 1996; Goldhaber et al. 1997, 2001). Similar results have been obtained analyzing Type Ia supernova spectra (Blondin et al. 2008).

It is now time to fix a bit of notation. The constant of proportionality between recession velocity and physical distance is now called the "Hubble constant" and is generally indicated with the letter H: $v = Hr + \mathcal{O}(r^2)$ or $z = Hr/c + \mathcal{O}(r^2)$ (both to first order in r). Note that the name refers to the constancy of H over space. In the Friedmann-Lemaître models H is linked to the scale factor of the universe $a(t)$ (a measure of how length scales change over time) by $H = \dot{a}/a$ (where the dot indicates the time derivative), therefore H is generally time dependent! Its present-day value is usually written as $H_0 = 100\,h$ km s^{-1} Mpc^{-1} with h a dimensionless number. The best direct measurement with the Hubble space telescope gave $h = 0.72 \pm 0.08$ (Freedman et al. 2001). Note that $1/H$ has the dimension of time. This quantity is dubbed "Hubble time" and gives the characteristic timescale for cosmic expansion. In numbers: $1/H_0 = 3.086 \times 10^{17} h^{-1}$ s $= 9.778 \times 10^9 h^{-1}$ yr. The current "age" of the universe in the Friedmann-Lemaître models is obtained multiplying the Hubble time by a number of order unity (that depends on the specific model).

The Hubble constant also determines the "critical density" ρ_c, i.e. the minimum density of a universe filled with matter that that would be needed to halt the cosmic expansion at some point in the future. General relativity gives: $\rho_\mathrm{c} = 3H_0^2/8\pi G \simeq 2.778 \times 10^{11} h^2 M_\odot$ Mpc^{-3} $= 1.88 \times 10^{-26} h^2$ kg m^{-3} $= 11.2h^2$ protons m^{-3}. In cosmology it is customary to express densities in units of ρ_c and to indicate them with the letter Ω followed by a subscript indicating what density one is speaking about. For instance, the baryon density will be $\Omega_\mathrm{b} = \rho_\mathrm{b}/\rho_\mathrm{c}$.

1.2.3 The cosmological debate

Following Hubble discovery, in 1931 Lemaître had suggested that the universe might have originated when a "primeval atom" or "cosmic egg" exploded in a spectacular fireworks, creating space-time. This proposal met skepticism from the scientists of the time. Lemaître was a Catholic priest and his theory was considered too strongly reminiscent of the dogma of creation.

The idea of an evolving universe was soon challenged by a new hypothesis, that the universe might be in a steady state after all. [3] In 1948, Fred Hoyle (1915-2001), Hermann Bondi (1919-2005), and Thomas Gold (1920-2004) formulated the so-called "steady-state model" based on the "perfect cosmological principle": the universe looks the same from every point in it and at every time. How could the universe continue to look the same when observations suggested it was expanding? To this regard, Hoyle wrote: "One tends to think of unchanging situations as being necessarily static. ...One can have unchanging situations that are dynamic, as for instance a smoothly

[3] Hubble's original measurement of the expansion rate gave $H \sim 500$ km s^{-1} Mpc^{-1}, corresponding to an expansion age of ~ 2 Gyr. This was shorter than the estimated age of the Earth from radioactive dating!

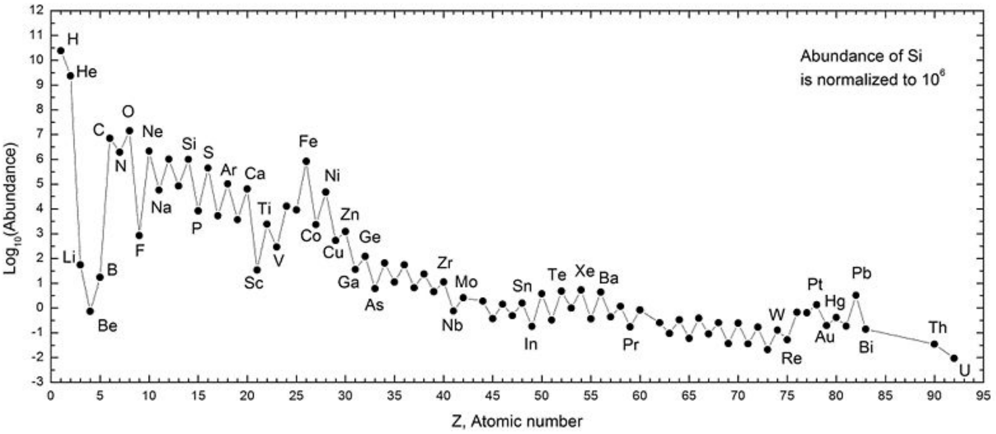


Figure 1.3: Relative abundance of the chemical elements in the Solar System.

flowing river." The key idea was to balance the ever-decreasing density of an expanding universe assuming that matter was continuously created in such a way that the cosmic density was kept always the same. The amount required was undetectably small: 1 nucleon every 50 years in a cubic km. While Bondi and Gold did not propose any mechanism for matter creation, Hoyle introduced the concept of the C-field (C for creation) a reservoir of negative energy which, because of energy conservation, was becoming more negative every time matter was created from its perturbations and then restored to the original value by cosmic expansion.[4] The negative pressure of the C-field drove the steady expansion of the cosmos.

During those years, the cosmological debate was very harsh and acquired religious and political aspects. Some people associated the steady-state model to the Communist party (when, in reality, Soviet astronomers rejected both western world models as idealistic and unsound). Hoyle saw in it the symbol of freedom and anti-communism, other loosely associated it to atheism. In this climate, during a BBC radio talk in 1949, Hoyle coined the name "big bang" for the competing theory. In 1952, Pope Pious XII announced that big-bang cosmology was in harmony with the Christian dogma.

1.2.4 The origin of the elements

With the development of nuclear physics another astrophysical problem became of great interest. 92 elements are naturally found on Earth, of which 80 have stable isotopes. Their relative abundance in the Solar System as a

[4] One might smile today but would you consider more reasonable to create a few atoms here and now or the entire universe out of some quantum fluctuations?

function of atomic number shows peculiar features (Figure 1.3). Can physics explain these trends?

In the 1920s Arthur Eddington suggested that stars obtain their energy from nuclear fusion of hydrogen into helium. In 1928, George Gamow (1904-1968) developed the basic theory that gives the probability for two nuclei to fuse at given conditions of the stellar interior (temperature and density). In the late 1930s, Hans Bethe (1906-2005) and Karl von Weizsäcker (1912-2007) individuated the proton-proton chain and the CNO cycle. This was enough to explain the source of energy that kept the stars hot and prevent them to collapse under their own weight. The creation of heavier nuclei was not addressed however.

In 1946, George Gamow (who had briefly studied with Friedman) started considering the implications of cosmic expansion and cooling from an initial state of nearly infinite density and temperature. He realized that sufficiently early on all matter would have been protons, neutrons and electrons swamped in an ocean of high-frequency radiation that dominated the energy budget.[5] Gamow thought that in these conditions all elements could be built up capturing neutrons one by one. Protons could capture neutrons and lead to the formation of deuterium atoms. Then the subsequent neutron captures resulted in the building up of heavier and heavier nuclei. β decay would have got rid of unstable atoms. He also realized that this could have happened for a relatively short time as the universe would have soon become too cold and too little dense because of its rapid expansion. In his mind, the slope of the abundance curve was related to the expansion history of the universe. His graduate student Ralph Alpher (1921-2007) made detailed calculations using one of the first computers (some neutron-capture cross sections stopped being classified material after the end of World War II). His results were published in the famous $\alpha\beta\gamma$ paper (Alpher, Bethe & Gamow 1948).[6] They roughly agreed with the observations of stars: Helium accounted for roughly a quarter of the mass, and Hydrogen for nearly all the rest. However, Enrico Fermi (1901-1954) noted an inconsistency in the calculation. The cross section for neutron capture of a Helium atom is basically zero. However, Alpher fitted a smooth curve to interpolate through the known cross sections and this made this particular one excessively high. Attempts to repeat the calculations with the correct reaction rates failed to get a sensible answer for heavier elements. Basically primordial nucleosynthesis could not produce anything heavier than Helium atoms.

Fred Hoyle dismissed the attempts of pre-stellar element buildup as "requiring a state of the universe for which we have no evidence" and continued

[5] Gamow dubbed this hypothetical mixture of particles "Ylem" (from obsolete Middle English phylosophical jargon indicating "the primordial matter from which all matter is formed" and deriving from the Greek hylem, "matter").

[6] The name of Bethe was added by Gamow just to make the list of authors appear funny. This, however, made Alpher deeply unhappy.

to pursue the possibility that elements were cooked up in stars. In 1946 he showed that hot star interiors could synthesize elements from Carbon up to Iron but the paper remained unnoticed for long time. In 1950 a revolutionary observational result solved part of the controversy. Martin Schwarzschild (1912-1997, son of Karl) together with his wife Barbara showed that population I stars (young stars in the disk of the Milky Way, originally selected by Walter Baade (1893-1960) for their low vertical velocity with respect to the disk of the Galaxy) have a greater abundance of Iron and other metals with respect to population II stars (old stars in the halo of the Milky Way, originally selected by Baade for their high vertical velocity with respect of the disk of the Galaxy). This striking evidence for metal production by stars removed the need for a pre-stellar mechanism for element formation. The subsequent work by Edwin Salpeter (1924-), Fred Hoyle, and Willy Fowler (1911-1995, Nobel prize winner in 1983) culminated in the seminal paper B^2FH (Burbidge, Burbidge, Hoyle & Fowler 1957) that showed that all atomic elements heavier than lithium up through uranium could be synthesized in stars.

Surprisingly enough, it was Hoyle himself to show that the total amount of energy released by the formation of all observed Helium is some ten times greater than the energy radiated by galaxies since their formation (Hoyle & Taylor 1964). So the work begun by George Gamow was revived, revised and merged with the stellar channel.

As you will discuss in detail in the cosmology class, the current tenant is that primordial nucleosynthesis took place after protons and neutrons condensed out of the primordial quantum soup (Baryogenesis) and lasted for a few minutes, until densities and temperatures were too low for nuclear reactions to happen. In the end, big-bang nucleosynthesis (BBN) produced a mixture dominated by Hydrogen ($\sim 76\%$ in mass) and Helium 4 ($\sim 24\%$ in mass), along with trace amounts of Deuterium, Helium 3 (any Hydrogen 3 decays to Helium 3), Lythium, and Berillium (Figure 1.4).

The key parameter that regulates the final elemental abundances is the current number of baryons per photon, $\eta = 3.4 \times 10^{-10}(\Omega_{\rm b}h^2/0.0125)$. Since the temperature at which the reactions happen is fixed (at the onset of nucleosynthesis $T \sim 70$ keV, η is simply a measure of the baryon density $\Omega_{\rm b}$. The great success of the theory is that a single value of $\Omega_{\rm b}$ is enough to simultaneously explain the observed abundance of all pre-stellar elements (Figure 1.4). The most recent studies give $\Omega_{\rm b}h^2 = 0.0214 \pm 0.002$.

We now firmly believe that all heavier elements have been synthesized in stars (in some sense then our bodies are made of stardust).

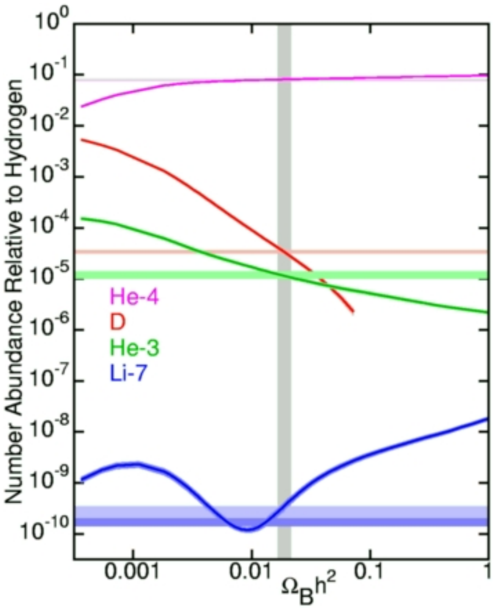


Figure 1.4: The light element abundance predictions from BBN theory plotted against the baryon density. From top to bottom are the mass fraction of ^{4}He and the relative fractions D/H, ^{3}He/H and ^{7}Li/H. The shaded bands enclose the 1σ experimental uncertainty.

1.2.5 The microwave background

In their 1949 paper, Alpher and Robert Herman (1914-1997) predicted that a remnant of the hot early universe would remain at late times: a cosmic background radiation permeating all space. The key reasoning was as follows. Cosmic hydrogen remained ionized until the temperature dropped below 3000 K because of the universal expansion. At this point, the photons could stream freely without interacting anymore with matter. As the universe expanded this radiation would cool. Alpher and Herman predicted that the temperature of this radiation now should be not higher than 5 K. They thought, however, that it would have been difficult to distinguish it from other forms of cosmic radiation including integrated starlight. In 1964, A Doroshkevich (1937-) and Igor Novikov (1935-) were the first to realize that the relic radiation should have had a blackbody spectrum[7] (as it had been in thermal equilibrium with matter through frequent interactions) and should be detectable with current technology in the microwave spectral range. They identified the ultra-sensitive horn antenna of Bell Labs at Crawford Hill (NJ) as the best available instrument for its detection. However, they misinterpred some published data taken with this instrument and

[7] Already in 1934 Tolman had demonstrated that black-body radiation in an expanding universe cools but remains thermal.

concluded that the "Gamow theory" was contradicted by experiments.

The year 1965 is a milestone of modern astrophysics and cosmology. Arno Penzias (1933-) and Robert Wilson (1936-) of Bell Labs were conducting radio astronomy experiments at Crawford Hill but were frustrated by a noise in its receiving system, a noise that remained constant no matter which direction they scanned. This made no sense but after carefully checking all the plausible sources (including chasing pigeons that lived in the antenna) the noise remained. Mentioning their unexplained noise to Bernard Burke (1929-) of MIT, they became aware of the theoretical work by Robert Dicke (1916-1997), a physicist at nearby Princeton University who had independently thought of the cosmic background radiation. In March 1965, Dicke and his former student Philip James Edwin ("Jim") Peebles (1935-) had published a paper explaining the origin and nature of this radiation. Together with some colleagues they had planned to build an experiment dedicated to its detection. An agreement was made: Penzias and Wilson published the data without attempting any interpretation while Dicke and collaborators wrote a different paper containing the interpretation as radiation of about 3 K left over from the big bang.

Many groups have now measured the intensity of the cosmic microwave background (CMB) at different wavelengths. Currently the best information on its spectrum comes from the FIRAS instrument onboard the COBE satellite (1974-1976, Figure 1.5). No spectral deviation from a black-body spectrum at $T = 2.725 \pm 0.002$ K was detected over the wavelength range from 0.5 to 5 mm. Moreover, the temperature of the CMB is isotropic to nearly one part per 10^5. Recent studies of the CMB temperature anisotropies provided a new measure of the baryon density in the universe: $\Omega_{\rm b}h^2 = 0.0223^{+0.0007}_{-0.0009}$ where the Hubble parameter is $h = 0.73^{+0.03}_{-0.04}$.

Penzias and Wilson were awarded the Nobel prize in Physics in 1978.[8] The leaders of the COBE experiment, John Mather (1946-) and George Smoot (1945-), got it in 2006 for having accurately measured the spectrum of the CMB and detected tiny temperature anisotropies, respectively.

The discovery of the CMB and the observation that quasars and radio sources were much more abudant at high redshift than in the local universe basically ended the era of the steady-state model.

[8] The first experimental evidence for CMB was actually obtained (but unrecognized) by Adams & Dunham in 1937 who detected several optical absorption lines due to the CN molecule rotationally excited by the radiation background.

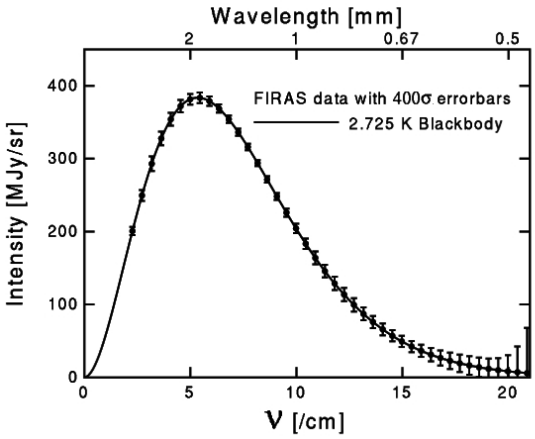


Figure 1.5: FIRAS spectrum of the CMB.

It is important to stress that, in its modern interpretation, the big bang is not an explosion localized in space and time. Rather it describes the origin of space-time and happens everywhere in the universe. At the enormous densities reached in the vicinity of the singularity quantum-gravity effects (that we do not know how to model) become extremely important. Therefore, the predictions of classical general relativity in this regime should not be taken too seriously. Inflationary theories give a speculative theoretical explanation of the origin of cosmic expansion without requiring a “bang”. Future experiments will test their predictions.

1.3 Historical remarks: The search for the IGM

We will now depart from the history of cosmology and focus on the search for the intergalactic medium.

1.3.1 The X-ray background

The energy density of the X-ray sky is dominated by a diffuse radiation which is mostly of cosmic origin: the X-ray background (XRB). It was first clearly detected in 1962 (before the discovery of the CMB) during a rocket flight intended to study X-rays from the Moon (Giacconi et al. 1962).[9]

At energies below 1 keV the background is patchy and clearly correlated with optical features in the Milky Way suggesting that it receives a sub-

[9] Riccardo Giacconi (1931-) was awarded the Nobel prize in 2002 for his pathbreaking work inventing the field of X-ray astronomy.

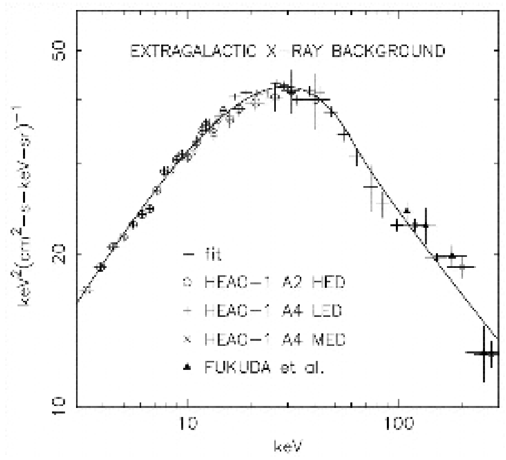


Figure 1.6: The spectrum of the X-ray background.

stantial contribution from Galactic emission (namely hot gas produced by supernova explosions in the Galaxy). On the other hand, at energies above 2 keV, the high degree of isotropy and the lack of correlation with Galactic features strongly suggest that the bulk of the background is of extragalactic origin.

For some time after the discovery of the cosmic XRB there was considerable controversy over its origin. Diffuse emission of thermal bremsstrahlung radiation from a hot intergalactic plasma at a temperature of $\sim 10^8$ K was a plausible source. Alternatively, the hard XRB could be attributed to the superposition of unresolved discrete X-ray sources (active galactic nuclei and quasars).

The spectrum of the XRB (Figure 1.6) was accurately measured in the late 1970s by the first of the High Energy Astronomy Observatories (HEAO-I, a NASA satellite). In the 2-10 keV energy range, this is well fit by a power-law model with a slope $\Gamma \sim 1.4$, significantly different from typical (Type I, showing broad optical spectral lines) active galactic nuclei (AGN) that are characterized by $\Gamma \sim 1.7$. This "spectral paradox" and the fact that the observed spectrum is well fit by a thermal bremsstrahlung model at a temperature of ~ 30 keV, were favouring the first possibility. However, the discovery of many soft X-ray sources (obscured, Type II AGN showing only narrow optical spectral lines) provided evidence for the second.

In order to explain the XRB with bremsstrahlung radiation, one has to assume that a density corresponding to $\Omega_\mathrm{b} \simeq 0.25 - 0.30$ is contributed by gas in a uniform IGM.[10] This gas was heated in the early universe by some unknown phenomenon and then cooled by adiabatic expansion, Compton

[10] This could be somewhat reduced by clumping the gas, but the isotropy of the microwave background forces the clump to be on a scale of less than 1 Mpc and so they must be confined in some way.

scattering against the microwave background and bremsstrahlung. This scenario suffers from a number of problems.

1. The gas density needed to generate the background is difficult to reconcile with the limits coming from standard primordial nucleosynthesis;

2. It requires a total energy nearly 40% of the cosmic microwave background to be injected into the IGM at early epochs (corresponding to a redshift $z > 6$). It is difficult to concieve how this could have happened.

3. A hot IGM perturbs the cosmic microwave background by inverse Compton scattering: the microwave spectrum is cooled by about 0.1 K in the Rayleigh-Jeans part of the spectrum but retains its exponential shape up to energies well into the exponential tail where it develops a quasi-power-law high-energy component which extends up to the infrared.

We now know that that the bulk of the XRB cannot originate in a uniform, hot intergalactic medium because a strong Compton distortion on the cosmic microwave background spectrum was not observed by the FIRAS instrument on the COBE satellite. Moreover, very deep images of the X-ray sky taken with the Chandra and the XMM-Newton satellites have shown that discrete sources can account for at least 75% of the hard XRB (and likely much more than that).

1.3.2 Intracluster gas

The existence of baryonic material outside galaxies became evident in the late 1970s and early 1980s when extended X-ray emission (Figure 1.7) has been detected from over a hundred local galaxy clusters ($z < 0.08$). In this case, there is little doubt that the dominant X-ray emission process is thermal bremsstrahlung. For instance, spectral lines resulting from transitions of higly ionized iron have been detected and found in good agreement with the hot-plasma hypothesis. The baryonic material that pervades the space between galaxies in a galaxy cluster has been dubbed the "intracluster medium" (ICM). Studies of the X-ray surface brightness have been used to define the cluster gravitational potential and thus its mass.

A number of sources can contribute gas to the ICM:

1. primordial gas can survive in the ICM without being ever included in galaxies as a consequence of the limited efficiency of the galaxy-formation process;

2. primordial gas can accrete onto the cluster at late epochs;

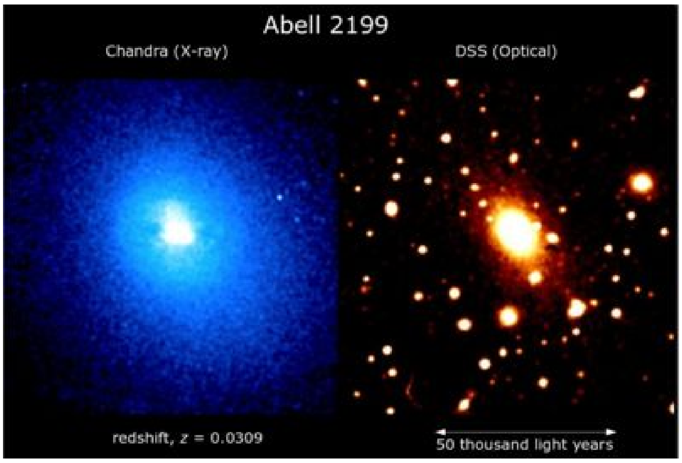


Figure 1.7: The map of the X-ray emission from the intracluster medium in the core of the Abell 2199 galaxy cluster (left) is compared with the corresponding optical emission of the galaxies (right).

3. gas injection of metal enriched gas from galaxies may result as a consequence of galactic winds, ram-pressure stripping or evaporation.

Evidence for galaxy-ICM interactions is provided by the fact that local galaxy clusters typically have a Fe/H abundance (in number) which is nearly one half of the solar value.

1.3.3 The Lyman-alpha forest

In 1971, Roger Lynds took the optical spectrum of the quasar 4C 05.34 (at redshift $z = 2.877$, the largest known at the time) and reported the presence of a much larger density of sharp absorption lines on the blue side of the Lyman-α emission line as compared with the red side.

He suggested that the absorption lines are due to the presence of intervening intergalactic clouds absorbing in the strongest hydrogen resonance line: the Lyman-α transition. The absorption lines appear at longer wavelengths due to the expansion of the universe and the cosmological redshift.

Further observations revealed that the phenomenon is widespread and applies to all high-redshift quasars (Figure 1.8), in some cases, there are associated Lyman-β and γ lines; however, only in rare cases, are there lines of heavier elements at the same redshift. This supports the idea that the absorption features are generated by large clouds of atomic hydrogen.

Several distinct possibilities were proposed to account for the origin of the lines. The absoption could in fact take place from:

- Truly intergalactic gas clouds and protogalaxies (Figure 1.9, originally proposed by Arons in 1972);

- Very extended, diffuse, hydrogen halos surrounding each galaxy;
- Strong winds generated by violent star-formation episodes in dwarf galaxies;
- Gas pervading superclusters of galaxies (a sort of intrasupercluster gas);
- Shockwaves propagating out of star-forming galaxies or quasars.

Detailed determinations of the number density, clustering properties, and metal content of the absorbers combined with studies of the line profiles ruled out all the possibilities but the first. The main arguments against the hypothesis that the Lyman-α forest is associated to galaxies are:

1. The objects responsible for the Lyman-α absorption are generally poor in heavy elements;
2. The number density of single Lyman-α absorbers exceeds that of the systems associated with heavy elements (which are thought to be largely due to intervening galaxies) by a factor of ~ 60 (but galaxies are rare objects, so they should have a huge cross-section or, equivalently, fill most of the volume);
3. The Lyman-α lines are much less clustered than the heavy element lines.

The Lyman-α forest has been extensively studied. We now believe that it is caused by relatively cold ($T \sim 10^4$ K), photo-ionized, diffuse intergalactic gas lying in the elaborate network of filaments forming the “cosmic web”.

1.3.4 The missing baryons

Gas in the Lyman-α forest at $z > 2$ accounts for at least three-quarters of the total baryon budget as inferred by both cosmic microwave background anisotropies and big-bang nucleosynthesis predictions when combined with observed light-element ratios at $z > 2$.

However, these clouds of photo-ionized intergalactic gas became more and more sparse moving towards the present and structures (galaxies, galaxy groups, and clusters) started to be assembled. Anyway, somewhat surprisingly, only a small fraction of the baryons that were present in the intergalactic medium at $z > 2$ are now found in stars, cold or warm interstellar matter, hot intracluster gas, and residual photo-ionized intergalactic medium. Nearly 50% of the baryons are now “missing” (e.g. Fukugita, Hogan & Peebles 1998, Fukugita 2003, Danforth & Shull 2005). Where are they? Is there something wrong with the big picture?

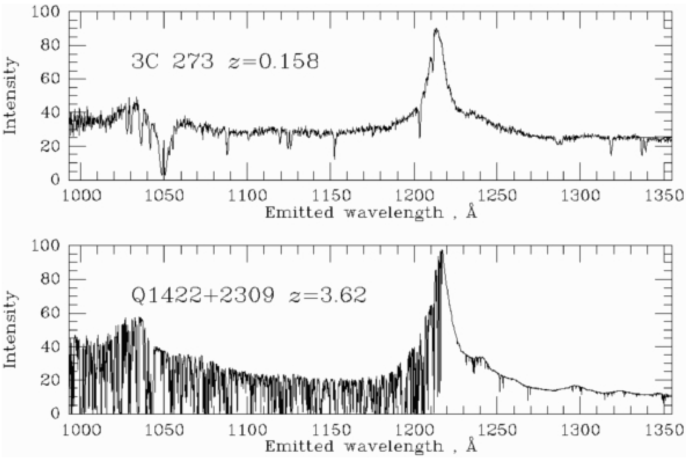


Figure 1.8: Spectra of low- and high-redshift quasars showing the thickening of the Lyman-α forest.

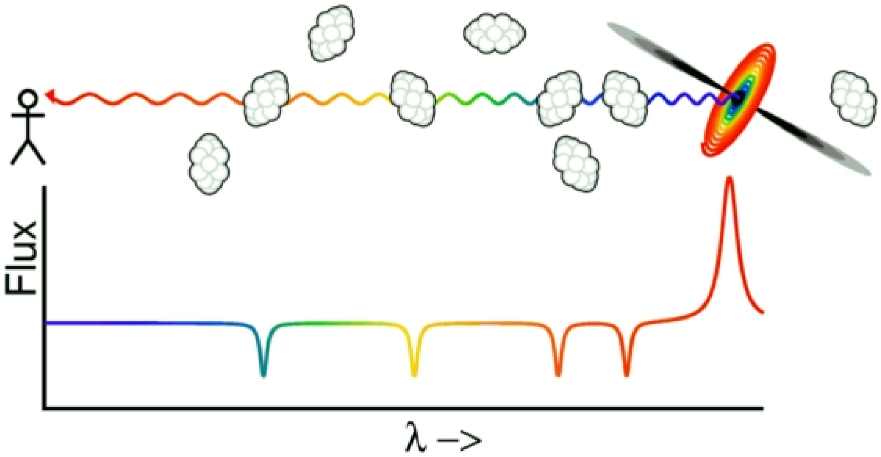


Figure 1.9: Arons explanation of the Lyman-α absorption lines. See text for further details.

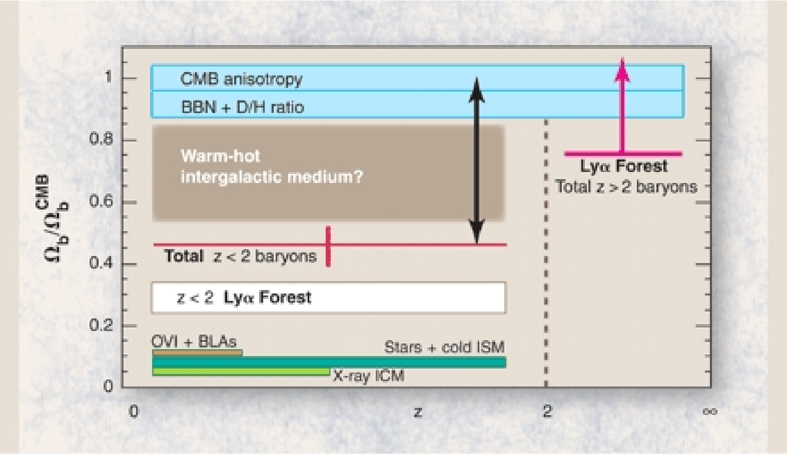


Figure 1.10: The cosmic baryon budget emphasizing the missing baryon problem. See text for details.

1.3.5 The warm-hot IGM

Given the paucity of observational findings, most of what we know about the IGM is based on numerical simulations. Within the last decade, a picture of the IGM has emerged whereby the growth of baryonic structure is regulated by the collapse of primordial perturbations via gravitational instability. According to this model, baryonic material exists in several different states. At high redshift, most of the gas is found in the Lyman-α forest, which is generally distributed and relatively cool at $T \sim 10^4$ K, its temperature governed by photo-ionization heating. As the universe evolves toward the present and density perturbations grow to form large-scale structures, baryons in the diffuse IGM accelerate toward the sites of structure formation under the influence of gravity and go through shocks that heat them to temperatures of millions of kelvin degrees. Being concentrated in a filamentary web of tenuous (baryon density $n \sim 10^6$ to 10^5 cm^3, corresponding to overdensities of $1 + \delta = n/\langle n \rangle \sim 5$ to 50. The cooling timescale for this shock-heated phase is so long that by $z = 0$ as many as 50% of the baryons may accumulate in gas with temperatures between 10^5 to 10^7 K. This matter is so highly ionized that it can only absorb or emit far-ultraviolet and soft X-ray photons, primarily at lines of highly ionized (Li-like, He-like, or H-like) C, O, Ne, and Fe (e.g. Cen & Fang 2006). Because of the extreme low density and relatively small size (1 to 10 Mpc) of the WHIM filaments, the intensity of any observable (either in emission or in absorption) is low. This makes the search for the missing baryons particularly challenging, if not impossible, with current facilities. As we will se in the course, a new generation of astronomical instruments is being developed to specifically detect the WHIM. Hopefully these will lead to a detection and confirm the simulation predictions or show a different picture altogether of where the missing baryons lie.

Chapter 2

Atomic physics

As we already discussed, astrophysical data and models of primordial nucleosynthesis indicate that matter which has never been processed by stars is almost entirely made of Hydrogen and Helium atoms. Hydrogen accounts for $\sim 75\%$ of the mass (i.e. 92% of the number of atoms) while Helium for the remaining $\sim 25\%$ (8% in number). In other words, there is approximately one atom of Helium every 12 of Hydrogen.

Before proceeding with the study of the physical processes taking place in the IGM, it is thus useful to review the basic quantum-mechanical properties of the Helium and Hydrogen atoms. This is the subject of this Chapter.

2.1 Hydrogen atom

We want to solve the quantum-mechanical problem of an electron and a proton interacting electromagnetically. As a starting point, we assume that their motion is non-relativistic.

2.1.1 Hamiltonian

Consider a proton (with charge $+e$ and mass $m_{\rm p}$) and an electron (with charge $-e$ and mass $m_{\rm e}$) which interact electromagnetically. The potential energy of the system is the usual central Coulomb potential,

$$U(r) = -\frac{e^2}{4\pi\epsilon_0 r} \;, \tag{2.1}$$

where r is the electron-proton distance. The Hamiltonian of the system is then obtained accounting for the kinetic energy of the two particles:

$$H = \frac{p_{\rm e}^2}{2m_{\rm e}} + \frac{p_{\rm p}^2}{2m_{\rm p}} + U(r) \;, \tag{2.2}$$

where p_{e} and p_{p} are the (linear) momenta of the particles. The Hamiltonian can be rewritten in terms of the momentum of the center of mass, P, and the relative momentum p

$$H = \frac{P^2}{2(m_{\mathrm{e}} + m_{\mathrm{p}})} + H' \qquad H' = \frac{p^2}{2\mu} + U(r) \; , \tag{2.3}$$

where $\mu = m_{\mathrm{e}} m_{\mathrm{p}}/(m_{\mathrm{e}} + m_{\mathrm{p}})$ is the reduced mass of the system. Note that $\mu \simeq m_{\mathrm{e}}$, since $m_{\mathrm{e}} \ll m_{\mathrm{p}}$, and the centre of mass nearly coincides with the proton. We are not interested in the translational motion of the whole system, therefore we drop the kinetic energy of the center of mass and only study the relative motion of the electron and the proton.

2.1.2 Schrödinger equation

We pass from a classical to a quantum description by applying canonical quantization. Dynamical variables (e.g. x, p) become Hermitian operators $(\hat{x}, \hat{p})$ acting on a Hilbert space of quantum states and Poisson brackets are replaced by commutators. In Schrödinger representation, a quantum state is represented by a complex-valued function of the eigenvalue of the position operator, $\psi(\mathbf{x})$. The probability that a measurement of position yields a result between $\mathbf{x}$ and $\mathbf{x} + d\mathbf{x}$ is $dP = |\psi|^2 d^3\mathbf{x}$. In this scheme, the momentum operator can be written as $\hat{p} = -i\hbar\nabla$.

The time-independent Schrödinger equation for the Hydrogen atom is obtained by requiring:

$$\hat{H}'\psi = E\psi \; . \tag{2.4}$$

Adopting spherical coordinates, $\psi = \psi(r, \theta, \phi)$, we obtain:

$$\frac{-\hbar^2}{2\mu}\nabla^2\psi + U(r)\psi = E\psi \tag{2.5}$$

with

$$\nabla^2 = \frac{1}{r^2 \sin\theta}\left[\sin\theta \frac{\partial}{\partial r}\left(r^2 \frac{\partial}{\partial r}\right) + \frac{\partial}{\partial\theta}\left(\sin\theta \frac{\partial}{\partial\theta}\right) + \frac{1}{\sin\theta}\frac{\partial^2}{\partial\phi^2}\right] \; . \tag{2.6}$$

Equation (2.5) is separable. The “spherical harmonic” functions of degree ℓ and order m, $Y_{\ell m}(\theta, \phi) = P_{\ell m}(\cos\theta)\, e^{im\phi}$ (with $P_{\ell m}$ the associated Legendre polynomials), satisfy the angular part of equation (2.5) with eigenvalue $\ell(\ell+1)$. Thus, substituting the form

$$\psi(r, \theta, \phi) = R(r)\, Y_{\ell m}(\theta, \phi) \; , \tag{2.7}$$

and writing $R(r) = u(r)/r$, one obtains

$$-\frac{\hbar^2}{2\mu}\left[\frac{d^2}{dr^2} - \frac{\ell(\ell+1)}{r^2} + \frac{2\mu e^2}{\hbar^2 4\pi\epsilon_0 r}\right] u(r) = E\, u(r) \; , \tag{2.8}$$

which is the Schrödinger equation of a fictitious particle of mass μ moving in the uni-dimensional effective potential $U_{\text{eff}}(r) = \ell(\ell+1)\hbar^2/(2\mu r^2) - e^2/(4\pi\epsilon_0 r)$. Note that, when $\ell \neq 0$, the "centrifugal potential" proportional to $\ell(\ell+1)/r^2$ counteracts the effect of the attractive Coulomb potential and becomes the dominant term for $r \to 0$.

2.1.3 Bound states

We are interested in the bound states of the electron-proton system, i.e. those with $E < 0$. In this case it is convenient to use the dimensionless radial variable $\rho = r/r_0$ with $r_0^{-1} = (-2\mu E/\hbar^2)^{1/2}$ and re-write the differential equation above in terms of the function $w(\rho)$ defined by $u(\rho) = e^{-\rho}\rho^{\ell+1}w(\rho)$. This gives

$$\rho\frac{d^2w}{d\rho^2} + 2(\ell+1-\rho)\frac{dw}{d\rho} + 2\left[\nu - (\ell+1)\right]w = 0\ , \tag{2.9}$$

where $2\nu = e^2/(4\pi\epsilon_0 r_0 E)$. It can be shown that this differential equation admits physically acceptable solutions (i.e. where $R(r)$ is finite, single valued, and square integrable) only when ν is a positive integer, $n \geq \ell+1$. In this case, (by using the new radial variable $x = 2\rho$) the radial equation reduces to the associated Laguerre equation $xw''+(1+j-x)w'+kw = 0$ (with j and k integer numbers) which is solved by the generalized (or associated) Laguerre polynomials $w(x) = L^j_{j+k}(x)$. [1]

Thus the Hydrogen atom only admits discrete energy levels for the bound states where

$$E = E_n = \frac{-\mu e^4}{8\epsilon_0^2 h^2}\frac{1}{n^2} = -\frac{13.6}{n^2}\ \text{eV} = -\frac{2.18\times10^{-18}}{n^2}\ \text{J} = -\frac{1}{n^2}\ \text{Ry} \tag{2.10}$$

with $n = 1, 2, 3, \ldots$. The corresponding solutions of the radial equation have the form[2]

$$R_{n\ell}(r) \propto \left(\frac{2r}{na_0}\right)^{\ell} L^{2\ell+1}_{n+\ell}\left(\frac{2r}{na_0}\right) e^{-\frac{r}{na_0}} \tag{2.11}$$

where the Bohr radius

$$a_0 = \frac{4\pi\epsilon_0\hbar^2}{\mu e^2} \simeq 5.29\times10^{-11}\,\text{m} = 0.529\,\text{Å} \tag{2.12}$$

[1] Unfortunately there exist two different definitions of the associated Laguerre polynomials which can generate some confusion when comparing different texts: $L^j_k = d^j/dx^j[e^x d^k/dx^k(x^k e^{-x})]$ and $L^j_k = e^x x^{-j} d^k/dx^k(e^{-x}x^{j+k})/(k!)$. We adopt the first definition. In terms of the second one, the Laguerre equation is solved by L^j_k.

[2] With the alternative definition of the associated Laguerre polynomials, the function $L^{2\ell+1}_{n+\ell}$ is replaced by $L^{2\ell+1}_{n-\ell-1}$.

defines the characteristic length scale. The first few radial eigenfunctions are:

$$\begin{aligned} R_{10}(r) &= \frac{2}{a_0^{3/2}} \exp\left(-\frac{r}{a_0}\right) , \\ R_{20}(r) &= \frac{2}{(2a_0)^{3/2}} \left(1 - \frac{r}{2a_0}\right) \exp\left(-\frac{r}{2a_0}\right) , \\ R_{21}(r) &= \frac{1}{3^{1/2}(2a_0)^{3/2}} \frac{r}{a_0} \exp\left(-\frac{r}{2a_0}\right) . \end{aligned} \tag{2.13}$$

Note that $R_{n\ell}$ has $n-\ell-1$ nodes (i.e. zero crossings) while $Y_{\ell m}$ has ℓ angular nodes (m of which around the ϕ direction and $\ell - m$ in the θ direction), so that the full wavefunction has $n-1$ nodes.

In summary, a bound state of the electron-proton system is fully specified by 4 quantum numbers:

1. The principal quantum number, n ($n = 1, 2, 3, \dots$);
2. The orbital quantum number, ℓ ($0 \leq \ell \leq n-1$);
3. The magnetic quantum number, m ($m = -\ell, -\ell+1, \dots, 0, \dots, \ell-1, \ell$ for a total of $2\ell+1$ possibilities);
4. The spin quantum number, s ($s = -1/2, +1/2$).

Since the energy of a given state only depends on the principal quantum number the degeneracy of each energy level is

$$\sum_{\ell=0}^{n-1}(2\ell+1) = n + 2\sum_{\ell=0}^{n-1}\ell = n + n(n-1) = n^2 \tag{2.14}$$

(i.e. there are n^2 possible states with the same energy). When we take into account also the two spin states of the electron, the degeneracy of the n-th energy level becomes $2n^2$ (and, if you also account for the proton spin, $4n^2$).

A special notation is commonly used to indicate bound energy levels in atomic systems. Consider the total orbital angular momentum of the atom $\mathbf{L} = \sum \mathbf{l}$ (i.e. the sum of the orbital angular momenta of all the electrons). The magnitude of the vector $\mathbf{L}$ is $\sqrt{L(L+1)}\hbar$ where L can assume non-negative integer values: 0, 1, 2, 3, ... Similarly, denote by $\mathbf{S} = \sum \mathbf{s}$ the total electronic spin of the atom (once again summing up the vector contribution of all its electrons). The magnitude of $\mathbf{S}$ is $\sqrt{S(S+1)}\hbar$. Finally, compute the total angular momentum of the atom $\mathbf{J} = \mathbf{L} + \mathbf{S}$, with magnitude $\sqrt{J(J+1)}\hbar$. A given quantum state is then associated with the letter $S, P, D, F, G, H, I, K, \dots$ according to whether its L value is $0, 1, 2, 3, 4, 5, 6, 7, \dots$. The value of $2S+1$ is written as an upper-left superscript. The value of J is written as a bottom-right subscript. For instance, the fundamental state of the hydrogen atom ($n = 0, \ell = 0$) corresponds to ${}^2S_{1/2}$. The first excited level ($n = 1$) instead includes terms ${}^2S_{1/2}$ (i.e. $\ell = 0$), ${}^2P_{1/2}$ (i.e. $\ell = 1$ where **L** and **S** are antiparallel), and ${}^2P_{3/2}$ (i.e. $\ell = 1$ where **L** and **S** are parallel).

2.1.4 Fine structure

As a first approximation, we treated the hydrogen atom as a non-relativistic system subject to Coulomb interaction. However, to achieve higher accuracy, correction to this model are required. We can distinguish three contributions.

1. Let us consider first relativistic effects. After factoring out the rest mass μc^2, the energy of the hydrogen bound states with principal quantum number n can be written as

$$|E_n| = \frac{1}{2}\mu c^2 \left(\frac{\alpha}{n}\right)^2 \tag{2.15}$$

where

$$\alpha = \frac{e^2}{\hbar c\, 4\pi\epsilon_0} = \frac{1}{137.036} \simeq 7.2974 \times 10^{-3} \tag{2.16}$$

is the dimensionless fine-structure constant. The fact that $|E_n| \ll \mu c^2$ justifies our initial assumption of a non-relativistic system. However, relativistic corrections are not completely negligible. The kinetic energy associated with the relative motion of the electron and the proton is

$$\begin{aligned} T &= (p^2c^2 + \mu^2c^4)^{1/2} - \mu c^2 = \mu c^2 \left[\left(\frac{p^2}{\mu^2c^2} + 1\right)^{1/2} - 1\right] \simeq \\ &\simeq \mu c^2 \left[\frac{1}{2}\left(\frac{p}{\mu c}\right)^2 - \frac{1}{8}\left(\frac{p}{\mu c}\right)^4 + \ldots\right] = \\ &= \frac{p^2}{2\mu} - \frac{p^4}{8\mu^3c^2} + \ldots \end{aligned} \tag{2.17}$$

where the series expansion holds for $|p| \ll \mu c$. The contribution of the term proportional to p^4 to the hydrogen energy levels can be computed perturbatively (i.e. using the unperturbed solution to calculate the correction). In this case one obtains:

$$\Delta E_n = -\frac{E_n^2}{2\mu c^2}\left(\frac{4n}{\ell + 1/2} - 3\right) , \tag{2.18}$$

which can be rewritten as

$$\frac{\Delta E_n}{E_n} = -\frac{\alpha^2}{n^2}\left(\frac{n}{\ell + 1/2} - \frac{3}{4}\right) . \tag{2.19}$$

Note that this removes the degeneracy between levels with the same principal quantum number but different orbital quantum number.

2. The electron has an intrinsic magnetic moment

$$\mathbf{M}_{\mathrm{e}} = \frac{g_{\mathrm{e}} e}{2m_{\mathrm{e}}} \mathbf{S}_{\mathrm{e}} = -g_{\mathrm{e}} \mu_{\mathrm{B}} \frac{\mathbf{S}_{\mathrm{e}}}{\hbar} , \tag{2.20}$$

where $\mathbf{S}_{\mathrm{e}}$ is the electron spin, $g_{\mathrm{e}} \simeq 2$ is the electron g factor, and $\mu_{\mathrm{B}} = e\hbar/(2m_{\mathrm{e}}) = 9.274\times10^{-24}$ J/T is the Bohr magneton. In the electron rest frame there is a magnetic field generated by the current due to the relative motion of the proton. Therefore, electromagnetic interactions are not limited to pure Coulomb attraction as assumed above. Rather, the Hamiltonian should contain an extra term $\Delta H = -\mathbf{M}_{\mathrm{e}} \cdot \mathbf{B}$ where $\mathbf{B} = -\mathbf{v} \times \mathbf{E}/c^2$ is the magnetic field felt by a charge moving with velocity $\mathbf{v}$ in the presence of an electric field $\mathbf{E} = e\mathbf{r}/(4\pi\epsilon_0 r^3)$. Therefore,

$$\Delta H = -\frac{g_{\mathrm{e}} e^2}{2m_{\mathrm{e}} c^2 4\pi\epsilon_0 r^3} \mathbf{S}_{\mathrm{e}} \cdot (\mathbf{v} \times \mathbf{r}) , \tag{2.21}$$

and using the definition of the orbital angular momentum $\mathbf{L} = m_{\mathrm{e}} \mathbf{r} \times \mathbf{v}$, we obtain the so-called spin-orbit interaction term:

$$\Delta H = \frac{e^2}{4\pi\epsilon_0} \frac{1}{2m_{\mathrm{e}}^2 c^2 r^3} \mathbf{L} \cdot \mathbf{S}_{\mathrm{e}} . \tag{2.22}$$

In reality, the expression above needs to be reduced by a factor of 2 due to the "Thomas precession" which takes into account the relativistic time dilation between the electron and the laboratory frames and the non-inertiality of the electron rest frame. The key idea is that two successive Lorentz transformations along different directions in the orbit (as the electron accelerates) are mathematically equivalent to a single Lorentz transformation plus a rotation in three-dimensional space. This rotation causes a precession of the spin vector of the electron.

Given the total angular momentum $\mathbf{J} = \mathbf{L} + \mathbf{S}$, the scalar product $\mathbf{L} \cdot \mathbf{S}$ can be written as $\mathbf{L} \cdot \mathbf{S} = (J^2 - L^2 - S^2)/2$. Therefore the energy eigenstates obtained accounting for the spin-orbit term will be most easily classified as states of definite total angular momentum where $\mathbf{L} \cdot \mathbf{S}$ assumes the values $[(j(j+1) - \ell(\ell+1) - s(s+1)]\hbar/2$. The corresponding energy shift is:

$$\Delta E_n = -\frac{1}{2n} \alpha^2 E_n \frac{j(j+1) - \ell(\ell+1) - 3/4}{\ell(\ell+1/2)(\ell+1)} . \tag{2.23}$$

3. Finally one has to consider a special term in the Hamiltonian that is different from zero only for states with $\ell = 0$.

$$\Delta H = 4\pi \frac{\hbar^2}{8m_{\mathrm{e}}^2 c^2} \left(\frac{e^2}{4\pi\epsilon_0} \right) \delta_{\mathrm{D}}(\mathbf{r}) \tag{2.24}$$

   with δ_D the Dirac-delta distribution. This is known as the "Darwin term" and naturally arises in a fully relativistic treatment of the Hydrogen atom (Dirac equation). The physical origin of the Darwin term is a phenomenon in Dirac theory called "Zitterbewegung", whereby the electron, on top of its usual steady motion, undergoes extremely rapid small-scale fluctuations on the order of the Compton wavelength $\lambda_{\mathrm{C}} = h/(m_{\mathrm{e}}c) \simeq 2.4 \times 10^{-12}$ m with a period of $\lambda_{\mathrm{C}}/c \simeq 4 \times 10^{-21}$ s. As a consequence of this "motion" the electron sees a smeared-out Coulomb potential of the nucleus which is not $U(r)$ but its average over a patch of size λ_{C}. Since the Compton wavelength is much smaller than the Bohr radius, the Zitterbewegung only matters for electrons which are very close to the nucleus. We have seen that all radial wavefunctions $R_{n\ell}$ vanish at $r = 0$ except those having $\ell = 0$. This is why the Darwin term only matters for s states.

The total energy shift (including all 3 effects) of the bound states comes out to be:

$$\Delta E_{n\ell m} = E_n \frac{\alpha^2}{n^2} \left(\frac{n}{j + 1/2} - \frac{3}{4} \right) \qquad j = \ell \pm 1/2 \tag{2.25}$$

Note that although the three separate contributions depend on ℓ, the total energy shift does not: it only depends on j. This degeneracy is present even in the exact solution of the Dirac equation for the Coulomb potential.

In brief, what at first approximation appears to be a single energy level in the hydrogen spectrum actually consists of two or more closely spaced level when analysed with high precision. The spacing between the levels of this "fine structure" is suppressed by a factor $\alpha^2 \simeq 10^4$ with respect to the principal levels.

2.1.5 Hyperfine structure and 21cm radiation

Also the proton, being a spin one-half particle, possesses an intrinsic magnetic moment with a g factor $g_{\mathrm{p}} = 5.586$. [3] The interaction between the magnetic moment of the proton with the magnetic field produced by the electron spin and the orbital motion of the electron (nuclear spin-orbit interacton) give rise to additional small perturbations in the energy levels of

[3] Note that the magnetic moment of a proton is much smaller than that of an electron. The ratio $\mu_{\mathrm{e}}/\mu_{\mathrm{p}}$ is of the same order of $m_{\mathrm{p}}/m_{\mathrm{e}} \sim 1836$.

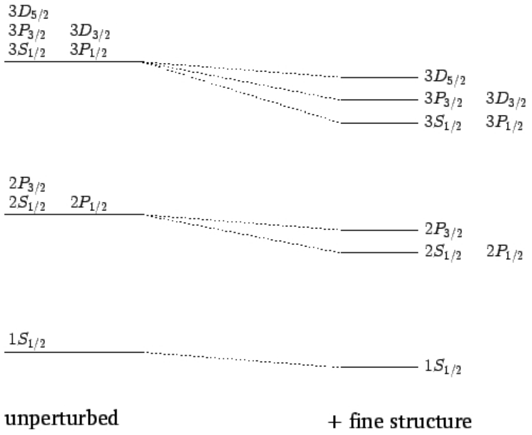


Figure 2.1: Effect of the fine-structure energy-shift on the $n = 1, 2$, and 3 states of the Hydrogen atom. Not to scale!

the hydrogen atom. Since the Hamiltonian contains terms which are proportional to the proton spin $\mathbf{S}_\mathrm{p}$, the operator that measures the angular momentum of the electron ($\mathbf{J} = \mathbf{L} + \mathbf{S}_\mathrm{e}$) does not commute with the full Hamiltonian. However, the operators F^2 and F_z (where $\mathbf{F}$ denotes the total angular momentum $\mathbf{J} + \mathbf{S}_\mathrm{p}$) do. Hence every energy level associated with a particular set of quantum numbers n, ℓ, and j will be split into two levels of slightly different energy depending on the relative orientation of the proton and electron spins. The amplitude of these splittings is typically a factor $m_\mathrm{e}/m_\mathrm{p}$ smaller than for the fine structure and for this reason it is dubbed "hyperfine structure".

For the specific case of the ground state of the hydrogen atom, the energy separation between the states $f = 1$ and $f = 0$ is 5.9×10^{-6} eV. This corresponds to radiative transitions with frequency $\nu = 1420.406$ MHz and wavelength $\lambda = 21.1$ cm. This is the source of the "21cm line" which has been used by astronomers to map the distribution of neutral hydrogen in our galaxy.

2.1.6 Lamb shift

In 1947 Willis Lamb (1913-2008, Nobel laureate in 1955) and Robert Retherford (1912-) showed that the $2\,S_{1/2}$ ($n = 2, \ell = 0, j = 1/2$) and $2\,P_{1/2}$ ($n = 2, \ell = 1, j = 1/2$) states of the hydrogen atom are not degenerate. The P state is slightly more bound with an energy difference $\Delta E = 4.372 \times 10^{-6}$ eV (corresponding to a transition frequency between the two states of 1057.864 MHz).

The effect is now explained by treating the electromagnetic field as a quantum system. In this case, the ground state of the electromagnetic field

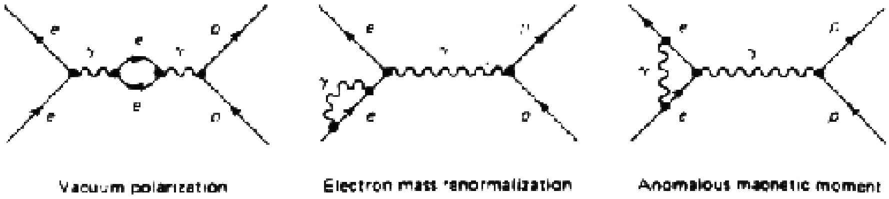


Figure 2.2: Feynman loop diagrams showing effects that contribute to the Lamb shift.

has non-vanishing energy (as in the case of the harmonic oscillator) and thus a non-vanishing field. Radiative coupling of the electron to the vacuum field produces the Lamb shift. At the lowest perturbative level (one loop) in quantum electrodynamics one recognizes three contributions to the Lamb shift (see Figure 2.2). The dominant one (4.2 μeV) comes from the fact that electrons subjected to the electromagnetic field spontaneously emit photons and quickly reabsorb them. [4] This "self-interaction" of the electron slightly changes the energy that binds the electron to the proton (electron mass renormalization). The second largest contribution (0.28 μeV) comes from the fact that one-loop corrections produce an anomalous magnetic dipole moment for the electron with a g-factor $g = 2.00232$ instead of the standard Dirac-value of 2. The last contribution (-0.11 μeV) comes from virtual electron-positron pairs which, in the presence of an electromagnetic field, align and create electric dipoles counteracting the external field (vacuum polarization).

For states with $\ell = 0$ the Lamb-shift correction to the energy levels is:

$$\Delta E_{\text{Lamb}} = \alpha^5 \, m_{\text{e}} c^2 \, \frac{k(n, \ell = 0)}{4n^3} \tag{2.26}$$

where $k(n, \ell = 0)$ is a tabulated function which varies slightly with n and assumes values between 12.7 (for $n = 1$) and 13.2 (as $n \to \infty$). For $\ell \neq 0$ the Lamb shift is very small. In this case:

$$\Delta E_{\text{Lamb}} = \alpha^5 \, m_{\text{e}} c^2 \, \frac{1}{4n^3} \left[k(n, \ell) \pm \frac{1}{\pi (j + 1/2)(\ell + 1/2)} \right] , \tag{2.27}$$

for $j = \ell \pm 1/2$, where $|k(n, \ell)| < 0.05$ is a small numerical factor which varies slightly with n and ℓ.

[4] Using the words by Gordon Kane: "Quantum mechanics allows, and indeed requires, temporary violations of conservation of energy, so one particle can become a pair of heavier particles (the so-called virtual particles), which quickly rejoin into the original particle as if they had never been there. But while the virtual particles are briefly part of our world they can interact with other particles."

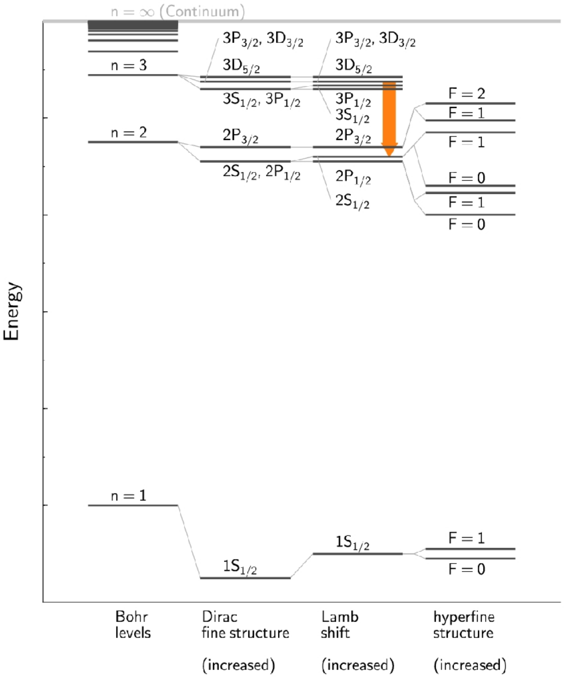


Figure 2.3: A scheme showing the energy levels of the Hydrogen atoms at different approximation levels. Moving from left to right one includes more and more terms in the Hamiltonian.

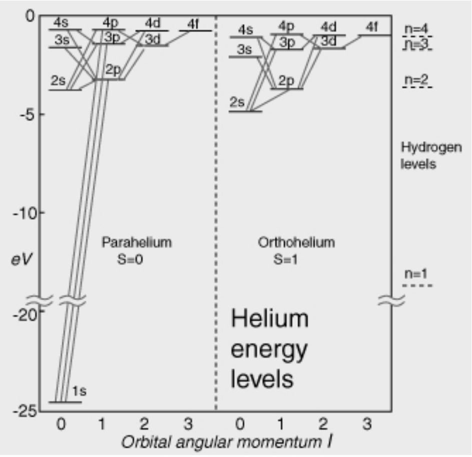


Figure 2.4: Energy levels for neutral Helium obtained assuming that the first electron is in the ground state $1s$.

2.2 Helium atom

A Helium atom consists of a nucleus of charge $+2e$ surrounded by two electrons. Although there are eight known isotopes of Helium (containing a different number of neutrons in the nucleus), only Helium 3 (two protons and one neutron) and Helium 4 (two protons and two neutrons) are stable. On Earth only 0.000137% of the Helium atoms are Helium 3 ($\sim$ 1 parts per million), all the rest is Helium-4. In outer space the Helium 3 abundance is higher, for instance in the atmosphere of Jupiter it amounts to 100 ppm.

2.2.1 Singly ionized Helium

The He$^+$ atom is just like a hydrogen atom with a nuclear charge $+2e$ ($Z = 2$). Since the energy levels depend upon the square of the nuclear charge, the energy levels of the atom are:

$$E_n = \frac{-\mu e^4 Z^2}{8\epsilon_0 h^2}\frac{1}{n^2} = -\frac{54.4}{n^2}\ \text{eV} \quad n = 1, 2, 3, \dots\ . \tag{2.28}$$

2.2.2 Neutral Helium

Adding a second electron one obtains the neutral Helium atom. In this case there is no analytic solution for the energy levels that can however be computed numerically or using perturbative techniques (Figure 2.4). One has to distinguish two cases:

1. “Parahelium”: where the spins of the electrons are antiparallel (i.e. the total spin quantum number is $S = 0$, a singlet state);

2. "Orthohelium": where the spins of the electrons are parallel (i.e. the total spin quantum number is $S = 1$, a triplet state).

Parahelium is energetically the lowest state of Helium. In this case, the ground state of the second electron corresponds to $E = -24.6$ eV. For all the other energy levels, however, orthohelium is a slightly more bound system than parahelium. Let us try to understand why the energy levels of ortho- and parahelium are different. In the case of orthohelium, the spin part of the wavefunction is symmetric (i.e. one can exchange the two electrons without noticing a difference). However, the total wavefunction for electrons (and, in general, for all fermions) must be anti-symmetric to obey the Pauli exclusion principle. Note that, neglecting to first approximation the spin-orbit interaction, the total wavefunction can be written as the product of the spin and space parts: $\psi_{\rm tot} = \psi_{\rm r}(\mathbf{r}_1, \mathbf{r}_2)\, \psi_{\rm s}(\mathbf{s}_1, \mathbf{s}_2)$. Therefore, the space part of the wavefunction must be anti-symmetric for orthohelium (and for this reason both electrons cannot be in the 1s state as shown in Fig. 2.4). An anti-symmetric function of position must vanish at zero separation. This suggests that electrons tend to be more separated for orthohelium than for parahelium. Therefore, the second electron in orthohelium will feel less shielding from the nucleus by the other electron and it will be more tightly bound. This reasoning is often called "spin-spin interaction" and lies at the base of the first Hund's rule for the ordering of energy levels in multi-electronic atoms.

2.3 Atoms and electromagnetic radiation

In this Section we will discuss how atomic systems interact with radiation. For simplicity, we will adopt a semi-classical treatment, where the atom is described quantum-mechanically and radiation is treated classically. We will stress the limits of this large-photon-number approximation and briefly discuss where it fails.

To simplify the notation, we adopt Gauss units for the electromagnetic quantities.

2.3.1 Time evolution

The time evolution of a (non-relativistic) quantum system is described in terms of its Hamiltonian, H, by the time-dependent Schrödinger equation

$$i\hbar \frac{\partial \psi}{\partial t} = \hat{H}\psi \; . \tag{2.29}$$

Let us consider a system (e.g. an atom) with an Hamiltonian H_0 that does not contain time explicitly. At an arbitrary time t_0, this system is characterized by discrete energy states E_n corresponding to the wavefunctions ψ_n (i.e.

the solutions of the time-independent Schrödinger equation $\hat{H}\psi_n = E_n\psi_n$) that satisfy the orthonormality relation

$$\int \psi_n^*\, \psi_m\, d^3x = \langle n|m\rangle = \delta_{nm} \equiv \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{2.30}$$

where the first equality introduces the bra-ket notation by Paul Dirac and defines the inner product in Hilbert space. For a state of defined energy, Eq. (2.29) gives

$$i\hbar \frac{\partial \psi_n}{\partial t} = E_n\, \psi_n \to \psi_n(t) = \psi_n \exp\left[-\frac{i\, E_n\, (t - t_0)}{\hbar}\right] . \tag{2.31}$$

Suppose that at time $t = t_0$ the state of the system is

$$\phi(t_0) = \sum_n c_n\, \psi_n \tag{2.32}$$

where the c_n are complex numbers. Its time evolution will thus be

$$\phi(t) = \sum_n c_n \exp\left[-\frac{i\, E_n\, (t - t_0)}{\hbar}\right] \psi_n \tag{2.33}$$

and the probability of finding the system in state n at time t is obtained by projecting the wavefunction ϕ onto the eigenspace generated by ψ_n:

$$P_n(t) = |\int \phi^*(t)\, \psi_n\, d^3x|^2 = |\langle \phi|n\rangle|^2 = |c_n|^2 = P_n(t_0) . \tag{2.34}$$

In this case, probabilities do not change with time. In particular, a system in an energy eigenstate $|n\rangle$ indefinitely stays in that state.

2.3.2 Time-dependent perturbations

Let us now perturb our system adding a small time-dependent Hamiltonian $H_1(t)$:

$$H = H_0 + H_1(t) . \tag{2.35}$$

You might recall from your class on quantum mechanics that the eigenfunctions of any Hermitian (self-adjoint) operator form a complete set of functions in Hilbert space. This is known as the "spectral theorem" and means that we can write any wavefunction as a linear combination of the elements forming the complete set. We can then still use the eigenfunctions of $\hat{H}_0$ to describe the evolution of the perturbed system but the coefficients of the expansion, c_n, will be time dependent in this case, i.e. $\phi(t) = \sum_n c_n(t)\, \psi_n(t)$. The initial state $\phi(t_0)$ will thus evolve according to

$$i\hbar \sum_m \frac{dc_m}{dt}\, \psi_m(t) = \sum_m c_m(t)\, \hat{H}_1\, \psi_m(t) \tag{2.36}$$

where we have used the unperturbed Schrödinger equation for $\psi_n(t)$ to eliminate $\hat{H}_0$ from the equation. Taking the inner product with a given ψ_n then gives

$$i\hbar\frac{dc_n}{dt} = \sum_m H_{nm}(t)\exp\left[i\,\omega_{nm}(t-t_0)\right]c_m(t) \tag{2.37}$$

with

$$H_{nm}(t) = \int \psi_n^*\,\hat{H}_1(t)\,\psi_m\,d^3x = \langle n|\hat{H}_1|m\rangle \tag{2.38}$$

the interaction matrix element and

$$\omega_{nm} = \frac{E_n - E_m}{\hbar}\ . \tag{2.39}$$

This is a set of coupled differential equations for the $c_n(t)$. In matrix form, it gives:

$$i\hbar\frac{d}{dt}\begin{pmatrix} c_1 \\ c_2 \\ . \\ c_n \end{pmatrix} = \begin{pmatrix} H_{11} & H_{12}\,e^{i\,\omega_{12}\,t} & . & H_{1n}\,e^{i\,\omega_{1n}\,t} \\ H_{21}\,e^{i\,\omega_{21}\,t} & H_{22} & . & H_{2n}\,e^{i\,\omega_{2n}\,t} \\ . & . & . & . \\ H_{n1}\,e^{i\,\omega_{n1}\,t} & H_{n2}\,e^{i\,\omega_{n2}\,t} & . & H_{nn} \end{pmatrix}\ . \tag{2.40}$$

The probability of finding the system in any particular state at any later time is $|c_n(t)|^2$. We can thus say that a (small) time-dependent perturbation causes the quantum system to make transitions between its unperturbed energy eigenstates. Note, however, that the states $|n\rangle$ have a definite energy only for the unperturbed Hamiltonian. They will not be energy eigenstates for the perturbed one. However, each of them will be representable as a superposition of the (unknown) "true" eigenstates of the full Hamiltonian. The energy of each state $|n\rangle$ is thus not exactly defined in the presence of the perturbation.

2.3.3 Harmonic perturbations

Let us now consider a quantum system in an initial state $|i\rangle$ perturbed by a periodic potential $H_1(t) = V\exp(-i\omega t)$ (where V does not depend explicitly on time) switched on at $t = 0$. We are interested to know the probability that the system will be in the state $|f\rangle$ at time $t > 0$. To first order (i.e. for small changes), we can assume that $c_i \simeq 1$ all the time and neglect transitions to $|f\rangle$ from other states than $|i\rangle$. Plugging this into equation (2.37) we get:

$$i\hbar\,\frac{dc_f}{dt} \simeq H_{fi}(t)\,\exp\left(i\,\omega_{fi}\,t\right) \tag{2.41}$$

which gives

$$c_f(t) = -\frac{i}{\hbar}\,\langle f|\hat{V}|i\rangle\,\int_0^t \exp\left[i\left(\omega_{fi}-\omega\right)t\right]dt\ . \tag{2.42}$$

Note that the time integral gives the Fourier transform of the pertubation. The final result is

$$c_f(t) = -\frac{i}{\hbar} \langle f|\hat{V}|i\rangle \frac{\exp\left[i\left(\omega_{fi}-\omega\right)t\right]-1}{i\left(\omega_{fi}-\omega\right)} \ . \tag{2.43}$$

The transition probability to the state $|f\rangle$ is then:

$$P_{i\to f} = |c_f|^2 = \frac{2}{\hbar^2} \left|\langle f|\hat{V}|i\rangle\right|^2 \frac{1-\cos\left[\left(\omega_{fi}-\omega\right)t\right]}{\left(\omega_{fi}-\omega\right)^2} \ , \tag{2.44}$$

and, using the identity $2\sin^2(x) = 1-\cos(2x)$, it can be rewritten as

$$P_{i\to f} = \frac{4}{\hbar^2} \left|\langle f|\hat{V}|i\rangle\right|^2 \frac{\sin^2\left[\left(\omega_{fi}-\omega\right)\frac{t}{2}\right]}{\left(\omega_{fi}-\omega\right)^2} \ . \tag{2.45}$$

The oscillatory function on the right-hand side (see Figure 2.5) has a maximum at $\omega = \omega_{fi}$ where it assumes the value $t^2/4$. The corresponding peak has a width of $2\pi/t$. As time passes, the transition probability integrated over all possible angular frequencies ω increases proportionally to t (note that after a short transient this probability mainly comes from the main peak around ω_{fi}). We are interested in the large t limit, where there are no effects due to the fact that we artificially “switched-on” the perturbation at $t = 0$. It makes sense, therefore, to define a “transition rate per unit time” as

$$R_{i\to f} = \lim_{t\to\infty} \frac{P_{i\to f}}{t} \ . \tag{2.46}$$

Taking into account that $\lim_{x\to\infty} \sin^2(\beta x)/(\beta^2 x) = (\pi/2)\,\delta_{\mathrm{D}}(\beta)$, we finally obtain:

$$R_{i\to f} = \frac{2\pi}{\hbar^2} \left|\langle f|\hat{V}|i\rangle\right|^2 \delta_{\mathrm{D}}(\omega_{fi}-\omega) \ . \tag{2.47}$$

This equality is known as the (second) *Fermi golden rule*. It states that, to first order in perturbation theory, the transition rate depends only on the square of the matrix element of the operator $\hat{V}$ between initial and final states, $|V_{fi}|^2 = |\langle f|\hat{V}|i\rangle|^2$. Also it enforces energy conservation via the Dirac delta distribution. The condition $\omega = \omega_{fi}$ is equivalent to the condition $E_f = E_i + \hbar\omega$. This means that the transition can only occur if the frequency of the field is exactly “tuned” to match the energy difference between the initial and final states.

In this context, it is interesting to address a couple of questions that might naturally arise.

Why at finite time t transitions are possible also for $\omega \neq \omega_{fi}$?

The issue is related to the fact that we suddenly switched on the

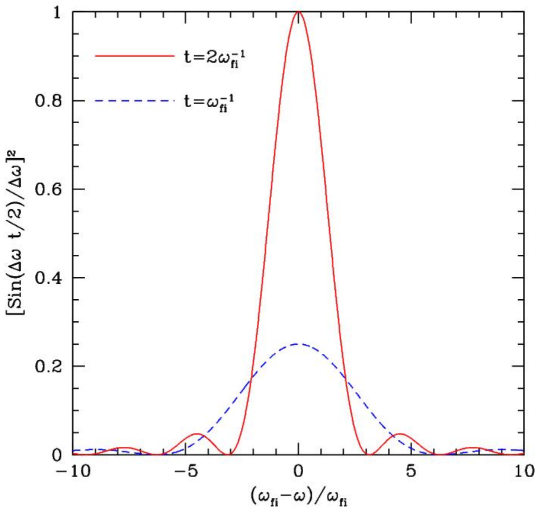


Figure 2.5: The oscillatory function in equation 2.45 at two different times. The integral of this function over the angular frequency of the incident radiation ω gives $\pi\, t/2$.

oscillating perturbation at $t = 0$. Because of this, the frequency spectrum of H_1 is not exactly monochromatic at the frequency $\omega/(2\pi)$ (i.e. $S(\nu)d\nu \propto \delta_{\rm D}[\nu - (\omega/2\pi)]\, d\nu$) at finite times. [5] Rather it follows:

$$S(\nu) \propto \frac{\sin^2\left[\left(\nu - \dfrac{\omega}{2\pi}\right)\dfrac{t}{2}\right]}{\left(\nu - \dfrac{\omega}{2\pi}\right)^2} \; . \tag{2.48}$$

Therefore the perturbation contains harmonic modes with $\nu = \omega_{fi}/(2\pi)$ (those that conserve energy during the transition) even though $\omega \neq \omega_{fi}$. The presence of these modes makes the transition possible and the transition probability is then proportional to $S[\omega_{fi}/(2\pi)]$. Only in the limit $t \to \infty$ the perturbation will have a truly monochromatic spectrum.

What does the limit $t \to \infty$ in equation (2.47) really mean in practice?

The width of the main peak in Figure 2.5 is $\Delta\omega\, t/2 = \pi$ which we can write as $\Delta\omega = 2\pi/t$. Requiring that this frequency width is much smaller that the energy difference between the initial and final states divided by the Planck constant, gives:

$$\frac{\Delta\omega}{\omega_{fi}} \ll 1 \quad \to \quad t \gg \frac{2\pi}{\omega_{fi}} = \frac{1}{\nu_{fi}} \; . \tag{2.49}$$

[5] Remember that the frequency spectrum is the square modulus of the Fourier transform of a time-dependent signal.

When this is true, one can extend the limit to infinity without appreciably changing the result. On the other hand, in our derivation, we assumed that the transition probability $P_{i\to f} \ll 1$. This condition allowed us to use first-order corrections (otherwise higher-order terms should be taken into account). Requiring that the peak of the transition probability is smaller than unity gives

$$\frac{|V_{fi}|^2}{\hbar^2}\,t^2 \ll 1 \quad \text{or} \quad t \ll \frac{\hbar}{|V_{fi}|}. \tag{2.50}$$

Therefore, our derivation only holds if

$$\frac{2\pi}{\omega_{fi}} \ll t \ll \frac{\hbar}{|V_{fi}|} \tag{2.51}$$

which requires

$$|V_{fi}| \ll \frac{\hbar\omega_{fi}}{2\pi} = \frac{E_f - E_i}{2\pi}\ . \tag{2.52}$$

In brief, the matrix element $|V_{fi}|$ must be much smaller than the energy difference between the quantum states involved in the transition. This generally holds when applied to atomic systems interacting with electromagnetic radiation in normal conditions.

The Fermi golden rule must be modified when one considers transitions to a continuum distribution of final states characterized by a (broad and smooth) density $\rho(E)$ (number of states per unit energy). In this case, it becomes: [6]

$$R_{i\to f} = \frac{2\pi}{\hbar}\ |\langle f|V|i\rangle|^2\ \rho(E_f)\ . \tag{2.53}$$

This for instance applies to states of an ionized atom where the electron is unbound to the nucleus.

2.3.4 Electromagnetic interactions

Let us now consider the special case of an atomic electron interacting with classical (i.e. non quantized) electromagnetic radiation which is switched on at time $t = 0$. In this case

$$H_1 \simeq -\frac{e\mathbf{A}(t)\cdot\mathbf{p}}{m_e\,c}\ , \tag{2.54}$$

with $\mathbf{A}$ the vector potential of the electromagnetic field. This is obtained as follows (see any basic quantum mechanics textbook for details):

1. Using the standard prescription $\mathbf{p} \to \mathbf{p} + q\mathbf{A}$, $H \to H - q\phi$ for writing the Hamiltonian of a particle of charge q in the presence of an electromagnetic field with scalar potential ϕ and vector potential $\mathbf{A}$.

[6] Remember that $\delta_{\rm D}(\omega) = \hbar\,\delta_{\rm D}(\hbar\omega) = \hbar\,\delta_{\rm D}(E)$.

2. Separating the electromagnetic field in a static component (the Coulomb potential due to the rest of the atom) and radiation. In the rest frame of the nucleus, this is done by adopting the Coulomb gauge $\nabla \cdot \mathbf{A} = 0$ (also known as transverse or radiation gauge). In this case, the radiation field is fully described by the (transverse) vector potential and has $\phi = 0$. The electric and magnetic fields of the radiation component are given by $\mathbf{E} = -(1/c)\,\partial \mathbf{A}/\partial t$ and $\mathbf{B} = \nabla \times \mathbf{A}$, respectively.

3. Supposing that the perturbation corresponds to a monochromatic plane wave with $\mathbf{A} = 2\,A_0\,\mathbf{u}\cos(\mathbf{k}\cdot\mathbf{r} - \omega t)$. The electromagnetic wave propagates along the direction of the wavevector $\mathbf{k}$ (whose modulus corresponds to the wavenumber $k = |\mathbf{k}| = \omega/c$) whilst the vector potential is harmonically oscillating along the direction $\mathbf{u} = \mathbf{A}/|\mathbf{A}|$. Note that $\mathbf{u}$ and $\mathbf{k}$ are perpendicular, i.e. $\mathbf{u} \times \mathbf{k} = 0$.

4. Neglecting a quadratic term in $\mathbf{A}$ (since we are after small perturbations):
$$H_2 = \frac{e^2}{2m_\mathrm{e}c^2}\,\mathbf{A}\cdot\mathbf{A}\ . \tag{2.55}$$
We will return to this term later, in Section 2.4.1.

Expliciting the time dependence in equation (2.54), we obtain

$$H_1 = -\frac{eA_0}{m_e\,c}\,\mathbf{u}\cdot\mathbf{p}\,[\exp(i\,\mathbf{k}\cdot\mathbf{r} - i\omega t) + \exp(-i\,\mathbf{k}\cdot\mathbf{r} + i\omega t)]\ . \tag{2.56}$$

This is the linear superposition of two harmonic perturbations and we can reason as in Section 2.3.3. The Fermi golden rule then gives

$$R_{i\to f} = \frac{2\pi}{\hbar^2}\left[|\langle f|V|i\rangle|^2\,\delta_\mathrm{D}(\omega_{fi} - \omega) + |\langle f|V^\dagger|i\rangle|^2\,\delta_\mathrm{D}(\omega_{fi} + \omega)\right]\ , \tag{2.57}$$

where
$$V = -\frac{eA_0}{m_e\,c}\,\mathbf{u}\cdot\mathbf{p}\,\exp(i\,\mathbf{k}\cdot\mathbf{r})\ . \tag{2.58}$$

Note that the transition rate scales as A_0^2 and thus as the square modulus of the electric field which is directly proportional to the intensity of radiation (the time average of the Poynting flux along the $\mathbf{k}$ direction).

The terms associated with the two delta distributions correspond to two different physical processes caused by radiation of angular frequency ω impinging onto an atomic system.

Absorption. The first term on the left describes a process by which the atomic system gains the energy $\Delta E = \hbar\omega$ from the perturbing field, whilst making a transition to a final state whose energy level exceeds that of the initial state by ΔE. This process is known as absorption (Fig. 2.6).

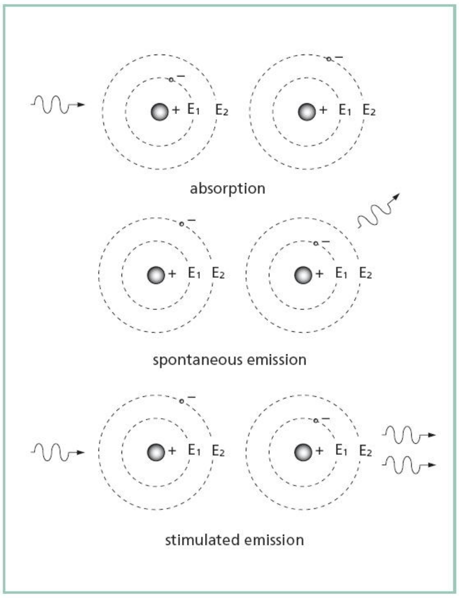


Figure 2.6: Schematic representation of stimulated emission, absorption and spontaneous emission.

Stimulated emission. The latter term on the right describes a process by which the atomic system gives up energy $\Delta E = \hbar\omega$ to the electromagnetic field while making a transition to a final state whose energy level is less than that of the initial state by ΔE. This process is called stimulated (or induced) emission (Fig. 2.6).

In both cases, the total energy (i.e. that of the atomic system plus the perturbing field) is conserved.

2.3.5 The electric dipole approximation

Typically, atomic energy levels have separations of a few eV. A radiative transition between two energy levels separated by $\Delta E = 1$ eV produces radiation with a wavelength $\lambda = hc/\Delta E = 12398.44$ Å. On the other, photons with an energy of 13.6 eV (the ionization threshold for Hydrogen in the ground state) have $\lambda = 911$ Å. This is still much larger that the typical size of a light atom (atomic wavefunctions typically only extend for a few Å).

Now note that the electromagnetic perturbation V in equation (2.58) scales proportionally to $\exp[i\,\mathbf{k}\cdot\mathbf{r}]$ where $k = 2\pi/\lambda$. The argument of the exponential function is therefore expected to be a small number of the order of the ratio between the atom size and the wavelength of the electromagnetic wave. It makes thus sense to Taylor expand it:

$$\exp[i\,\mathbf{k}\cdot\mathbf{r}] \simeq 1 + i\,\mathbf{k}\cdot\mathbf{r} - \frac{1}{2}(\mathbf{k}\cdot\mathbf{r})^2 + \ldots \ . \qquad (2.59)$$

Keeping only the leading term (i.e. 1) corresponds to assuming that the electric field is constant in space (but not in time!) within the atomic volume. This is known as the *electric dipole approximation*:

$$V_{fi} \simeq -\frac{eA_0}{m_e\, c}\, \mathbf{u} \cdot \langle f|\hat{\mathbf{p}}|i\rangle \ . \tag{2.60}$$

The electron linear momentum $\mathbf{p}$ can be expressed in terms of the commutator $[r, H_0]$ as $\mathbf{p} = i\,(m_\mathrm{e}/\hbar)\,[r, H_0]$ (where H_0 is the usual unperturbed Hamiltonian including a non-relativistic kinetic term and the Coulomb interaction). Therefore,

$$\langle f|\hat{\mathbf{p}}|i\rangle = i\, m_\mathrm{e}\, \omega_{fi}\, \langle f|\hat{\mathbf{r}}|i\rangle \tag{2.61}$$

and

$$V_{fi} \simeq -\frac{eA_0}{c}\, i\, \omega_{fi}\, \hat{\mathbf{u}} \cdot \langle f|\hat{\mathbf{r}}|i\rangle \ . \tag{2.62}$$

The matrix element V_{fi} is thus proportional to

$$\langle f|\hat{\mathbf{d}}|i\rangle = -\langle f|e\,\hat{\mathbf{r}}|i\rangle \tag{2.63}$$

which gives the (effective) electric dipole moment for the transition.

2.4 Selection rules and forbidden transitions

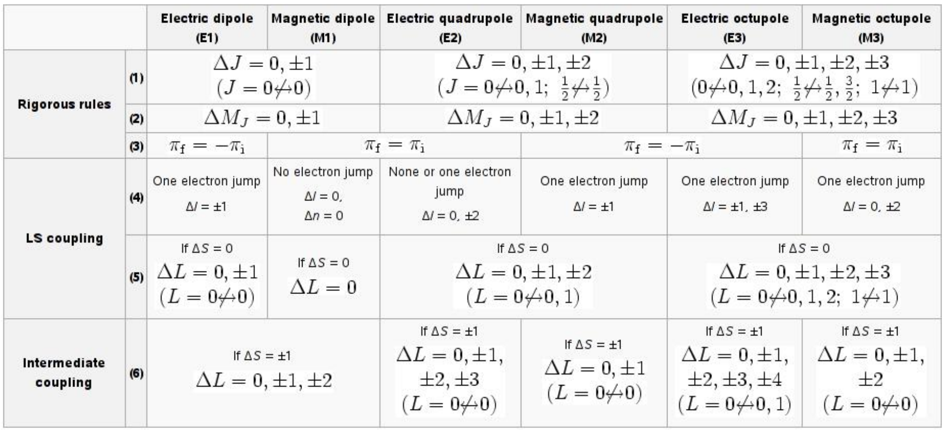
In the electric dipole approximation, radiative atomic transitions can only happen if the associated electric dipole moment does not vanish. Consider a transition from an energy state corresponding to the quantum numbers n, ℓ, m to one with n', ℓ', m'. Using the fact that the position $\mathbf{r}$ is a vector (i.e. a spin-one tensorial operator), it can be shown that the dipole matrix element vanishes unless the initial and final states satisfy (see any quantum mechanics book for details):

$$\Delta\ell = \ell' - \ell = \pm 1 \ , \qquad \Delta m = m' - m = 0, \pm 1 \ . \tag{2.64}$$

Moreover, since the perturbing Hamiltonian does not contain any spin operators, the spin quantum number m_s cannot change during a transition. Hence, we have the additional selection rule that $m_s' = m_s$. These are termed the *selection rules* for electric dipole transitions. They follow from the Wigner-Eckart theorem and derive from the properties of the spherical harmonics.

Strictly speaking the above selection rules are only valid when the spin-orbit interaction is not considered in the unperturbed Hamiltonian of the atom. In the general case, the eigenstates of H_0 are classified by the quantum numbers n, ℓ, m, j, m_j. The dipole selection rules then become:

$$\Delta J = 0, \pm 1 \text{ (except } 0 \to 0) \ , \qquad \Delta M_J = 0, \pm 1 \qquad \Delta L = \pm 1 \qquad \Delta S = 0 \ , \tag{2.65}$$



		Electric dipole (E1)	Magnetic dipole (M1)	Electric quadrupole (E2)	Magnetic quadrupole (M2)	Electric octupole (E3)	Magnetic octupole (M3)
Rigorous rules	(1)	$\Delta J = 0, \pm 1$ $(J = 0 \not\leftrightarrow 0)$		$\Delta J = 0, \pm 1, \pm 2$ $(J = 0 \not\leftrightarrow 0, 1;\ \frac{1}{2} \not\leftrightarrow \frac{1}{2})$		$\Delta J = 0, \pm 1, \pm 2, \pm 3$ $(0 \not\leftrightarrow 0, 1, 2;\ \frac{1}{2} \not\leftrightarrow \frac{1}{2}, \frac{3}{2};\ 1 \not\leftrightarrow 1)$	
	(2)	$\Delta M_J = 0, \pm 1$		$\Delta M_J = 0, \pm 1, \pm 2$		$\Delta M_J = 0, \pm 1, \pm 2, \pm 3$	
	(3)	$\pi_\mathrm{f} = -\pi_\mathrm{i}$	$\pi_\mathrm{f} = \pi_\mathrm{i}$		$\pi_\mathrm{f} = -\pi_\mathrm{i}$		$\pi_\mathrm{f} = \pi_\mathrm{i}$
LS coupling	(4)	One electron jump $\Delta l = \pm 1$	No electron jump $\Delta l = 0$, $\Delta n = 0$	None or one electron jump $\Delta l = 0, \pm 2$	One electron jump $\Delta l = \pm 1$	One electron jump $\Delta l = \pm 1, \pm 3$	One electron jump $\Delta l = 0, \pm 2$
	(5)	If $\Delta S = 0$ $\Delta L = 0, \pm 1$ $(L = 0 \not\leftrightarrow 0)$	If $\Delta S = 0$ $\Delta L = 0$	If $\Delta S = 0$ $\Delta L = 0, \pm 1, \pm 2$ $(L = 0 \not\leftrightarrow 0, 1)$		If $\Delta S = 0$ $\Delta L = 0, \pm 1, \pm 2, \pm 3$ $(L = 0 \not\leftrightarrow 0, 1, 2;\ 1 \not\leftrightarrow 1)$	
Intermediate coupling	(6)	If $\Delta S = \pm 1$ $\Delta L = 0, \pm 1, \pm 2$		If $\Delta S = \pm 1$ $\Delta L = 0, \pm 1,$ $\pm 2, \pm 3$ $(L = 0 \not\leftrightarrow 0)$	If $\Delta S = \pm 1$ $\Delta L = 0, \pm 1$ $(L = 0 \not\leftrightarrow 0)$	If $\Delta S = \pm 1$ $\Delta L = 0, \pm 1,$ $\pm 2, \pm 3, \pm 4$ $(L = 0 \not\leftrightarrow 0, 1)$	If $\Delta S = \pm 1$ $\Delta L = 0, \pm 1,$ ± 2 $(L = 0 \not\leftrightarrow 0)$

Figure 2.7: Selection rules for radiative transitions.

and express the fact that the photon is a spin-1 particle carrying one unit of angular momentum.

Transitions that are not allowed are called (dipole) “forbidden”. Among these we must distinguish two cases.

1. Those for which the true radiative matrix $\langle f|\mathbf{u}\cdot\mathbf{p}\ \exp\left[i\,\mathbf{k}\cdot\mathbf{r}\right]|i\rangle$ is zero are absolutely forbidden at the one-photon level. For instance, the $2\,S \rightarrow 1\,S$ transition in the Hydrogen atom is one of these and cannot take place.

2. On the other hand, the dipole-forbidden transitions for which the true radiative matrix does not vanish are not strictly forbidden. They take place at a far lower rate than transitions which are allowed according to the electric dipole approximation. In fact, the higher order terms in equation (2.59) give rise to transitions with different selection rules. For instance, the linear correction proportional to $\mathbf{r}$ generates two new families of radiative transitions.

   - *Magnetic dipole transitions* (interactions between the electron spin and orbital angular momentum with the oscillating magnetic field of the incident radiation) are typically 10^5 times more unlikely than similar electric dipole transitions.
   - *Electric quadrupole transitions* (interactions between the electric quadrupole of the atom and the oscillating electric field of the incident radiation) are typically about 10^8 times more unlikely than electric dipole transitions.

   The corresponding selection rules are summarized in Figure 2.7.

2.4.1 Two-photon emission

In “normal” conditions the H_2 term in equation (2.55) is negligible. It can be shown that its contribution matches that of H_1 only for very intense electromagnetic fields (that are not found in the intergalactic medium). In terms of photon densities, one needs 10^{24} photons per cm^3 for H_2 being as important as H_1 (this density corresponds to nearly 1 photon per atomic volume). Consider, however, that even at the surface of the Sun the photon density is “only” of $\sim 10^{14}$ cm^{-3}.

There are, however, situations in which the contribution of H_1 vanishes because of particular symmetries of the states involved in the transition. In this cases, H_2 becomes the leading term. Being quadratic in $\mathbf{A}$, H_2 leads to two-photon transitions. These make the decay of the $2S$ state of Hydrogen possible.

2.5 Spontaneous emission

Contrary to our discussion in Section 2.3.1, an atom in an excited state $|i\rangle$ can decay spontaneously (i.e. without the influence of external radiation) to a lower-energy level $|f\rangle$ by emitting a photon of energy $E_i - E_f$ (see Figure 2.6). This is called *spontaneous emission.*

From the theoretical point of view, spontaneous emission becomes manifest only by treating the electromagnetic field as a quantum-mechanical entity. In quantum electrodynamics, the electromagnetic field has a ground state (the vacuum) which couples with the excited levels of the atom. Spontaneous emission results from this interaction. In some sense, it can be considered as a stimulated emission caused by the zero-point energy (i.e. the energy of the vacuum state[7]) of the field.

Computing the rate of spontaneous emission from quantum electrodynamics goes beyond the scope of this course. We will however derive this quantity in a future class using an elegant method due to Einstein.

Stimulated emission always generates radiation in phase with the incident one. On the contrary, spontaneous emission takes place with random phases and directions.

Spontaneous emission takes place at an unpredictable time and thus requires a statistical treatment.

2.5.1 Radiative decay widths

In reality, radiative atomic transitions are not infinitely sharp, rather they have finite width. This can be easily understood as follows. Consider an atomic system in an excited state $|i\rangle$. Because of spontaneous emission, the

[7] For an analogous case, think of the ground state of a quantum harmonic oscillator.

probability to be in the initial state will satisfy the differential equation

$$\left.\frac{d|c_i|^2}{dt}\right|_{\text{spont.em.}} = -\Gamma\,|c_i|^2 \,, \tag{2.66}$$

which is solved by

$$|c_i|^2 = e^{-\Gamma t} \,. \tag{2.67}$$

The lifetime of the level is defined as $\tau = \Gamma^{-1}$ and indicates the time needed to reduce the probablity by a factor e. The lifetime of an excited state is computed by summing up the allowed spontaneous emission rates to all possible lower levels. In the presence of external radiation, one should also add the induced rates to this.

Since the probability of being in a given excited state decays exponentially with time, also the intensity of the emitted radiation will decrease. It is then evident that the spectrum of the emitted radiation cannot be monochromatic.

Repeating the calculations in Section 2.3.3 but allowing also for spontaneous emission in equation (2.41) gives a Fermi rule where the Dirac distribution is replaced by

$$\phi(\omega) = \frac{1}{\pi}\,\frac{\dfrac{\Gamma}{2}}{(\omega-\omega_{fi})^2 + \left(\dfrac{\Gamma}{2}\right)^2} \,, \tag{2.68}$$

(the Fourier transform of a damped oscillator). This is a Lorentzian lineshape [8] and it is fully characterized by the central frequency ω_{fi} and its Full Width at Half Maximum (FWHM) Γ. This lineshape will be observed whenever the density of final states is nearly constant over the width of the line. Note that if the final level $|f\rangle$ is not the ground state of the system it will decay as well further broadening the linewidth. In this case,

$$\Gamma = \Gamma_i + \Gamma_f \,. \tag{2.69}$$

In summary, since the excited states only live for a finite (random) time, there is a spread in the frequency (or energy) of the transition lines. For optical transitions the natural linewidth is typically 10^{-7} eV (with an associated lifetime of 0.66×10^{-8} s). Strong lines, which correspond to fast transitions, are smeared out more that weak lines.

[8] The distribution $[\pi(1+x^2)]^{-1}$ is known as Cauchy distribution is statistics, as Lorentz distribution in spectroscopy, and as Breit-Wigner distribution in nuclear physics.

2.6 Bound-bound transitions

2.6.1 Hydrogen spectrum

Electromagnetic transitions between bound states of the Hydrogen atom will emit or absorb photons with energy

$$E = 13.6 \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right) \text{ eV} \tag{2.70}$$

where n_1 and n_2 ($n_2 > n_1$) are the principal quantum number of the states involved in the transition (we neglect the fine and hyperfine structure here). Some transitions are particularly prominent.

Lyman series. Spectral lines from/to the fundamental level ($n = 1$) form the Lyman series and lie in the ultraviolet region of the electromagnetic spectrum. For instance, the Lyman α line ($n_2 = 2, n_1 = 1$) corresponds to a photon energy of 10.2 eV and to a wavelength of 1215.87 Å. Similarly, the Lyman β line ($n_2 = 3, n_1 = 1$) gives 12.09 eV or 1025.18 Å, and so on, up to the Lyman-series limit (or Lyman edge, $n_2 \to +\infty, n_1 = 1$) with 13.6 eV or 911.267 Å.

Balmer series. Spectral lines from/to the first excited level ($n = 2$) form the Balmer series and lie in the visible/ultraviolet region of the electromagnetic spectrum (Hα at 6563 Å (the strongest optical line) is red, Hβ at 4861 Å is blue/green, Hγ at 4341 Å is violet, the Balmer-series limit at 3646 Å lies in the ultraviolet);

Other series. Spectral lines from/to the excited levels with $n > 2$ form the Paschen ($n = 3$), Brackett ($n = 4$) and Pfund ($n = 5$) series and all lie in the infrared region of the electromagnetic spectrum.

As an example of electric-dipole-forbidden transition, it is important to mention the 21cm spin-flip transition of neutral hydrogen. This magnetic-dipole transition happens at the exceptionally low rate of $\Gamma = 2.9 \times 10^{15}$ s^{-1}. This means that the lifetime of the excited level is $\tau \sim 10^7$ yr!

2.6.2 Ionized-Helium spectrum

Radiative transitions in singly-ionized Helium atoms will create or absorb photons with energy

$$E = 54.4 \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right) \text{ eV} . \tag{2.71}$$

In this case, the Lyman α transition corresponds to an energy of 40.8 eV and to a wavelength of 303.78 Å. On the other hand, the Lyman edge (54.4 eV) lies at nearly 228 Å.

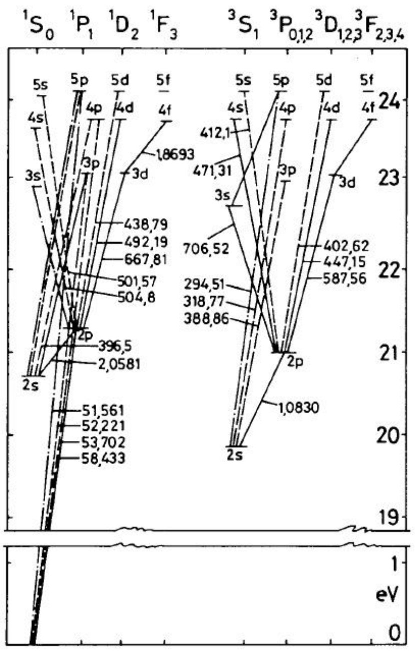


Figure 2.8: Electric dipole transitions for the neutral Helium atom (Parahelium on the left and Orthohelium on the right). Most of the labels indicate wavelengths in nm (with some exceptions, as for the 10830 Å line which is expressed in units of 10^{-7} m).

2.6.3 Neutral Helium spectrum

Since radiative transitions from triplet states to singlet states are forbidden (at the electric dipole level which allows only $\Delta S = 0$), the lowest state of orthohelium is metastable. This implies that, at very low densities, ortho- and parahelium behave as two different elements with different spectra. At higher densities, however, collisions make transitions between the two states possible.

Electric-dipole transitions for ortho- and parahelium are listed in Figure 2.8. For Parahelium, the equivalent of the Lyman-α line lies at 584.33 Å (21.22 eV, ultraviolet), the equivalent of Hα at 20581 Å (near infrared), and the Lyman edge at 504 Å (24.6 eV, ultraviolet). The strongest optical lines are $3\,^1D \rightarrow 2\,^1S$ (5015.7 Å, green) and $3\,^1D \rightarrow 2\,^1P$ (6678.1 Å, red). For Orthohelium, the equivalent of Hα lies at 10830 Å (near infrared). The strongest optical lines are $3\,^3D \rightarrow 2\,^3P$ (5875.7 Å, yellow-orange), $3\,^3S \rightarrow 2\,^3P$ (7065.2 Å, red), and $4\,^3D \rightarrow 2\,^3P$ (4471.5 Å, violet).

Chapter 3

Radiative transfer

Electromagnetic radiation provides most of the information we have on the astrophysical universe. Electromagnetic waves emitted by cosmic sources travel through vast distances before being detected by our instruments. The space they cross is not empty; rather it is filled with an extremely diluted (by terrestrial standards) mixture of ions, atoms, molecules and larger "dust" grains (somewhat similar to soot or sand). Typical densities in the interstellar medium (the diffuse material between stars in a galaxy) range between 10^{-4} and 10^{6} particles per cm^3, with typical values of 1 particle per cm^3. In the intergalactic medium, densities are even lower reaching values of 10 to 1000 particles per cubic meter. To appreciate the real meaning of these values we should remember that the air we breathe has a density of about 10^{19} molecules per cm^3, while the very best laboratory vacuum ever achieved contains about 1000 particles per cm^3. Imagine a volume of space big enough to stretch halfway to the Moon, in intergalactic space such a box would contain about as many atoms as the air in your refrigerator!

In this course we will discuss the physics of the intergalactic medium. As you can imagine, the properties of matter at such a low density are very different from typical Earth conditions (for instance, atomic collisions are much rarer). We will start by discussing the propagation of energy by electromagnetic waves through a dilute medium. This goes under the name of radiative-transfer or radiation-transport problem.

3.1 Radiation fields and the specific intensity

3.1.1 Definitions

In free space or homogeneous media, electromagnetic radiation can be considered to travel in straight lines called rays (at least on distances that greatly exceed the wavelength). If one adopts a particle description of electromagnetic phenomena (where the radiant energy is quantized into fundamental units associated with particles named photons), the rays mark the photon trajectories. We define a *radiation field* as a region of space where rays come from all directions at every point. Radiation of different frequencies ν propagates along each ray.

We can write the radiant energy crossing an infinitesimal area dA within a solid angle $d\Omega$ centred around the direction perpendicular to dA in time dt and in frequency range $d\nu$ as

$$dE = I_\nu \, dA \, dt \, d\Omega \, d\nu \tag{3.1}$$

where I_ν is the *specific intensity* or *brightness.* This is the fundamental quantity we will use to describe radiation fields. At a given time, I_ν is a function of position in space, direction with respect to dA, and frequency. Therefore I_ν depends on 7 variables: 3 spatial coordinates, 2 directional coordinates, frequency, and time. A radiation field is called isotropic if I_ν does not depend on the direction while it is called homogeneous if I_ν does not depend on the spatial location.

The *mean intensity* of radiation, J_ν, is defined by averaging I_ν over all solid angles

$$J_\nu = \frac{1}{4\pi} \int_{4\pi} I_\nu \, d\Omega \tag{3.2}$$

and is a function of 5 variables (position, frequency and time). For an isotropic radiation field $J_\nu = I_\nu$.

The differential energy flux along some arbitrary direction forming an angle θ with the normal to dA is

$$dF_\nu = I_\nu \, \cos\theta \, d\Omega \tag{3.3}$$

and the net flux (per unit frequency) is found by integrating dF_ν over the solid angle. Note that for an isotropic radiation field $F_\nu = 0$ since $\int_{4\pi} \cos\theta \, d\Omega = 0$. This means that equal amounts of energy cross dA per unit time in opposite directions.

The linear momentum of a photon is $p = E/c$ and lies along its direction of propagation (i.e. its ray). Then the momentum flux along a ray at angle θ is dF_ν/c and the specific (i.e. per unit frequency range) radiation pressure (flux of linear momentum in the perpendicular direction to dA) is

$$P_\nu = \frac{1}{c} \int I_\nu \, \cos^2\theta \, d\Omega \; . \tag{3.4}$$

The radiant energy density can also be written in terms of I_ν. Consider an infinitesimal area dA and, on one side of it, build a cylinder of length $c\,dt$ along the perpendicular direction. The radiant energy within the cylinder due to rays in the direction $d\Omega$ is be $dE = u_\nu \, dA \, c \, dt \, d\Omega \, d\nu$. However, since in the time dt all this energy will cross dA, we also have $dE = I_\nu \, dA \, d\Omega \, dt \, d\nu$. Equating these two expressions one derives

$$\frac{dE}{dV \, d\Omega \, d\nu} = \frac{I_\nu}{c} \; , \tag{3.5}$$

and, integrating over angles, one obtains the *specific energy density*

$$u_\nu = \frac{1}{c}\int I_\nu \, d\Omega = \frac{4\pi}{c} J_\nu \; . \tag{3.6}$$

The total intensity, flux, pressure, and energy density are obtained by integrating the corresponding specific quantities over frequency:

$$I = \int I_\nu \, d\nu \; , \tag{3.7}$$

$$F = \int F_\nu \, d\nu \; , \tag{3.8}$$

$$P = \int P_\nu \, d\nu \; , \tag{3.9}$$

$$u = \int u_\nu \, d\nu \; . \tag{3.10}$$

3.1.2 The photon distribution in phase space

Let $f_\epsilon(\mathbf{x}, \mathbf{p}, t)$ be the photon distribution in phase space for particles of elicity ϵ (remember that photons are spin-one particles but, being fully relativistic objects, have only two polarization states: spin parallel or anti-parallel to the direction of propagation). The total phase-space distribution function is obtained summing over the different polarizations:

$$f(\mathbf{x}, \mathbf{p}, t) = \sum_{\epsilon=1}^{2} f_\epsilon(\mathbf{x}, \mathbf{p}, t) \; . \tag{3.11}$$

Then the number density of photons with linear momentum in the infinitesimal element d^3p is

$$dn = f\, d^3p = f\, p^2\, dp\, d\Omega = f\, \left(\frac{h}{c}\right)^3 \nu^2\, d\nu\, d\Omega \; , \tag{3.12}$$

and the associated energy flux through a normal unit surface is

$$h\nu\, dn\, c = f\, \frac{h^4}{c^2}\, \nu^3\, d\nu\, d\Omega \; . \tag{3.13}$$

The latter quantity can also be expressed in terms of the specific intensity as $I_\nu\, d\nu\, d\Omega$, implying that:

$$I_\nu = \frac{h^4\, \nu^3}{c^2} f \; . \tag{3.14}$$

It is often convenient to define the occupation number for each photon spin state as $\mathcal{N}_\epsilon = h^3 f_\epsilon$ so that

$$I_\nu = \frac{h\, \nu^3}{c^2} \sum_{\epsilon=1}^{2} \mathcal{N}_\epsilon \; . \tag{3.15}$$

The occupation number is a dimensionless quantity expressing the mean occupation of a phase-space cell of volume h^3 (the fundamental unit of action in quantum mechanics).

3.2 The equation for radiative transfer

3.2.1 Radiative transfer in vacuum

Consider a light ray in vacuum and select two random points along it separated by the distance r. Now consider the infinitesimal surfaces dA_1 and dA_2 orthogonal to the ray in these points. The energy of the rays that cross both surfaces in the frequency range $d\nu$ and time dt can be written as $dE_1 = I_\nu(1)\, dA_1\, dt\, d\Omega_{12}\, d\nu$ and $dE_2 = I_\nu(2)\, dA_2\, dt\, d\Omega_{21}\, d\nu$ where $d\Omega_{ij}$ indicates the solid angle subtended by dA_j at point i. Energy conservation requires $dE_1 = dE_2$ and geometry gives $d\Omega_{ij} = dA_j/r^2$. Therefore we obtain that $I_\nu(1) = I_\nu(2)$ or, equivalently, that the specific intensity is constant along a ray. This is the reason why we choosed to describe radiation fields in terms of this quantity. We can express the constancy of I_ν in differential form by writing:

$$\frac{dI_\nu}{ds} = 0\ , \tag{3.16}$$

where ds is the differential element of length along the ray. This is the radiative-transfer equation in vacuum.

This equation can be derived also starting from the photon phase-space density. Liouville theorem states that for Hamiltonian systems, the phase-space density is constant along trajectories of the system (i.e. phase space is incompressible). This implies that, along a ray, $df/dt = 0$. Hence, using equation (3.14) and the fact that $ds = c\, dt$,

$$\frac{d}{ds}\frac{I_\nu}{\nu^3} = 0 \quad \rightarrow \quad \frac{I_\nu}{\nu^3} = \mathrm{const} \tag{3.17}$$

which at any given frequency is equivalent to equation (3.16).

3.2.2 Radiative transfer through matter

In our summary of quantum mechanics, we have shown that the presence of matter leads to emission and absorption of radiation. We have identified three processes: spontaneous emission, absorption, and stimulated emission. Therefore the specific intensity of radiation which propagates through matter cannot be expected to remain constant.

We define the monochromatic (spontaneous) *emission coefficient* j_ν such that the energy emitted by matter per unit volume, per unit solid angle, per unit time and per unit frequency is:

$$dE = j_\nu\, dV\, d\Omega\, dt\, d\nu\ . \tag{3.18}$$

In travelling a distance ds, a beam of rays of cross section dA travels through a volume $dV = dA\,ds$. Thus the intensity added to the beam by spontaneous emission is:

$$dI_\nu = j_\nu \, ds \; . \tag{3.19}$$

Similarly, we define the *absorption coefficient* α_ν by considering the loss of intensity in a beam as it travels a distance ds. Since the quantum-mechanical probability of absorption was found proportional to the intensity of the incident radiation, we write:

$$dI_\nu = -\alpha_\nu \, I_\nu \, ds \; . \tag{3.20}$$

However, also the process of stimulated emission is proportional to the intensity of the incoming beam. It is thus customary to include its effects in α. In consequence, the absorption coefficient may be positive or negative, depending on whether absorption or stimulated emission dominates. Note that α_ν has the dimensions of (length)$^{-1}$.

Collecting all the pieces together we can write the fundamental equation for radiative transfer (in the absence of scattering processes that will be discussed separately)

$$\frac{dI_\nu}{ds} = -\alpha_\nu \, I_\nu + j_\nu \; . \tag{3.21}$$

It can be easily solved in two ideal limiting cases.

Medium that only emits radiation. If $\alpha_\nu = 0$, the equation becomes

$$\frac{dI_\nu}{ds} = j_\nu \; , \tag{3.22}$$

and the solution is

$$I_\nu(s) = I_\nu(s_0) + \int_{s_0}^{s} j_\nu(x) \, dx \; . \tag{3.23}$$

In words, the integral of the emission coefficient along the line of sight is added to the brightness.

Medium that only absorbs radiation. If $j_\nu = 0$, we have

$$\frac{dI_\nu}{ds} = -\alpha_\nu \, I_\nu \; , \tag{3.24}$$

and the solution is

$$I_\nu(s) = I_\nu(s_0) \, \exp\left[- \int_{s_0}^{s} \alpha_\nu(x) \, dx \right] \; . \tag{3.25}$$

In words, the brightness is attenuated by a factor that coincides with the exponential of the integral of the absorption coefficient along the line of sight.

3.2.3 Optical depth

The transfer eqution takes a particularly simple form if, instead of the path length along the line of sight s, we use another variable defined by

$$d\tau_\nu = \alpha_\nu \, ds \qquad \text{or} \qquad \tau_\nu(s) = \int_{s_0}^{s} \alpha_\nu(x) \, dx \; . \tag{3.26}$$

This is called the *optical depth* and it is measured along the path of a ray. Its zero point, s_0, is arbitrary.

Note that equation (3.25) can be written as $I_\nu(s) = I_\nu(s_0) \exp(-\tau_\nu)$. If you think of the radiation as composed by a stream of photons (so that I_ν is simply proportional to the number of quanta of frequency ν) then this equation says that the probability of a photon travelling an optical depth τ_ν without being absorbed is $\exp(-\tau_\nu)$.

A medium is called *optically thick* or *opaque* when τ_ν across the medium changes by $\Delta\tau_\nu > 1$. This means that, on average, a photon of frequency ν cannot traverse the entire medium without being absorbed. On the other hand, a medium with $\Delta\tau_\nu < 1$ is said to be *optically thin* or *transparent.* In this case, photons of frequency ν are generally able to cross the entire medium without being absorbed.

3.2.4 Mean free path

It is useful to introduce the concept of *mean free path* of a photon as the average distance a photon can travel through a material without being absorbed. Remember that the probability of a photon to be absorbed at optical depth τ_ν is $\exp(-\tau_\nu)$, so that the mean optical depth travelled by a photon is

$$\langle \tau_\nu \rangle = \int_0^\infty \tau_\nu \, \exp(-\tau_\nu) \, d\tau_\nu = 1 \; . \tag{3.27}$$

The mean free path in a homogeneous medium is thus

$$\lambda = \frac{1}{\alpha_\nu} \; . \tag{3.28}$$

3.2.5 Formal solution of the radiative-transfer equation

After dividing by α_ν, the radiative-transfer equation can be rewritten as

$$\frac{dI_\nu}{d\tau_\nu} = -I_\nu + \mathcal{S}_\nu \; , \tag{3.29}$$

where the *source function* $\mathcal{S}_\nu$ is defined as the ratio of the emission and the absorption coefficients:

$$\mathcal{S}_\nu = \frac{j_\nu}{\alpha_\nu} \; . \tag{3.30}$$

Equation (3.29) is the most used form of the radiative-transfer equation. This is for two reasons:

1. The optical depth maps more clearly the important intervals along a ray where changes in the radiation field happen with respect to the physical path length.

2. In many cases, the source function has a simpler form than the emission coefficient.

We can now give a formal solution of the radiative-transfer problem. Multiply equation (3.29) by $\exp(\tau_\nu)$ and define the auxiliary quantities $H = I_\nu \exp(\tau_\nu)$ and $K = S_\nu \exp(\tau_\nu)$. In this case the RT equation becomes

$$\frac{dH}{d\tau_\nu} = K \tag{3.31}$$

with the solution

$$H(\tau_\nu) = H(0) + \int_0^{\tau_\nu} K(x)\, dx \ . \tag{3.32}$$

Expressing the solution in terms of I_ν and S_ν we finally have:

$$I_\nu(\tau_\nu) = I_\nu(0)\exp(-\tau_\nu) + \int_0^{\tau_\nu} \exp[-(\tau_\nu - x)]\, S_\nu(x)\, dx \ . \tag{3.33}$$

This is composed of two terms:

1. The first term on the left gives the initial intensity attenuated by absorption;

2. The second term gives the sum of all the source terms along the line of sight, each attenuated by the corresponding absorption.

As an example, consider the case of a medium with a constant source function (i.e. assuming the same value everywhere). Then, the solution (3.33) becomes

$$I_\nu(\tau_\nu) = I_\nu(0)\exp(-\tau_\nu) + S_\nu\,[1 - \exp(-\tau_\nu)] \ . \tag{3.34}$$

As $\tau_\nu \to \infty$, $I_\nu \to S_\nu$ i.e. the radiation field looses memory of its initial intensity and becomes more and more determined by the properties of the surrounding medium. Note that if $I_\nu > S_\nu$ then $dI_\nu/d\tau_\nu < 0$ and I_ν decreases along the ray (net absorption). On the other hand, if $I_\nu < S_\nu$ then $dI_\nu/d\tau_\nu > 0$ and I_ν increases along the ray (net emission). Therefore the source function is the quantity that the specific intensity relaxes to if given sufficient optical depth.

3.2.6 Blackbody and thermal radiation

A blackbody is an ideal construction referring to an object that absorbs all the electromagnetic radiation impinging onto it. Radiation in thermal equilibrium with a blackbody enclosure at temperature T has special properties.

As you certainly have learned during your previous studies, simple arguments of thermodynamics can be used to show that the specific intensity of radiation in a blackbody cavity is a universal function of temperature and frequency, $I_\nu = B_\nu(T)$. Based on similar reasonings, one can also show that for any material in thermal equilibrium $\mathcal{S}_\nu = B_\nu(T)$ (Kirchoff's law).

The correct analytic form of the blackbody spectrum has been derived by Planck in 1900 and is now known as the Planck function:

$$B_\nu(T) = \frac{2h\nu^3/c^2}{\exp(h\nu/k_\mathrm{B}T) - 1} \; . \tag{3.35}$$

His model required that the energy of radiation in the cavity had to be quantised in small packets each of energy $h\nu$. Einstein built upon this idea and proposed the quantisation of electromagnetic radiation itself in 1905 to explain the photoelectric effect. All this triggered the development of quantum electrodynamics.

In modern language, the Planck function corresponds to the specific intensity of a photon gas where the occupation number of each spin state at energy E follows [cf. equation (3.15)]

$$\mathcal{N}_\epsilon = \frac{1}{\exp(E/k_\mathrm{B}T) - 1} \tag{3.36}$$

i.e the Bose-Einstein distribution with vanishing chemical potential.

Among the properties of $B_\nu(T)$, it is worth remembering that

1. At low frequencies, such that $h\nu \ll k_\mathrm{B}T$,

   $$B_\nu(T) \simeq \frac{2\nu^2}{c^2}\, k_\mathrm{B}T \; , \tag{3.37}$$

   which is known as the Rayleigh-Jeans regime. Note that the Planck constant does not appear in the asymptotic form of $B_\nu(T)$, indicating that quantum physics is not relevant when the number density of photons with a given energy is high. Equation (3.37) coincides with what one would obtain assuming energy equipartition for the classical electromagnetic modes inside the cavity. Extrapolating this behaviour to arbitrarily high frequencies produces an “ultraviolet catastrophe” where the energy density would diverge.

2. At high frequencies such that $h\nu \gg k_\mathrm{B}T$,

   $$B_\nu(T) \simeq \frac{2h\nu^3}{c^2}\, \exp\left[-\left(\frac{h\nu}{k_\mathrm{B}T}\right)\right] \; . \tag{3.38}$$

   This is the Wien limit and describes the quantum regime where the number density of photons with a given energy is low.

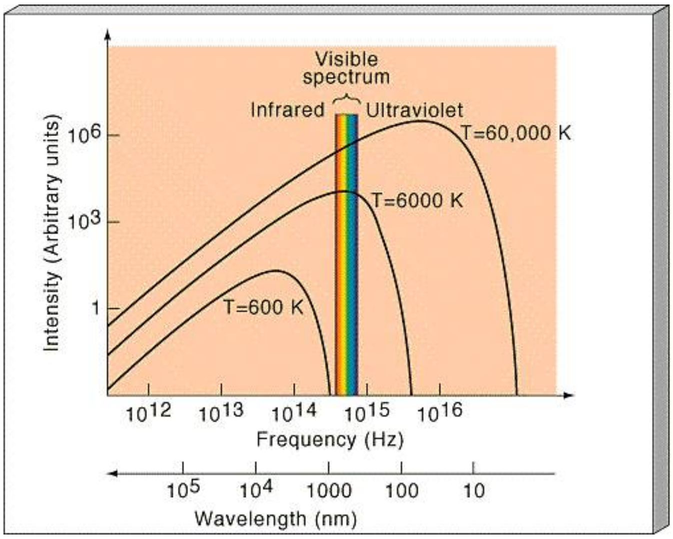


Figure 3.1: The specific intensity of blackbody radiation at different temperatures.

3. Two blackbody curves at different temperatures never cross. The spectrum at higher temperature lies entirely above the other one (see Figure 3.1).

4. The frequency $\nu_{\rm max}$ at which the peak of $B_\nu(T)$ occurs is related to T by $h\nu_{\rm max} = 2.82\, k_{\rm B}T$ (which gives $\nu_{\rm max} = 5.879 \times 10^{10}$ Hz/K T). This is the Wien displacement law.

5. The total energy radiated per unit area and per unit time by the surface of a blackbody is σT^4 with $\sigma = 5.67 \times 10^{-5}$ erg cm^{-2} deg^{-4} s^{-1} the Stefan-Boltzmann constant.

6. The energy density of radiation within a blackbody enclosure is $u(T) = a\,T^4$ with $a = 4\sigma/c = 7.56 \times 10^{-15}$ erg cm^{-3} deg^{-4} (Stefan-Boltzmann law).

It is important to stress that blackbody radiation has $I_\nu = B_\nu$ while thermal radiation is characterized by $\mathcal{S}_\nu = B_\nu$. Only in optically thick media thermal radiation becomes blackbody radiation (see equation 3.34).

Note that the source function for thermal radiation not only has a particular frequency distribution but also a fixed amplitude for every temperature. For this reason, it is useful to introduce the concept of a graybody as an object emitting with a source function $\mathcal{S}_\nu = Q\, B_\nu(T)$ with Q a real number such that $0 < Q < 1$. Hence, the source function of a graybody has the same frequency distribution of blackbody radiation but a different amplitude.

3.3 Connecting macro and microscopic models

The formal solution of the radiative-transfer equation is a powerful tool but cannot be used in practice until we know how to compute the absorption and emission coefficient for a given material. To do this one has to link the macroscopic description of radiation transport given by the radiative-transfer equation with the microscopic modelling of matter-radiation interactions in terms of quantum probabilities for absorption and emission.

3.3.1 Cross-sections

It is often convenient to picture absorption processes as follows. Consider a medium composed by some randomly distributed "particles" (for instance hydrogen atoms) with number density n (number per unit volume). Imagine that each particle is able to completely absorb radiation of frequency ν within a sphere of cross section σ_ν. Now consider a beam of radiation of area dA and length ds passing through the medium. The total (transverse) absorbing area presented by the absorbers is $d\Sigma = \sigma_\nu \, dN = \sigma_\nu \, n \, dA \, ds$ with dN the number of particles in the beam.[1] Therefore, a fraction $d\Sigma/dA$ of the incident energy will be absorbed out of the beam or, equivalently,

$$dI_\nu = -n \, \sigma_\nu \, I_\nu \, ds \, . \tag{3.39}$$

Using the definition of the absorption coefficient - Eq. (3.20) - we obtain

$$\alpha_\nu = n \, \sigma_\nu \, , \tag{3.40}$$

and the photon mean free path is

$$\lambda = \frac{1}{n \, \sigma_\nu} \, . \tag{3.41}$$

In reality atoms and molecules are not opaque spheres. We can anyway use this microscopic model for absorption after replacing the geometric cross section of the spheres with a (frequency dependent) effective cross section. As we will see later, this quantity can be computed using the quantum probabilities for absorption discussed in the previous Chapter.

3.3.2 Einstein coefficients

In 1917 Einstein provided a beautifully simple interpretation of the Planck spectrum for blackbody radiation based on matter-radiation interactions. In

[1] We are assuming here that the radius of the absorbing spheres, $\sim \sigma_\nu^{1/2}$, is much smaller than the mean interparticle distance, $n^{-1/3}$, so that overlapping of cross-sections can be neglected. This corresponds to the condition: $\alpha_\nu n^{-1/3} \ll 1$ which holds in nearly all astrophysical conditions.

particular, he showed that the analytic form of the Planck function implies the existence of spontaneous emission (which was unknown at the time) beyond stimulated emission and absorption. He also derived two equations linking the rates of these three phenomena so that, if one knows one of them, it is easy to derive the remaining two. It would have taken another decade before the quantum mechanics of Heisenberg and Schrödinger for the quantized energy levels of atoms, and the time-dependent perturbation theory of Dirac for radiative transitions would verify by direct computation the relations derived by Einstein. The argument develops as follows.

Consider transitions involving atoms in the bound states $|i\rangle$ and $|j\rangle$ with energies $E_i < E_j$ and statistical weights (i.e. degeneracy factors) g_i and g_j. For a single atom, the *transition probabilities per unit time* are [2]

- $B_{ij}\, J_\nu$ for absorption,
- $B_{ji}\, J_\nu$ for stimulated emission,
- A_{ji} for spontaneous emission,

with $\nu = (E_j - E_i)/h$. The quantities A_{ji}, B_{ji}, and B_{ij} are called Einstein coefficients for bound-bound phototransitions.[3] For each pair of energy levels, the Einstein coefficients are 3 real numbers quantifying transition-rate probabilities. They can be either measured experimentally or computed using quantum mechanics.

Let us now consider a system with many atoms of the same chemical element such that n_i and n_j indicate the number density (number per unit volume) of atoms in level $|i\rangle$ and $|j\rangle$, respectively. In this case, the total number of transitions per unit time is

- $n_i\, B_{ij}\, J_\nu$ for absorption,
- $n_j\, B_{ji}\, J_\nu$ for stimulated emission,
- $n_j\, A_{ji}$ for spontaneous emission.

3.3.3 Detailed balance

This atomic gas is now put inside a cavity with black walls, and the system is kept at temperature T until thermodynamic equilibrium between matter

[2] In this section we are implicitly assuming (as Einstein did in his original derivation) that atomic transitions are infinitely narrow in frequency. As we have already discussed, real transitions always have a line shape $\phi(\nu)$ (such that $\int \phi(\nu)\, d\nu = 1$) with a finite width. One should then write the transition probabilities as $B_{ij}\, \bar{J}_\nu$ with $\bar{J}_\nu = \int_0^\infty J_\nu\, \phi(\nu)\, d\nu$. This would slightly complicate the derivation but the final results would be unchanged with respect to the simpler version given here.

[3] Sometimes the Einstein coefficients are defined using the radiant energy density u_ν instead of the mean intensity of radiation J_ν. This leads to definitions differing by a factor $c/4\pi$.

and radiation is reached. We assume that the gas is so rarefied that the atoms only interact with the radiation filling the cavity and not directly with one another. In thermodynamic equilibrium,

1. upward and downward transitions should be equally frequent;
2. the level populations n_i and n_j should follow the Boltzmann distribution;
3. the specific intensity of radiation coincides with the blackbody solution.

Condition 1) requires that

$$n_i\,B_{ij}\,J_\nu = n_j\,B_{ji}\,J_\nu + n_j\,A_{ji}\ , \tag{3.42}$$

i.e. the mean intensity of radiation

$$J_\nu = \frac{A_{ji}/B_{ji}}{(n_i/n_j)(B_{ij}/B_{ji}) - 1}\ . \tag{3.43}$$

Condition 2) means that

$$\frac{n_i}{n_j} = \frac{g_i}{g_j}\exp\left(\frac{h\nu}{k_\mathrm{B}T}\right)\ , \tag{3.44}$$

so that, together with 1), gives

$$J_\nu = \frac{A_{ji}/B_{ji}}{(g_iB_{ij}/g_jB_{ji})\exp(h\nu/k_\mathrm{B}T) - 1}\ . \tag{3.45}$$

Finally, the fact that $J_\nu = B_\nu$ implies that, for all temperatures,

$$\begin{aligned} g_iB_{ij} &= g_jB_{ji} && (3.46)\\ A_{ji} &= \frac{2h\nu^3}{c^2}\,B_{ji}\ . && (3.47)\end{aligned}$$

If we determine any one of the Einstein coefficients, these relations allow us to derive the other two. These are examples of *detailed balance equations* connecting the rates of a microscopical process and its inverse. Although we have assumed thermodynamic equilibrium to derive them, these equations connect intrinsic atomic properties and must have general validity.

3.3.4 Absorption and emission coefficients

We want to express the absorption and emission coefficients in terms of the Einstein ones.

Let us start from the emission coefficient. Each atom which undergoes the transition $j \to i$ emits the energy $h\nu$ in a random direction. Therefore, the amount of energy emitted per unit volume, solid angle, frequency, and time is

$$j_\nu = \frac{h\nu}{4\pi} n_j \, A_{ji} \, \phi(\nu) \ . \tag{3.48}$$

Similarly, the probability per unit time for absorption is

$$\alpha_\nu^{(\mathrm{unc})} = \frac{h\nu}{4\pi} n_i \, B_{ij} \, \phi(\nu) \ . \tag{3.49}$$

this gives the *absorption coefficient uncorrected for stimulated emission* and is a strictly positive coefficient. However, the full absorption coefficient also includes the "negative absorption" due to stimulated emission. By reasoning entirely analogous to that above, one finds

$$\alpha_\nu = \frac{h\nu}{4\pi} \phi(\nu) \, (n_i B_{ij} - n_j B_{ji}) \ . \tag{3.50}$$

Finally, using the detailed balance relations derived by Einstein, this gives

$$\alpha_\nu = \frac{h\nu}{4\pi} \phi(\nu) \, g_i B_{ij} \left(\frac{n_i}{g_i} - \frac{n_j}{g_j} \right) \tag{3.51}$$

and

$$S_\nu = \frac{2h\nu^3}{c^2} \frac{g_i n_j}{g_j n_i - g_i n_j} \ . \tag{3.52}$$

In brief, the emission and absorption coefficients are determined by:

- the quantum probabilities of a given transition as quantified by the Einstein coefficients;
- the ratio between the statistical populations of the energy levels involved in the transition.

These populations are determined by both collisional and radiative transitions between the levels $|i\rangle$ and $|j\rangle$ but also by radiative transitions to/from other levels. Collisions tend to drive the population ratio to the equilibrium value at the kinetic temperature of the gas. Radiative transitions, instead, relate the population ratio with the frequency spectrum of the radiation field. In order to compute n_i for a given physical system one has to consider all possible transitions to and from level $|i\rangle$ and solve a complex network of coupled differential rate equations.

In general, one can distinguish three cases.

Local thermodynamic equilibrium (LTE). If the matter is in thermal equilibrium with itself (but not necessarily with the radiation) at the kinetic temperature T then equation (3.44) holds and we have

$$\alpha_\nu = \frac{h\nu}{4\pi} \phi(\nu) \, n_i B_{ij} \left[1 - \exp\left(-\frac{h\nu}{k_\mathrm{B} T} \right) \right] \ , \tag{3.53}$$

$$\mathcal{S}_\nu = B_\nu(T) \ . \tag{3.54}$$

This represents the situation where collisions are a very efficient way to exchange energy between the atoms thus maintaining thermal equilibrium (i.e. a Maxwellian velocity distribution).

Non thermal emission with normal populations. This is the case when equation (3.44) does not hold (and/or the velocity distribution of the atoms is not Maxwellian) but

$$\frac{n_i}{g_i} > \frac{n_j}{g_j} \tag{3.55}$$

i.e. there are, on average, more atoms in each of the lower energy levels and the absorption coefficient is positive.

Inverted populations: lasers and masers. When it happens that the number of atoms in the upper state is such that

$$\frac{n_i}{g_i} < \frac{n_j}{g_j} \tag{3.56}$$

one speaks of inverted populations. In this case, the absorption coefficient is negative as can be seen from equation (3.51). Therefore the intensity of radiation increases along a ray due to stimulated emission. This is the phenomenon at the base of laser (Light Amplification by Stimulated Emission of Radiation) and maser (Microwave Amplification by Stimulated Emission of Radiation) phenomena.

3.3.5 Natural linewidth revisited

Following directly from its definition, the natural linewidth for a transition between the levels $i \to f$ can be expressed in terms of the Einstein coefficient for spontaneous emission as

$$\Gamma_{if} = \sum_{n<f} A_{fn} + \sum_{n<i} A_{in} \ . \tag{3.57}$$

The demonstration is left as an exercise.

3.3.6 Putting all together

We have described the "strength" of an atomic transition in three different ways. For instance, speaking of absorption,

1. In Chapter 2, we used the transition rate $R_{i\to j}$ which is proportional (through the Fermi golden rule) to the perturbation matrix element $|\langle i|\hat{H}_1|j\rangle|^2$.

2. In Section 3.3.1, we used the cross section σ_{ij}.

3. In Section 3.3.2, we used the Einstein coefficient B_{ij}.

All these coefficients quantify the rate of the same physical phenomenon and must therefore be related. In this Section we present the equations linking the different coefficients. [4]

There are two main differences between the models we adopted in Chapter 2 and in this Chapter:

1. In Chapter 2, we obtained $R_{i\to j}$ by considering a single, linearly polarized, plane electromagnetic wave. On the other hand, in Chapter 3, σ_{ij} and B_{ij} have been defined in the case of broadband, unpolarized, and isotropic radiation.

2. In Chapter 2, the intensity of electromagnetic radiation was described using the amplitude of the vector potential $\mathbf{A}$, while the mean specific intensity J_ν was adopted in Chapter 3.

Therefore, the transition rate deriving from the Fermi golden rule has to be revised if we want to compare it with the calculations presented in this Chapter. This can be done as follows.

Let us start by finding out the relation between $\mathbf{A}$ and J_ν. In classical terms, the energy density of the radiation field is

$$u_{\rm em} = \frac{1}{8\pi}(|\mathbf{E}|^2 + |\mathbf{B}|^2) \ . \tag{3.58}$$

In the Coulomb gauge (after a little algebra), it can be shown that the energy density of radiation with propagation vector $\mathbf{k}$ and polarization state ϵ corresponds to

$$u_{\rm em} = \frac{1}{2\pi}\, k^2 |A_\epsilon(\mathbf{k})|^2 \ , \tag{3.59}$$

where $A_\epsilon(\mathbf{k})$ is the amplitude of the corresponding vector potential. The total energy density of radiation is obtained by summing up the contributions of the different wavevectors $\mathbf{k}$ and polarization states (2 values of ϵ for each $\mathbf{k}$):

$$u_{\rm em} = \frac{1}{2\pi} \int d^3k \sum_\epsilon k^2 |A_\epsilon(\mathbf{k})|^2 \ . \tag{3.60}$$

On the other hand, in quantum terms, the radiant energy density is represented by the sum over the contributions of all photons

$$u_{\rm em} = \frac{1}{V} \int d^3k \sum_\epsilon h\nu\, \mathcal{N}_\epsilon(\mathbf{k}) \ , \tag{3.61}$$

[4] The following calculations are not easy and it is not necessary to remember them for successfully attending the class. If you are curious, however, you might want to glimpse through them and gain some deeper understanding.

where V is the volume of the system. Comparing the last two expression for the radiant energy density, we may identify $|A_\epsilon(\mathbf{k})|$ as

$$|A_\epsilon(\mathbf{k})| = \left[\frac{2\pi h\nu \mathcal{N}_\epsilon(\mathbf{k})}{Vk^2}\right]^{1/2} = c\left[\frac{h\mathcal{N}_\epsilon(\mathbf{k})}{V\,\omega}\right]^{1/2} \tag{3.62}$$

(remember that $\omega = 2\pi\nu$ and $k = \omega/c$). This equation allows us to eliminate the amplitude of the vector potential from the expression of the transition rate obtained from the semi-classical perturbation theory presented in Chapter 2.

Now we want to derive the transition rate for photoabsorption in the presence of radiation with different polarizations and wavevectors. This is obtained by summing up the contributions of each polarized plane wave. Integration over the photon angular frequency ω leads to the following version of the Fermi golden rule:

$$R_{i\to f} = \sum_{\epsilon=1}^{2}\int w_\epsilon\, d\Omega\ , \tag{3.63}$$

with

$$w_\epsilon = \frac{2\pi}{\hbar}\,|\langle f|\hat{H}_1(\mathbf{k})|i\rangle|^2\,\rho_{\epsilon,\omega}(\mathbf{k}) \tag{3.64}$$

the probability rate per unit solid angle of making the transition with photons with polarization state ϵ, angular frequency ω, and propagation direction $\mathbf{k}$ (indicated by the variable Ω). Note that, for each direction of propagation, the number of photon states within an infinitesimal energy interval centred at $\hbar\omega = E_f - E_i$ is (see also equation 3.12)

$$\rho_{\epsilon,\omega}(\mathbf{k})\,d(\hbar\omega) = \frac{V}{h^3}\,\frac{d^3p}{dE\,d\Omega}\,dE = \frac{V}{(2\pi)^3}\,\frac{\omega^2}{c^3}\,d\omega\ . \tag{3.65}$$

For electric dipole transitions and in the presence of an isotropic and unpolarized radiation field ($\mathcal{N}_1 = \mathcal{N}_2 = \mathcal{N}_{\rm tot}/2$), one can perform the angular integral in equation (3.63) and obtain

$$R_{i\to f} = \frac{4\pi e^2\omega_{fi}^3}{3hc^3}\,\mathcal{N}_{\rm tot}(\omega_{fi})\,|\langle f|\hat{\mathbf{r}}|i\rangle|^2\ . \tag{3.66}$$

Eventually one can use equation (3.15) to express the transition rate in terms of the mean specific intensity of radiation

$$R_{i\to f} = \frac{8\pi^2}{3}\,\frac{e^2}{\hbar c}\,\frac{1}{\hbar}\,J_{\nu_{fi}}\,|\langle f|\hat{\mathbf{r}}|i\rangle|^2\ . \tag{3.67}$$

Note that the second fraction on the right-hand side is the fine-structure constant. The transition rate is therefore proportional to the intensity of radiation and to the square modulus of the matrix element $\langle f|\hat{\mathbf{r}}|i\rangle$ measuring the distance of the "jumping" electron from the nucleus. Allowing for a finite line width, one finally finds:

$$R_{i\to f} = \frac{8\pi^2}{3}\,\frac{e^2}{\hbar c}\,\frac{1}{\hbar}\,|\langle f|\hat{\mathbf{r}}|i\rangle|^2 \int_0^\infty J_\nu\,\phi(\nu)\,d\nu\;. \tag{3.68}$$

We are now ready to connect transition rates, cross sections and Einstein coefficients.

Quantum transition rate and cross section. We can think that the transition rate $|i\rangle \to |f\rangle$ arises from a flux of photons Φ_ν as they encounter an atom of radiative cross section σ_ν. In this case:

$$R_{i\to f} = \int_0^\infty d\nu\,\sigma_\nu\,\Phi_\nu = \int_0^\infty d\nu\,\frac{J_\nu}{h\nu}\,\sigma_\nu \int_{4\pi} d\Omega\;, \tag{3.69}$$

where the factor $h\nu$ at the denominator is needed to pass from the photon energy to the photon counts. Finally, matching equations (3.68) and (3.69), we find for electric dipole transitions:

$$\sigma_\nu = \frac{4\pi^2}{3}\,\frac{e^2}{\hbar c}\,|\langle f|\hat{\mathbf{r}}|i\rangle|^2\,\nu\,\phi(\nu)\;. \tag{3.70}$$

Note that σ_ν has the dimensions of a surface.

Cross section and Einstein coefficient. Matching equations (3.40) and (3.49), one directly obtains:

$$\sigma_\nu = \frac{h\nu}{4\pi}\,\phi(\nu)\,B_{ij}\;. \tag{3.71}$$

Quantum transition rate and Einstein coefficient Combining equations (3.70) and (3.71), we have:

$$B_{ij} = \frac{8\pi^2}{3}\,\frac{e^2}{\hbar c}\,\frac{1}{\hbar}\,|\langle f|\hat{\mathbf{r}}|i\rangle|^2\;, \tag{3.72}$$

which, inserted in equation (3.68), gives:

$$R_{i\to f} = B_{ij} \int_0^\infty J_\nu\,\phi(\nu)\,d\nu\;. \tag{3.73}$$

3.3.7 Oscillator strength

In order to facilitate the comparison between different lines, it is customary to quantify the strength of atomic transitions using a dimensionless number called the *oscillator strength*. This is defined by comparing the emission or absorption rates of an atomic transition with those of an ideal, classical, single electron, harmonic oscillator.

A classical oscillator of charge e and mass $m_{\rm e}$ invested by a plane electromagnetic wave extracts some energy from the wave. One can then define the absorption cross-section as the ratio of the power absorbed by the oscillator to the incident power per unit area in the electromagnetic field. This comes out to be

$$\sigma_{\rm osc} = \frac{\pi e^2}{m_{\rm e} c}\, \delta_{\rm D}(\nu - \nu_0) \ , \tag{3.74}$$

with ν_0 the natural frequency of the oscillator.

The oscillator strength for a given transition $|i\rangle \to |j\rangle$ is then defined as

$$f_{ij} = \frac{\int \sigma_\nu \, d\nu}{\int \sigma_{\rm osc}\, d\nu} = \frac{2}{3} \frac{m_{\rm e}}{\hbar^2} \left(E_j - E_i\right) |\langle j|\hat{\mathbf{r}}|i\rangle|^2 \ , \tag{3.75}$$

where we have assumed that the linewidth is negligibly small. Note that the oscillator strength is positive for transitions from a lower-energy state to an upper-energy state (i.e. for absorption processes).

The Einstein coefficients can then be written as (for $i < j$)

$$B_{ij} = \frac{4\pi^2 e^2}{m_{\rm e}\, h\nu_{ij}\, c}\, f_{ij} \tag{3.76}$$

$$B_{ji} = \frac{4\pi^2 e^2}{m_{\rm e}\, h\nu_{ij}\, c}\, \frac{g_i}{g_j}\, f_{ij} \tag{3.77}$$

$$A_{ji} = \frac{8\, \nu_{ij}^2\, \pi^2 e^2}{m_{\rm e} c^3}\, \frac{g_i}{g_j}\, f_{ij} \ . \tag{3.78}$$

For transitions between degenerate states, it is conventional to use oscillator strengths obtained averaging over initial substates and summing over final substates. Sometimes it is convenient to define oscillator strengths for emission processes as:

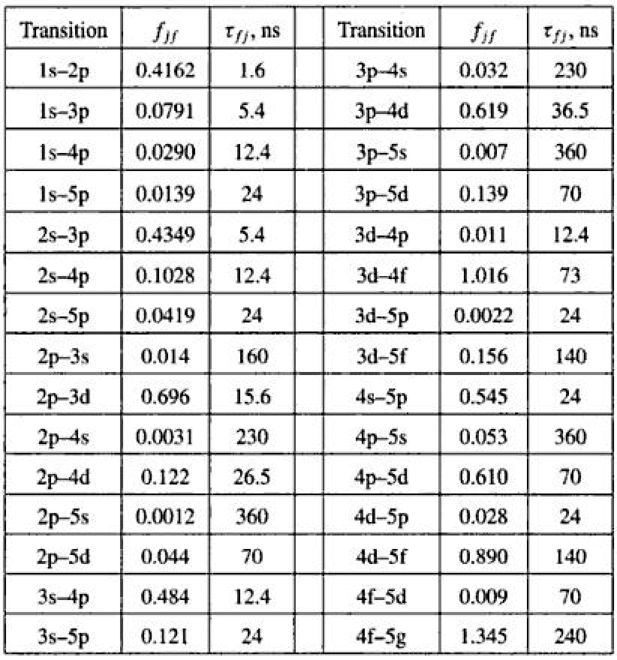
$$\tilde{f}_{ji} = -\frac{g_i}{g_j}\, f_{ij} \ , \tag{3.79}$$

(for $i < j$). In other words, the oscillator strengths have been defined so that, if $g_j = 3$, $g_i = 1$ and the Einstein coefficient A_{ji} is equal to the classical decay amplitude of the harmonic oscillator, then the absorption $f_{12} = 1$ and the emission $\tilde{f}_{21} = -1/3$.

Atomic transitions cannot be arbitrarily strong. For electric dipole transitions involving a single electron, the sum of the oscillator strength from one state to all other states must exactly equal 1:

$$\sum_{j \neq i} f_{ij} = 1 \ . \tag{3.80}$$

This relation is known as the Thomas-Reiche-Kuhn sum rule and can be generalized to transitions involving N electrons where the value 1 is replaced by N. The summation is made over both the discrete and the continuum



Transition	f_{jf}	τ_{fj}, ns		Transition	f_{jf}	τ_{fj}, ns
1s–2p	0.4162	1.6		3p–4s	0.032	230
1s–3p	0.0791	5.4		3p–4d	0.619	36.5
1s–4p	0.0290	12.4		3p–5s	0.007	360
1s–5p	0.0139	24		3p–5d	0.139	70
2s–3p	0.4349	5.4		3d–4p	0.011	12.4
2s–4p	0.1028	12.4		3d–4f	1.016	73
2s–5p	0.0419	24		3d–5p	0.0022	24
2p–3s	0.014	160		3d–5f	0.156	140
2p–3d	0.696	15.6		4s–5p	0.545	24
2p–4s	0.0031	230		4p–5s	0.053	360
2p–4d	0.122	26.5		4p–5d	0.610	70
2p–5s	0.0012	360		4d–5p	0.028	24
2p–5d	0.044	70		4d–5f	0.890	140
3s–4p	0.484	12.4		4f–5d	0.009	70
3s–5p	0.121	24		4f–5g	1.345	240

Figure 3.2: The oscillator strength and the lifetime for radiative transitions of the hydrogen atom.

energy states (this is an integral, of course). For instance, for the fundamental state of hydrogen, a contribution of 0.564 comes from transitions to bound states while the remaining 0.436 comes from transitions to unbound states.

The oscillator strengths of atomic transitions are available in tabulated form. A stronger transition is associated with a higher value of f. In general, for hydrogen-like atoms, oscillator strengths of transitions between states with principal quantum numbers i and j (corresponding to the absorption of radiation) are approximately given by

$$f_{ij} \simeq \frac{32}{3\pi\sqrt{3}} \frac{1}{i^5\, j^3} \frac{1}{(i^{-2} - j^{-2})^3} \tag{3.81}$$

(Menzel & Pekeris 1935; Bethe & Salpeter 1957) which is accurate within a factor of 2. For instance, for the Lyman α transition, the formula above gives 0.5808 while the correct value is 0.4162. Note that, for a fixed initial state, the oscillator strength rapidly decreases with the principal quantum number of the final state, j.

The exact oscillator strengths and lifetimes for the principal levels of the hydrogen atoms are listed in Figure 3.2. Note that f_{ij} depends upon the principal and the orbital quantum numbers of the energy levels. Therefore, caution should be taken in computing the total oscillator strength associated with a transition between degenerate levels. In general,

$$f_{nn'} = \frac{1}{n'^2} \sum_{\ell,\ell'} (2\ell + 1) \; f_{n,\ell \to n',\ell'} \; , \tag{3.82}$$

(remember that oscillator strengths are conventionally computed by averaging over initial substates and summing over final substates). For instance, for the Hα line in absorption,

$$\begin{aligned} f_{23} &= \frac{1}{4} [1 \cdot f_{2s\to 3p} + 3 \cdot (f_{2p\to 3s} + f_{2p\to 3d})] = \\ &= \frac{1}{4} [1 \cdot 0.4349 + 3 \cdot (0.014 + 0.696)] = 0.6407 \; . \end{aligned} \tag{3.83}$$

Chapter 4

Spectral lines

In 1835, Auguste Comte, a prominent French philosopher, stated the humans would never be able to understand the chemical composition of astronomical objects. He was soon proven wrong. In the latter half of the 19$^{\text{th}}$ century, the work of the pioneers of spectroscopy as Fraunhofer, Bunsen, and Kirchoff (and many others like Huggins, Secchi, Pickering) helped bring about a revolution in people's understanding of the cosmos. For the first time, scientists could investigate what the universe was made of. The advent of spectroscopy marked the birth of astrophysics as an observational science.

A hot thermal source emits a continuum spectrum of electromagnetic radiation with a particular frequency distribution. If this radiation crosses a cloud of cooler gas, it will be partially absorbed. In particular, since atoms and molecules have discrete energy levels, the gas will absorb radiation at a set of given frequencies. Therefore, the spectrum of the thermal source seen through the gas cloud will present a series of absorption lines (narrow wavelength regions with reduced intensity). At the same time, since a given atom will absorb and emit the same frequencies of electromagnetic radiation, a spectrum containing emission lines will be detected along lines of sight that intersect the gas cloud but not the background thermal source (see Figure 4.1).

The physics of spectral lines is the subject of this class.

4.1 Line broadening

4.1.1 Thermal broadening

Let us consider the absorption feature produced by a specific atomic transition in a medium in LTE. Atoms are in thermal motion and the frequency of emission or absorption in their own rest frame corresponds to a different frequency for an observer. The change in frequency associated with an atom

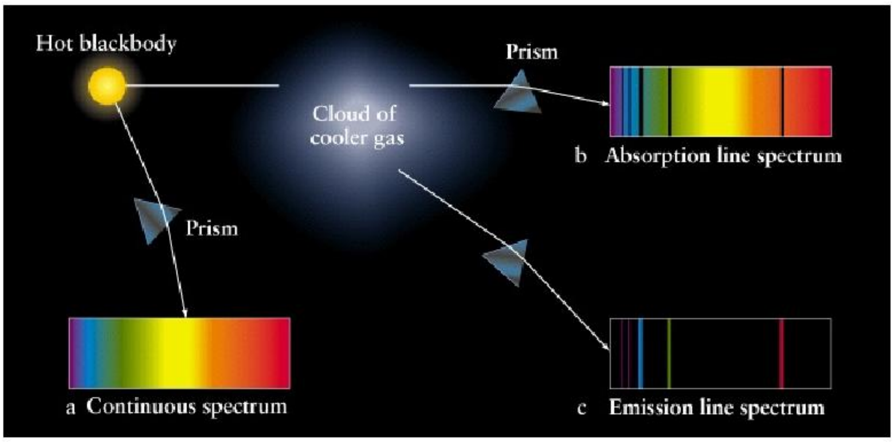


Figure 4.1: Schematic description of the production of absorption and emission lines.

with velocity component $v_{\rm z}$ along the line of sight is, to first order in $v_{\rm z}/c$,

$$\nu(v_{\rm z}) \simeq \nu_0 \left(1 + \frac{v_{\rm z}}{c}\right) \ , \tag{4.1}$$

where ν_0 indicates the rest-frame frequency of the transition. Each atom has its own Doppler shift, so that the net effect is to spread the line out without changing its total strength.

When LTE is established or, more generally, when the velocity distribution of the atoms is Maxwellian, the fraction of atoms having velocities in the range $v_{\rm z}$ to $v_{\rm z} + dv_{\rm z}$ follows a Gaussian distribution:

$$\begin{aligned} P(v_{\rm z})\, dv_{\rm z} &= \left(\frac{m_{\rm a}}{2\pi k_{\rm B} T}\right)^{1/2} \exp\left(-\frac{m_{\rm a} v_{\rm z}^2}{2 k_{\rm B} T}\right) dv_{\rm z} & (4.2) \\ &= \frac{1}{\sqrt{\pi}\, b_{\rm th}} \exp\left(-\frac{v_{\rm z}^2}{b_{\rm th}^2}\right) dv_{\rm z} \ , & (4.3) \end{aligned}$$

where $m_{\rm a}$ is the mass of an atom and

$$b_{\rm th} = \left(\frac{2 k_{\rm B} T}{m_{\rm a}}\right)^{1/2} \tag{4.4}$$

is the Doppler parameter (which has the dimension of a velocity and is normally measured in km s^{-1}).[1]

[1] The amplitude of the three-dimensional velocity, $v = (v_{\rm x}^2 + v_{\rm y}^2 + v_{\rm z}^2)^{1/2}$, follows the Maxwell-Boltzmann distribution $P(v)\, dv = 4\pi\, (1/\pi b_{\rm th})^{3/2}\, v^2\, \exp[-(v/b_{\rm th})^2]\, dv$. It is easy to show that $b_{\rm th}$ corresponds to the most probable speed, the mean velocity is $\langle v \rangle = \sqrt{4/\pi}\, b_{\rm th}$, and the root mean square (rms) speed is $v_{\rm rms} = \langle v^2 \rangle^{1/2} = \sqrt{3/2}\, b_{\rm th}$.

The resulting line profile is then

$$\phi(\nu) = \int_{-\infty}^{+\infty} P(v_{\rm z})\, \phi_{\rm int}[\nu(v_{\rm z})]\, dv_{\rm z} \tag{4.5}$$

where $\phi_{\rm int}(\nu)$ indicates the absorption profile in the rest-frame of the atom. If $\phi_{\rm int} = \delta_{\rm D}(\nu - \nu_0)$ (i.e. the rest-frame line profile is infinitesimally sharp) and $b_{\rm th}/c \ll 1$, then

$$\phi(\nu) = \frac{1}{\sqrt{\pi}\,\Delta\nu} \exp\left[-\left(\frac{\nu-\nu_0}{\Delta\nu}\right)^2\right] \tag{4.6}$$

where

$$\Delta\nu = \frac{b_{\rm th}}{c}\,\nu_0 \tag{4.7}$$

is the Doppler width. A medium in LTE thus produces Gaussian absorption lines with a width which is proportional to the characteristic speed of the atoms at temperature T. This phenomenon is called thermal (or Doppler) broadening of spectral lines.

In addition to thermal motions there can also be turbulent velocities associated with macroscopic velocity fields. When the scale of the turbulence is small in comparison with the photon mean free path (microturbulence), these motions can be accounted for by an effective Doppler parameter

$$b = \left(b_{\rm th}^2 + b_{\rm turb}^2\right)^{1/2} \tag{4.8}$$

where $b_{\rm turb}$ is the rms of the turbulent velocities. This assumes that the turbulent velocities also have a Gaussian distribution (which might or might not be true).

4.1.2 Natural broadening

We have already shown in Section 2.5.1, that radiative bound-bound transitions in atoms are not infinitely sharp. Because of spontaneous emission, the wavefunction of a given state $|n\rangle$ decays over time as $\exp(-\Gamma t/2)$ and the line profile is of the form [see equation (2.58)]

$$\phi(\nu) = \frac{1}{\pi}\,\frac{\dfrac{\Gamma}{4\pi}}{(\nu-\nu_0)^2 + \left(\dfrac{\Gamma}{4\pi}\right)^2}\,. \tag{4.9}$$

This is called a Lorentz (or natural) profile. Remember that in terms of the Einstein coefficient A_{ij} one has:

$$\Gamma_{ij} = \sum_{n<j} A_{jn} + \sum_{n<i} A_{in}\,. \tag{4.10}$$

4.1.3 Collisional broadening

The Lorentz profile applies more generally to certain types of collisional broadening mechanisms. For example, if the atom suffers collisions with other particles while it is emitting, the phase of the emitted radiation can be altered suddenly. If the phase changes completely randomly at collision times, then it can be shown that the emerging line profile is Lorentzian with

$$\Gamma_{\text{eff}} = \Gamma + 2\nu_{\text{coll}} \tag{4.11}$$

where ν_{coll} indicates the mean number of collisions experienced by an atom per unit time.

As we have already mentioned, collisions are hardly important in the intergalactic medium. To substantiate this, we provide an example for the Lyα line of hydrogen. Consider a medium in LTE containing n_{e} electrons per cm^3 at temperature T. The quantum-mechanical cross section of the Lyα transition induced by electron collisions is $\sigma \sim 6 \times 10^{-21}$ m^2 (slightly dependent on the electron energy). Therefore, the collisional Lyα transition rate per hydrogen atom is

$$\nu_{\text{coll}} = n_{\text{e}} \langle v \rangle \sigma = 3.7 \times 10^{-7} \left(\frac{n_{\text{e}}}{100 \text{ cm}^{-3}}\right) \left(\frac{T}{10,000 \ K}\right)^{1/2} \text{s}^{-1} \ , \tag{4.12}$$

where we have used $\langle v \rangle = 621.25\,(T/10,000 \ K)^{1/2}$ km s^{-1} (note that the above rate roughly corresponds to one collisionally induced transition per month). This must be compared with the natural linewidth $\Gamma = 6.265 \times 10^8$ s^{-1} (corresponding to a lifetime $\tau \simeq 1.6$ ns). Therefore, collisional Lyα broadening will be important only when

$$n_{\text{e}} T^{1/2} \simeq 10^{15} \text{cm}^{-3} K^{1/2} \ , \tag{4.13}$$

(or larger) which is commonly found in stellar interiors but not in the intergalactic medium.

4.1.4 The Voigt profile

In general, one has to take simultaneously into account thermal and natural broadening. From equation (4.5) the line profile will then be

$$\phi(\nu) = \frac{\Gamma}{4\pi^{5/2}\, b} \int_{-\infty}^{+\infty} \frac{\exp[-(v_{\text{z}}/b)^2]}{[\nu - \nu_0(1 + v_{\text{z}}/c)]^2 + (\Gamma/4\pi)^2} \, dv_{\text{z}} \ , \tag{4.14}$$

which is known as the Voigt profile. This can be written more compactly introducing the Hjerting function (the convolution of a Gaussian and a Lorentzian distribution which has no closed analytical form)

$$H(a, u) = \frac{a}{\pi} \int_{-\infty}^{+\infty} \frac{\exp(-y^2)}{a^2 + (u - y)^2} \, dy \tag{4.15}$$

(several fast and accurate numerical algorithms for calculating the Hjerting function are available). Then, the line profile can be written as

$$\phi(\nu) = \frac{1}{\sqrt{\pi}\,\Delta\nu}\,H(a,u) \tag{4.16}$$

with

$$\begin{aligned} a &= \frac{\Gamma}{4\pi\Delta\nu} = \frac{\Delta v_{\rm nat}}{b} && (4.17)\\ u &= \frac{\nu-\nu_0}{\Delta_\nu}\,, && (4.18)\end{aligned}$$

where $\Delta v_{\rm nat} = \Gamma c/4\pi\nu_0$.

Generally, the natural linewidth is much smaller than the Doppler one (i.e. $a \ll 1$; for instance $\Delta v_{\rm nat} = 6.06\times10^{-3}$ km s^{-1} for hydrogen Lyα and $a \simeq 2-3\times10^{-4}$ for intergalactic absorption lines). In this case,

- for $|\nu-\nu_0| < 3\,(b/c)\,\nu_0$ (the "core" of the line) the Voigt profile is dominated by the Doppler (Gaussian) profile;

- for $|\nu-\nu_0| > 3\,(b/c)\,\nu_0$ (the "wings" of the line) the Voigt profile is dominated by the natural (Lorentz) profile.

Examples are shown in Figures 4.2 and 4.3.

4.1.5 Spectral resolution

What is recorded by an instrument is the convolution of the line shape with the instrumental response. The *resolving power* of a spectrograph is usually expressed as

$$R = \frac{\lambda}{\Delta\lambda} = \frac{c}{\Delta v} \tag{4.19}$$

where $\Delta\lambda$ is the Full Width at Half Maximum (FWHM) of the instrumental broadening function in wavelength. Roughly speaking, $\Delta\lambda$ is the smallest difference in wavelength that can be distinguished by the spectrograph at a wavelength λ. Similarly Δv gives the minimum velocity difference that can be distinguished.

Lines which are intrinsically narrower than $\Delta\lambda$ will be distorted by the instrument and seen with a FWHM of $\Delta\lambda$. For these *unresolved* lines, one looses most of the information encoded in the line profile. Substantially broader lines, instead, will be *resolved* and will show their intrinsic profile.

Modern spectrographs are classified as:

- Low resolution: $R \sim 100$;

- Medium resolution: $R \sim 1,000$;

- High resolution: $R > 20,000$.

Typical values for b of intergalactic Lyα absorption lines are $b \sim 20 - 30$ km s^{-1}. These correspond to a FWHM of 1.665 $b \sim 35 - 50$ km s^{-1}. To resolve these lines an instrument with $R > 6,000 - 10,000$ is required. This only became possible in the mid 1990s with the advent of echelle spectrographs on 8-10m telescopes, particularly the High Resolution Echelle Spectrometer (HIRES, $25,000 < R < 85,000$ depending on configuration) at the W.M. Keck Observatory (Mauna Kea, Island of Hawai'i) and the Ultraviolet and Visual Echelle Spectrograph (UVES, $40,000 < R < 110,000$) on the Very Large Telescope (VLT, Cerro Paranal, Chile).

4.2 Spectral lines and gas properties

4.2.1 The column density

The optical depth of a gas cloud is proportional to the number of atoms (in the initial energy level of the transition) per unit area around the line of sight. This is so whether or not the cloud is homogeneous. The number of atoms per unit area (i.e. the number density integrated along the line of sight through the absorbing material) is called the *column density*,

$$N = \int n \, ds \, . \tag{4.20}$$

The optical depth is then $\tau_\nu = \sigma_\nu N$.

4.2.2 The equivalent width

In order to compare the strength of different spectral lines it is useful to introduce the concept of equivalent width. Ideally, this is done in three steps (see also Figure 4.2):

1. We first estimate the continuum intensity level in the spectral range covered by the line.

2. We then measure the area A of the spectral line below (for absorption) or above (for emission) the continuum intensity level.

3. Finally, we replace the spectral profile with an artificial one where the continuum light is fully absorbed and there are sharp boundaries jumping back to the unabsorbed level. We make sure that the new profile covers the area A and we define the equivalent width (EW) as the width of this "rectangular" profile.

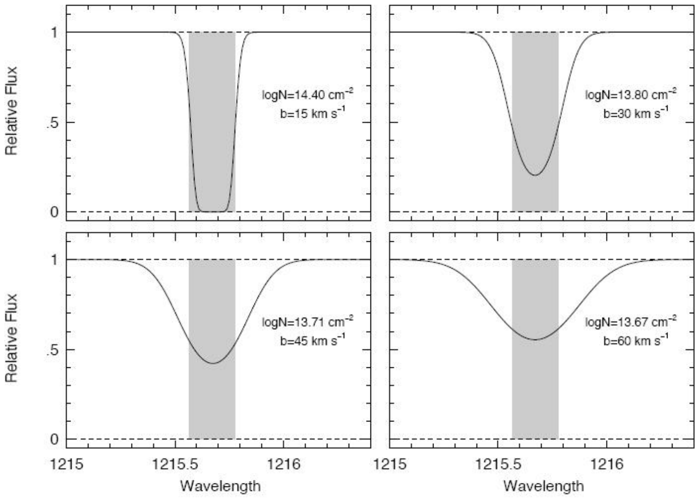


Figure 4.2: The equivalent width (EW) of an absorption line corresponds to the width of the shaded area. Note that the four Lyα lines shown have the same EW but different column densities and Doppler widths.

In formula,

$$W_\lambda = \int \frac{|I_{\rm cont} - I_\lambda|}{I_{\rm cont}}\, d\lambda = \int [1 - \exp(-\tau_\lambda)]\ d\lambda\ , \tag{4.21}$$

where the second equality holds for absorption lines.[2] The equivalent width has the dimension of length (it is generally measured in Å) and is independent of the instrumental resolution of the spectrograph (modulo complications in the case of spectra measured with low signal-to-noise ratio). This property makes it a particularly useful quantity to compare the intensity of spectral lines.

4.2.3 The curve of growth

The relation between the equivalent width of a spectral line and the column density of the absorbing atoms is known as the *curve of growth.* The precise functional dependence of W_λ on N is sensitive to the optical depth at the line core, τ_0 (i.e. to the oscillator strength f of the transition).

We can distinguish three regimes which are obtained integrating equation (4.14).

1. When the column density is low and the line is optically thin, $\tau_0 < 1$, the equivalent width is directly proportional to N irrespectively of the value of the Doppler parameter b.

[2] We are using here the specific intensity of radiation per unit wavelength I_λ instead of the usual I_ν. Since $\nu = c/\lambda$, $d\nu = -c/\lambda^2\, d\lambda$, and $I_\lambda = I_\nu\, |d\nu/d\lambda| = (c/\lambda^2)\, I_\nu$.

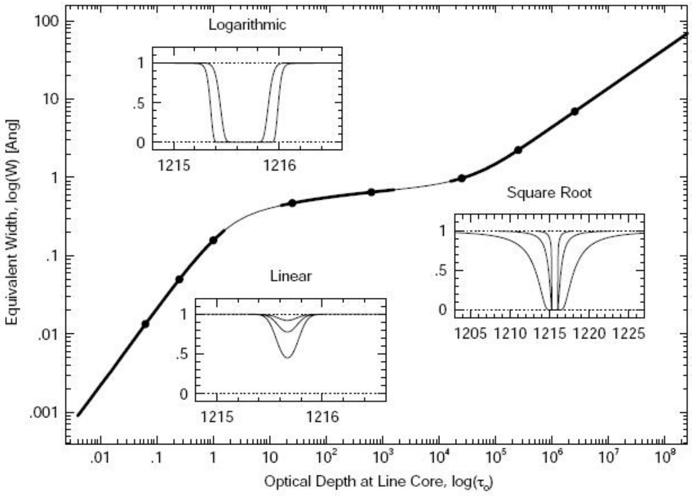


Figure 4.3: The curve of growth (COG) for a hydrogen Lyα line with $b = 30$ km s^{-1}. The line profiles corresponding to the points marked with filled dots on the COG are shown in the small panels. Note that the x-axis scale on the rightmost panel has been expanded to illustrate the large extent of damping wings.

2. For a high enough abundance of atoms, the line becomes optically thick and saturates (i.e. it completely removes all the light at the center of the line). In these conditions, the equivalent width of the line increases only moderately, and only by growing the wings (note that growing the wings means broadening the line). At first, Doppler broadening dominates the increase in the line strength and $W_\lambda \propto b\sqrt{\ln(N/b)}$. In this regime the equivalent width is NOT a good measure of the column density but is sensitive to the Doppler parameter (see Figure 4.4).

3. Eventually, as the density of atoms increases even more, the Lorentzian wings start dominating the growth of the equivalent width. In this regime, $W_\lambda \propto N^{1/2}$ and the equivalent width provides an accurate measure of the column density.

4.3 Retrieving information

What physical information about the intervening medium can we retrieve from absorption lines? The presence of a given chemical element can be determined by detecting its characteristic lines in a spectrum. Similarly, the physical conditions of the absorbing gas cloud (temperature, density, velocity) can be inferred from the line profile of the spectral lines.

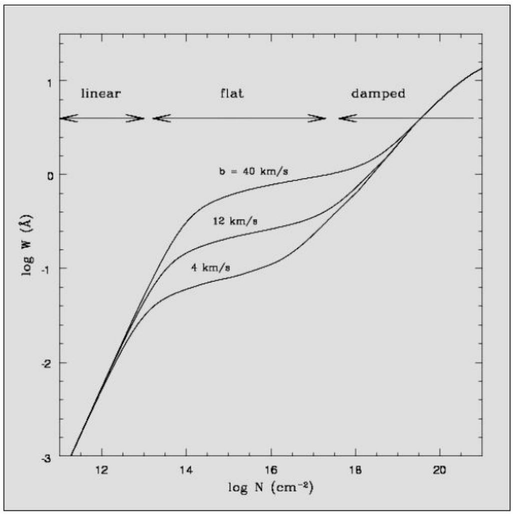


Figure 4.4: The curve of growth for the hydrogen Lyα transition with different b parameters.

4.3.1 Unresolved lines

For a single, unresolved line the relationship between N and W_λ is degenerate as it depends of the exact value of the Doppler parameter. Higher values of b combined with lower values of N produce the same W_λ (see Figure 4.4).

However, if several atomic transitions with different values of $f\lambda$ originating from the same atomic level are available, one can construct an empirical curve of growth and measure both N and b. An example is shown in Figure 4.5.

Many of the most commonly observed absorption lines are doublets (i.e. transitions between the ground state and an excited state consisting of two closely spaced sublevels). The ratio of the equivalent widths of such pairs of lines (the doublet ratio) can be a useful pointer to the region of the curve of growth where the lines fall.

4.3.2 Resolved lines

For resolved lines, we can avoid to use the equivalent width as we can measure the optical depth directly at each velocity (or wavelength). Each velocity bin (defined by the instrumental resolution) then gives a contribution to the column density of

$$\frac{dN}{dv_{\rm z}}(v_{\rm z}) = 3.77 \times 10^{14}\, \frac{\tau(v_{\rm z})}{f\lambda}\, {\rm cm}^{-2}\, ({\rm km\ s}^{-1})^{-1}\ . \tag{4.22}$$

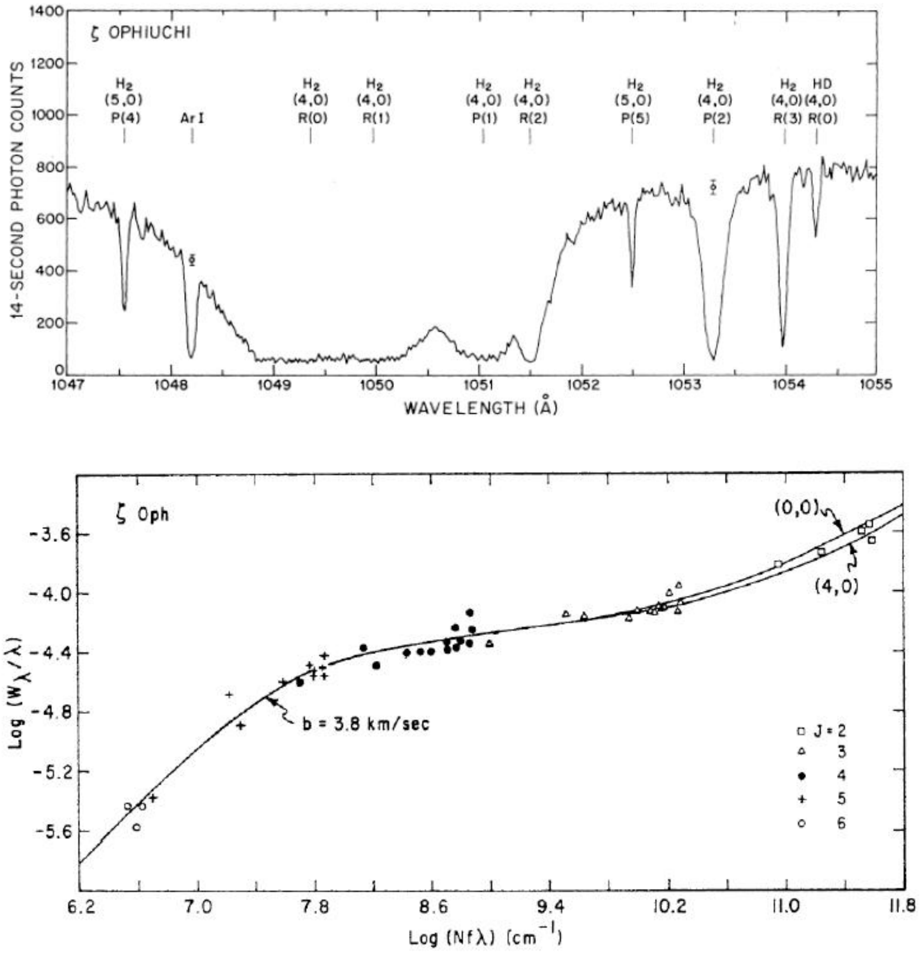


Figure 4.5: Top: The far-ultraviolet spectrum of the star ζ Ophiuchi showing several absorption lines of molecular hydrogen, H_2. Bottom: Empirical curve of growth obtained from the H_2 lines in the spectrum of ζ Ophiuchi. The best-fitting model represents a single absorbing cloud with a Doppler parameter $b = 3.8$ km s^{-1}.

Integration over the absorption line profile then leads to the column density

$$N = \int \frac{dN}{dv_\mathrm{z}}(v_\mathrm{z})\, dv_\mathrm{z} \; . \tag{4.23}$$

This is known as the *optical depth method.*

For the Lyα line

$$\frac{dN}{dv_\mathrm{z}}(v_\mathrm{z}) = 7.45 \times 10^{11}\, \tau(v_\mathrm{z})\, \mathrm{cm}^{-2}\, (\mathrm{km\ s}^{-1})^{-1} \; . \tag{4.24}$$

The Doppler parameter can be derived by fitting the observed spectrum with a linear superposition of Voigt profiles. The Voigt Profile FItting Program (VPFIT by R. F. Carswell, J. K. Webb, A. J. Cooke, M. J. Irwin) is a widespread code that enables us to do this. This is particularly useful to deblend lines that overlap in wavelength (see for example the region around 1049.5 Å in Figure 4.5).

4.3.3 A note on density measurements

There are two possible ways to increase the optical depth of a gas cloud of size L:

- increasing the number density of atoms within L;
- increasing the thickness of the cloud while keeping the number density constant.

The resulting absorption lines will be undistinguishable as long as they correspond to the same column density. This example shows that it is impossible to directly estimate the number density of atoms in a gas cloud just based on the absorption feature they generate. The only information we can extract is the column density. To measure the density one needs an independent estimate of the size of the cloud.

Also note that from a spectral line we can only determine the abundance of an element in the initial state of the transition. The total number of atoms can be determined only by knowing the fraction of atoms in the absorbing state, e.g. by using the Boltzmann distribution when LTE holds.

Chapter 5

The expanding universe

Observations suggest that we live in an expanding Universe and the cosmic expansion is expected to affect in many ways the physics of the intergalactic medium. Therefore, before discussing observational data and models for the IGM, we need to learn how to describe mathematically an expanding Universe. This is the subject of today's class.

5.1 Building a world model for the Universe

5.1.1 The cosmological principle

The observable Universe exhibits a wealth of structures on "small" spatial scales (in cosmic terms) as like as galaxies and clusters of galaxies. However,

- Temperature fluctuations of the cosmic microwave background (CMB) are of order $\Delta T/T \sim 10^{-5}$ on scales of a few degrees;
- The galaxy distribution is rather smooth on scales much larger than 100 Mpc;
- The X-ray background and the distribution of radio galaxies are highly isotropic.

We can summarize these statements as follows: *When smoothed on large scales, the Universe is isotropic as seen from the Earth.* At the same time, it seems reasonable to accept the Copernican principle which states that: *The Earth is not a privileged location in the Universe.* Therefore we are forced to conclude that: *The universe, when viewed on sufficiently large distance scales, has no preferred directions or preferred places.* Or, in other words, on large scales the Universe looks the same in all directions for an observer at any place. This statement is usually referred to as the cosmological principle and in mathematical terms says that the Universe is homogeneous and isotropic on large scales. It is interesting to note that the cosmological principle was first adopted by Arthur Milne in 1933 when observational cosmology was in its infancy. At that time, it was little more than a conjecture, embodying Occam's razor for the simplest possible model. The cosmological principle had hard times in the 1930s, very distinguished scientists (like Eddington) were not ready to accept models based on philosophical speculation. Nowadays the cosmological principle is considered a useful working hypothesis which no observation has contradicted.

5.2 The Friedmann-Robertson-Walker metric

In this section we will assume that space-time forms a Riemannian manifold (a differentiable space in which each tangent space is equipped with a scalar product which varies smoothly across the manifold) and that the local geometry of our Universe can be fully described by the corresponding metric tensor (that we will express in terms of the proper time element $d\tau$). Note that this does not imply that we are assuming General Relativity as the theory of gravitation.

The cosmological principle puts very tight constraints on the geometry of spacetime. Actually the functional form of the metric can be entirely determined by symmetry considerations. In mathematical terms one says that if the cosmological principle holds there exists a natural 1+3 foliation of spacetime in three-dimensional uniform (i.e. homogeneous and isotropic) hypersurfaces. In this section we will explain what is the meaning of this sentence.

The cosmological principle implies that there exists a set of fundamental observers which are at rest relative to the mean motion of the nearby matter and for which the CMB radiation is isotropic.[1] These observers can synchronize their clocks (e.g. by using the local density of the Universe or the CMB temperature and exchanging light signals). We can then introduce a cosmic time coordinate t, which is the proper time measured by the clocks of these fundamental observers. By definition, then, the time-time component of the metric is $g_{00} = 1$ (since the time coordinate coincides with the proper time).

Let us consider one spatial hypersurface containing all the events with a given value of t and choose a three-dimensional coordinate system x^j to labels these points. Focus on the particle of the cosmological fluid which lies at x^i on the surface and use the same three spatial coordinates to label the location of that particle at all times. For obvious reasons, the coordinates x^j are called *comoving coordinates.*

[1] Sometimes the existence of this set of fundamental observers is called "Weyl's postulate". Actually this postulate states that there is one and only one geodesic passing through each point of spacetime. Hence fundamental observers possess a unique velocity at each point and form a sort of "cosmological perfect fluid".

Now consider another hypersurface of homogeneity to the future of the first. Suppose one fundamental observer reaches the second surface after a proper time t. Not to violate the assumption of homogeneity, all other observers starting from different spatial positions must reach the second surface in the same proper time. Proper time of fundamental observers can thus be used as the coordinate that labels the spacelike hypersurfaces (since it assumes a constant value on each of them).

The four-velocity of a fundamental observer at $\mathbf{x} = (ct, x_1, x_2, x_3)$ is $\mathbf{u} = d\mathbf{x}/d\tau = (u^t, 0, 0, 0)$. This must be orthogonal to the surface of homogeneity at cosmic time t. Were it not, its component in the surface would single out some direction violating the assumption of isotropy. Mathematically, we must have that the scalar product $\langle \mathbf{u}, \mathbf{v} \rangle = 0$ for any tangent vector to the hypersurface ($\mathbf{v} = (0, v_1, v_2, v_3)$). This corresponds to requiring $g_{0i} u^t v^i = 0$ and implies that $g_{0i} = 0$.

The conclusion is that the cosmological principle allows us to write the line element in the form

$$c^2 d\tau^2 = (cdt)^2 - d\Sigma^2(t) \tag{5.1}$$

where $d\Sigma^2$ denots the separation between events lying on a spatial hypersurface at constant cosmic time t.

Let us now consider a specific reference time t_1 and compute the spatial distance between two fundamental observers at x^k and $x^k + \Delta x^k$: $\Delta\Sigma^2 = \gamma_{ij}(x^k)\Delta x^i \Delta x^j$ with γ_{ij} the three-dimensional metric tensor. Not to violate the assumption of homogeneity, the ratio of the distance between the same observers at a different time t_2 must be independent of x^k and of Δx^k (i.e. considering all different pairs). The only form of the three-dimensional metric tensor which allows this is $\gamma_{ij}(t, x^k) = a^2(t)\,\sigma_{ij}(x^k)$. In other words the time evolution of the three-dimensional metric is uniform in space.

Therefore, the line element of a homogeneous and isotropic spacetime is

$$c^2 d\tau^2 = (cdt)^2 - a^2(t)\,\sigma_{ij}(x^k)\,dx^i\,dx^j \tag{5.2}$$

where $\sigma_{ij}(x^k)$ defines a time independent homogeneous and isotropic spatial geometry. From the mathematical point of view, using Schur's theorem, it is relatively easy to show that one must have

$$d\sigma^2 = \frac{dr^2}{1 - kr^2/R^2} + r^2\left(d\theta^2 + \sin^2(\theta)\,d\phi^2\right) = \frac{dr^2}{1 - kr^2/R^2} + r^2\,d\Omega^2\ , \tag{5.3}$$

with $k = 0, \pm 1$. Therefore we can distinguish three cases:

- **Positive curvature** Geodesics "accelerate" (in second derivative sense) towards each other. Initially parallel geodesics converge. The sum of the angles of a triangle is larger than 180^o. Visual example in 2D: great circles on a sphere.

- **Flat space** No geodesic acceleration. Initially parallel geodesics stay parallel. The sum of the angles of a triangle equals 180^o. Visual example in 2D: straight lines on a plane.

- **Negative curvature** Geodesics “accelerate” away from each other. Initially parallel geodesics diverge. The sum of the angles of a triangle is smaller than 180^o. Visual example in 2D: geodesics on a saddle.

We can then conclude that: *any spacetime obeying the cosmological principle is (locally) uniquely characterized by an integer number, k, a real number R, and a time-dependent function $a(t)$.*

Note that the coordinates in eq. (5.2) have been chosen in such a way to make the symmetries of space-time self-evident. It is easy to show that the world lines with x^i =constant are indeed geodesics (as assumed to build the metric).

Independently, Robertson (1935) and Walker (1936) demostrated that this is the most general form for the line element in a spatially homogeneous and isotropic space-time, (independent of general relativity theory). This metric was first used by Friedmann in 1922 and for this reason it is often called the Friedmann-Robertson-Walker (FRW) metric.

5.2.1 Some comments

The FRW metric has been derived assuming perfect isotropy and homogeneity at constant cosmic time. On the other hand, strictly speaking, we measure isotropy only from a specific point (the Earth). Can then we confidently use the FRW metric for our everyday applications? In 1968, Ehlers, Geren & Sachs proved that if a family of freely falling observers measure self-gravitating background radiation to be everywhere exactly isotropic, then the universe is exactly FRW. However, we know that the CMB is not exactly isotropic, should we then expect strong deviations from the FRW form in the real Universe? In the 1990s, Stoeger, Maartens & Ellis have shown that if one observer sees a nearly isotropic background radiation (for all times) then

$$g^{\mu\nu} \simeq \eta^{\mu\nu} + h^{\mu\nu} \tag{5.4}$$

where $\eta^{\mu\nu}$ is the FRW background metric and $h^{\mu\nu}$ is a small perturbation about it.[2]

5.2.2 Topology

The metric provides information on the local geometry of a manifold but does not say anything regarding its global properties. Therefore, from the metric we cannot say whether a given space is finite or infinite. Similarly, we

[2] Strictly speaking this results is limited to universes dominated by pressure-free matter.

cannot answer questions like: Has the universe holes or handles? Is it connected or not? To do this one needs to know the topology of the space under analysis. Most cosmology books assume the simplest topology (known as "simply-connected"). In this case, positively curved spaces are finite (from this the definition of "closed universes") while flat and negatively curved ones are infinite. However, by considering "multi-connected" topologies it is possible to consider universe models where space is finite whatever its curvature.

To build multi-connected spaces, mathematics teach us that one can start from one of the three types of "ordinary" (simply connected) spaces. Then, identification between some points change the shape of space and makes it multi-connected. From this one can build universe models where space is finite (although the curvature can be negative or zero) and of a really small volume. They are called "small universes". The simplest example is when our space would be a hypertorus having a radius lower than (say) ~ 1000 Mpc. In this case, the light rays would have had time to turn a few times "around" the universe. That would imply that each cosmic object (each galaxy for example) should produce many "ghost" images on the sky. The observed universe thus appears made up of the repetition of a same set of galaxies, although viewed at different look-back times.

An active branch of cosmology looks for signature of special topologies in the pattern of temperature fluctuations of the cosmic microwave background.

5.3 Kinematics of the FRW metric

We have shown that in a universe where the cosmological principle holds it is possible to choose a comoving set of coordinates (t, r, θ, ϕ) such that the spacetime metric assumes the form:

$$c^2 d\tau^2 = c^2 dt^2 - a^2(t) \left[\frac{dr^2}{1 - k\dfrac{r^2}{R^2}} + r^2 \left(d\theta^2 + \sin^2\theta \, d\phi^2 \right) \right] . \tag{5.5}$$

Here k is a constant which determines the geometry of the the three-dimensional hypersurfaces at constant t:

- if $k = +1$ the hypersurfaces of homogeneity are positively curved with curvature radius $a(t)\,R$;
- if $k = 0$ the hypersurfaces of homogeneity are flat (i.e. Euclidean);
- if $k = -1$ the hypersurfaces of homogeneity are negatively curved with curvature radius $a(t)\,R$.

The dimensionless factor $a(t)$ denotes the overall scale of the spatial part of the metric and, in general, can be a function of the time coordinate (which coincides with the proper time measured by fundamental observers).

Since only the product $a(t)r$ gives the radial coordinate distance on the hypersurface at cosmic time t, all the variables $a(t)$, r and R can be arbitrarily normalized up to a constant scaling factor. In observational cosmology, it is customary to normalize the expansion factor such that, at the present epoch t_0, $a(t_0) = 1$.[3] In this case, the FRW metric assumes the form

$$c^2 d\tau^2 = c^2 dt^2 - a^2(t) \left[\frac{dr^2}{1 - k\dfrac{r^2}{R_0^2}} + r^2 \left(d\theta^2 + \sin^2\theta \, d\phi^2 \right) \right] . \tag{5.6}$$

where the radial coordinate r is measured on the homogeneity hypersurface at t_0 (which has curvature radius R_0).

5.4 Observations in a FRW universe

In this section we discuss several phenomena related to the propagation of radiation in a FRW universe. These results are independent of the explicit functional form of $a(t)$.

5.4.1 Light propagation in a FRW universe

Cosmological observations are mainly based on electromagnetic radiation that is received from faraway sources. Since light travels at a finite speed, we observe all astronomical sources along our past light cone. In other words, the radiation we receive at time t_0 must have been emitted at some earlier time $t_{\rm e}$ such that the two events (photon emission and photon observation) are connected by a null geodesics. Therefore, given that cosmological distances are quite large, telescopes can be considered powerful time machines.

To describe this phenomenon in a quantitative way, let us consider two fundamental observers in a FRW universe (the light source and the "real" observer) and adopt the coordinate system (t, r, θ, ϕ). It is convenient to associate the origin of the coordinate system ($r = 0$) with the observer (the Cosmological Principle guarantees that we can do it) and consider an electromagnetic wave emitted at $(t_{\rm e}, r_{\rm g}, \theta_{\rm g}, \phi_{\rm g})$ and traveling along the $-r$ direction, with constant θ and ϕ. Since light travels along null geodesics ($d\tau = 0$), one has

$$c\,dt = -a(t)\,\frac{dr}{\sqrt{1 - kr^2/R_0^2}} \qquad\qquad \frac{c\,dt}{a(t)} = -\frac{dr}{\sqrt{1 - kr^2/R_0^2}} \, . \tag{5.7}$$

[3] From now on the subscript $_0$ will always indicate quantities evaluated at the present cosmic time.

It is convenient to introduce the new radial coordinate $\tilde{r}$ such that $d\tilde{r} = dr/\sqrt{1-kr^2/R_0^2}$. Therefore, integrating between emission and observation,

$$\int_{t_\mathrm{e}}^{t_\mathrm{o}} \frac{c\,dt}{a(t)} = \int_0^{r_\mathrm{g}} \frac{dr}{\sqrt{1-kr^2/R_0^2}} \equiv \tilde{r}_\mathrm{g} = \tag{5.8}$$

$$= \begin{cases} R_0 \ \mathrm{arcsin}(r_\mathrm{g}/R) & \text{if } k=+1 \\ r_\mathrm{g} & \text{if } k=0 \\ R_0\,\mathrm{arcsinh}(r_\mathrm{g}/R) & \text{if } k=-1 \end{cases} .$$

Using $\tilde{r}$ as the radial coordinate, the FRW metric can be written as

$$c^2d\tau^2 = c^2dt^2 - a^2(t)\left[d\tilde{r}^2 + S_k^2\left(\frac{\tilde{r}}{R_0}\right)\left(d\theta^2 + \sin^2\theta\,d\phi^2\right)\right] , \tag{5.9}$$

with

$$S_k(\tilde{r}) = \begin{cases} R_0 \ \sin(\tilde{r}/R_0) & (k=+1) \\ \tilde{r} & (k=0) \\ R_0 \ \sinh(\tilde{r}/R_0) & (k=-1) \end{cases} \tag{5.10}$$

Although the metrics in equations (5.5) and (5.9) look different, they represent the same spaces. They apper different just because of the different choice of radial coordinates. With $\tilde{r}$ as radial coordinate, radial distances are “Euclidean” but angular distances are not (unless $k=0$). With r as radial coordinate, the reverse is true.

5.4.2 Cosmological time dilation and redshift

Consider a wave packet emitted within the coordinate-time interval $(t_\mathrm{e}, t_\mathrm{e} + \Delta t_\mathrm{e})$ by a distant galaxy (intended as a fundamental observer). This wave packet is received by another fundamental observer in the cosmic time interval $(t_\mathrm{o}, t_\mathrm{o} + \Delta t_\mathrm{o})$. Using the same emitter-observer configuration as in eq. (5.7), we get (since we are using comoving coordinates):

$$\int_{t_\mathrm{e}}^{t_\mathrm{o}} \frac{c\,dt}{a(t)} = \tilde{r}_\mathrm{g} \qquad \text{and} \qquad \int_{t_\mathrm{e}+\Delta t_\mathrm{e}}^{t_\mathrm{o}+\Delta t_\mathrm{o}} \frac{c\,dt}{a(t)} = \tilde{r}_\mathrm{g} , \tag{5.11}$$

which, assuming that $a(t)$ changes very little during a short time interval, implies

$$\frac{c\,\Delta t_\mathrm{o}}{a(t_\mathrm{o})} - \frac{c\,\Delta t_\mathrm{e}}{a(t_\mathrm{e})} = 0 \qquad \Longrightarrow \qquad \Delta t_\mathrm{o} = \frac{a(t_\mathrm{o})}{a(t_\mathrm{e})}\Delta t_\mathrm{e} . \tag{5.12}$$

If the Universe is expanding (as indicated observationally by the Hubble’s law), distant galaxies emitted the light we receive now at an epoch when the expansion factor was smaller. Therefore phenomena are observed to take longer in our reference frame than they do in that of the source. This phenomenon is called *cosmological time dilation* and provides a direct way of testing the FRW formalism. As we have already mentioned in our introductory class, Goldhaber et al. measured this effect in 1997 by studying the light-curve of SNae Ia.

Consider now the emission of a monochromatic electromagnetic wave, and indicate with $\Delta t_{\rm e} = \lambda_{\rm e}/c$ the time interval between the emission of two wave crests. Because of the cosmic time dilation effect, the observer would detect radiation at a different wavelength

$$\lambda_{\rm o} = c\Delta t_{\rm o} = \frac{a(t_{\rm o})}{a(t_{\rm e})}\,\lambda_{\rm e}\ . \tag{5.13}$$

In an expanding universe, the observed wavelength is always longer (i.e. shifted towards the red) with respect to $\lambda_{\rm e}$. This is conventionally expressed in terms of a *redshift parameter* z defined as the fractional increase with wavelength:[4]

$$z = \frac{\lambda_{\rm o} - \lambda_{\rm e}}{\lambda_{\rm e}} = \frac{\nu_{\rm e} - \nu_{\rm o}}{\nu_{\rm o}}\ . \tag{5.14}$$

Therefore, we can write

$$1 + z = \frac{a(t_{\rm o})}{a(t_{\rm e})}\ . \tag{5.15}$$

In simple terms, the cosmological redshift is a measure of the ratio between the scale factors of the Universe at $t_{\rm o}$ and at $t_{\rm e}$. Note that the cosmological redshift does not depend on the functional form of $a(t)$ (i.e. on whether it is monotonic or oscillates or jumps suddenly) but only on its initial and final values at emission and observation.

Cosmological redshift is an observational phenomenon discovered by V. M. Slipher in 1912 when taking spectra of distant galaxies. If we identify the time of observation with the present, we can then write

$$1 + z = \frac{a(t_0)}{a(t_{\rm e})} = \frac{1}{a(t_{\rm e})}\ . \tag{5.16}$$

When a cosmologist says that a given galaxy is “at redshift 3” he/she means that the photons we now receive have been emitted by the galaxy when the universe was a factor of 4 smaller than today. Note that this is only a statement regarding the size of the universe at emission and not about cosmic time at emission.

5.4.3 Measures of distance

In this section we discuss how we can define (and measure) distances between fundamental observers in a FRW universe. For analytical convenience, we place the origin of the coordinate system on the Milky Way (that we consider a fundamental observer).

[4] Note that if $z < 0$ one speaks of a *blueshift*.

Proper distance

Imagine that all the fundamental observers of the FRW metric measure the distance to their closest neighbour at the same cosmic time t (for instance by measuring the travel time for a light signal). We call *proper distance* the distance obtained by summing up all these contributions. For instance, the proper distance of a galaxy with radial comoving coordinate r_g at cosmic time t is

$$d_\mathrm{prop} = a(t)\int_0^{r_\mathrm{g}} \frac{dr}{\sqrt{1-kr^2/R_0^2}} = a(t)\begin{cases} R_0 \operatorname{arcsin}(r_\mathrm{g}/R_0) & \text{if } k=+1 \\ r_\mathrm{g} & \text{if } k=0 \\ R_0 \operatorname{arcsinh}(r_\mathrm{g}/R_0) & \text{if } k=-1 \end{cases} . \tag{5.17}$$

This is more easily expressed in terms of the radial coordinate $\tilde{r}$:

$$d_\mathrm{prop} = a(t)\,\tilde{r}_\mathrm{g} \,. \tag{5.18}$$

In simple words, the proper distance is the actual physical distance that separates two events on the same hypersurface of homogeneity at constant cosmic time. Due to the expansion (or contraction) of the universe, d_prop scales proportionally to the expansion factor $a(t)$.

Note that the proper distance is impossible to measure in practice and therefore its notion is not very relevant for observational cosmology.

Hubble's law

It is interesting to study how the proper distance of a galaxy evolves with time. From eq. (5.18) we obtain

$$\dot{d}_\mathrm{prop} = \dot{a}\,\tilde{r}_\mathrm{g} = \frac{\dot{a}}{a}\,d_\mathrm{prop} = H(t)\,d_\mathrm{prop} \,, \tag{5.19}$$

where the dot denotes derivatives with respect to cosmic time t. Eq. (5.19) encodes the Hubble's law and $H(t)$ is called the Hubble parameter. Note that $H(t)$ is constant on a hypersurface at constant t but evolves with cosmic time.

The Hubble's law is very well established observationally and constitutes one of the pillars of the standard big bang model.

Comoving distance

The comoving distance between two fundamental observers of the FRW spacetime is the proper distance intercurring between them at a given (prefixed) cosmic time t_ref. For analytical convenience, it is common practice to identify t_ref with the present epoch t_0. In this case:

$$d_\mathrm{com} = \frac{a(t_0)}{a(t)}\,d_\mathrm{prop} = (1+z)\,d_\mathrm{prop} \,. \tag{5.20}$$

Therefore the comoving distance of a galaxy at redshift z is the proper distance that the galaxy would have at redshift 0 (i.e. at the present time) if it would move following the overall expansion of the Universe. Note that the comoving distance of a galaxy from the Earth coincides with its $\tilde{r}$-coordinate distance.

Angular diameter distance

In Euclidean space, we can measure the distance d of an object by comparing its angular apparent size $\Delta\theta$ to its proper length perpendicular to the line of sight Δl (also called the transverse size). For small angles, this simply gives $\Delta\theta = \Delta l/d$. In cosmology, we can define an *angular diameter distance*, $d_{\rm A}$, such that the relation between $d_{\rm A}$ and $\Delta\theta$ looks like the standard Euclidean relation:

$$d_{\rm A} = \frac{\Delta l}{\Delta\theta} \, . \tag{5.21}$$

The fact that we see an object with transverse size Δl means that there are two null geodesics connecting the opposite extremes of the object to us. By construction, in a FRW universe, world lines with constant spatial coordinates are geodesics. Therefore, we can rotate the coordinate system to place the two extremes of the object at coordinates $(t_{\rm g}, r_{\rm g}, \Delta\theta_{\rm g}/2, 0)$ and $(t_{\rm g}, r_{\rm g}, -\Delta\theta_{\rm g}/2, 0)$. According to the FRW metric, the proper distance between these events is $\Delta l = a(t_{\rm g})\, r_{\rm g}\, \Delta\theta$ so that

$$d_{\rm A} = a(t_{\rm g})\, r_{\rm g} \, . \tag{5.22}$$

Remember that $t_{\rm g}$ and $r_{\rm g}$ are not independent: $t_0 - t_{\rm g}$ is the cosmic time interval during which light travels a coordinate distance $r_{\rm g}$ as in eq. (5.8). In an expanding universe, $a(t_{\rm g})$ decreases as $r_{\rm g}$ increases. In particular, using the redshift

$$d_{\rm A} = \frac{r_{\rm g}}{1+z} \tag{5.23}$$

Therefore, in some models, $d_{\rm A}$ does not monotonically increase with $r_{\rm g}$ and the angular size of very distant objects (with fixed proper size) can increase with their distance.

Equations (5.22) and (5.23) show that the coordinate distance $r_{\rm g}$ acts as a sort of “comoving angular-diameter distance” which can be used to associate angle sizes to the comoving transverse size of an object.

Luminosity distance

In Euclidean space, we can determine the distance d of a light source by comparing its flux f (energy received per unit time per unit surface) with its absolute luminosity L (energy emitted per unit second): $f = L/4\pi\, d^2$.

In cosmology, we can define a *luminosity distance*, d_L, such that f and L follow the same relation as in Euclidean space:

$$d_\mathrm{L} = \sqrt{\frac{L}{4\pi\, f}} \ . \tag{5.24}$$

Consider a light source at coordinate distance r_g.

- Due to the expansion of the Universe, each photon emitted with energy E will be redshifted to energy

$$E\, a(t_\mathrm{e})/a(t_0) = E/(1+z) \ ,$$

- At the same time, photons emitted at time intervals Δt_e will be received at time intervals

$$\Delta t_\mathrm{e}\, a(t_0)/a(t_\mathrm{e}) = \Delta t_\mathrm{e}\, (1+z) \ ;$$

- Moreover, at t_0, the photons from the source will be distributed over a sphere of proper surface area $4\pi\, r_\mathrm{g}^2$.

In summary, the observed flux at t_0 will be

$$f_\nu(\nu_\mathrm{o}) = \frac{L_\nu[(1+z)\nu_\mathrm{o}]}{4\pi\, r_\mathrm{g}^2\, (1+z)} \ . \tag{5.25}$$

Integrating over frequencies one obtains

$$f_\mathrm{bol} = \frac{L_\mathrm{bol}}{4\pi\, r_\mathrm{g}^2\, (1+z)^2} \ . \tag{5.26}$$

which, combined with eq. (5.24), gives

$$d_\mathrm{L} = r_\mathrm{g}\, (1+z) \ . \tag{5.27}$$

When expressed in terms of the $\tilde{r}$ radial coordinate, the luminosity distance assumes the form

$$d_\mathrm{L} = S_k(\tilde{r}_\mathrm{g}/R_0)\, (1+z) \ . \tag{5.28}$$

Finally, the observed flux density can be written as:

$$f_\nu(\nu_\mathrm{o}) = \frac{(1+z)\, L_\nu[(1+z)\nu_\mathrm{o}]}{4\pi\, d_\mathrm{L}^2} \ . \tag{5.29}$$

Surface brightness

Astronomers often deal with extended (i.e. non point-like) sources of light. To characterize the emission properties from these sources it is convenient to use their absolute luminosity per unit transverse area, $\mathcal{I}_\nu$, (emitted energy per unit time per unit surface per unit frequency per unit solid angle). On the other hand, the photon flux we receive from them is quantified in terms of the specific intensity of radiation, I_ν, (received energy per unit time per unit surface per unit frequency per unit solid angle). As we have shown in the class dedicated to radiative transfer, in Euclidean space these two quantities coincide numerically and are generally indicated with the term of *surface brightness.* What happens in a FRW universe? Consider an extended source with proper transverse area $\Delta S_\perp$ which subtends a solid angle $\Delta\Omega$. Therefore, combining our previous results for the luminosity distance and for the angular diameter distance, we obtain

$$I_\nu(\nu_\mathrm{o}) = \frac{f_\nu(\nu_\mathrm{o})}{\Delta\Omega} = \frac{(1+z)\, d_\mathrm{A}^2\, L_\nu[(1+z)\,\nu_o]}{4\pi\, d_\mathrm{L}^2\, \Delta S_\perp} = \frac{\mathcal{I}_\nu[(1+z)\nu_\mathrm{o}]}{(1+z)^3}\ , \tag{5.30}$$

and

$$I_\mathrm{bol} = \frac{\mathcal{I}_\mathrm{bol}}{(1+z)^4}\ . \tag{5.31}$$

These results have been first derived by Tolman in the 1930s. Cosmological surface brightness dimming effects play a dominant role in setting what is observed at high redshifts.

5.4.4 Radiative transfer in the expanding universe

Consequently, the equation of cosmological radiative transfer in comoving coordinates is:

$$\frac{1}{c}\frac{\partial I_\nu}{\partial t} + \frac{\hat{n}\cdot\nabla I_\nu}{\bar{a}} - \frac{H(t)}{c}\left(\nu\frac{\partial I_\nu}{\partial\nu} - 3I_\nu\right) = -\alpha_\nu\, I_\nu + j_\nu\ , \tag{5.32}$$

where $\bar{a} = (1+z_\mathrm{em})/(1+z)$ is the ratio of cosmic scale factors between photon emission at frequency ν and the time t. With respect to the non-expanding case we recognize two modifications:

- the denominator in the second term, which accounts for changes in the path length along a ray due to cosmic expansion;

- the third term, which accounts for cosmological redshift and dilution.

5.4.5 The age of the Universe

Let us consider eq. (5.16) and take the first derivative with respect to cosmic time:

$$\begin{aligned} \frac{da}{dt} = -\frac{1}{(1+z)^2}\frac{dz}{dt} \qquad &\Longrightarrow \qquad \frac{1}{a}\frac{da}{dt} = -\frac{1}{1+z}\,\frac{dz}{dt} \\ &\Longrightarrow \qquad dt = -\frac{dz}{(1+z)\,H(z)}\ . \end{aligned} \tag{5.33}$$

This equation relates cosmic time to redshift. Consider two galaxies at redshifts z_1 and $z_2 > z_1$, what is the time lag between the epochs at which they emitted the light that we now receive from them? This is easily computed by integrating our last equation:

$$t_1 - t_2 = \int_{z_1}^{z_2} \frac{dz}{(1+z)\,H(z)}\ . \tag{5.34}$$

Note that, contrary to redshift itself, this depends on the expansion history of the Universe which is encoded in the function $H(z)$. Similarly, assuming that redshift can assume all positive values (i.e. that the Universe started from a state with extremely high density), we can compute the age of the Universe as:

$$t_{\rm age} = \int_0^\infty \frac{dz}{(1+z)\,H(z)}\ . \tag{5.35}$$

5.4.6 Distance-redshift relations

It is often convenient to express all the distances we have defined in terms of the redshift parameter. This is easily done by using eq. (5.8) which gives

$$d_{\rm com} = \tilde{r} = c\int_0^z \frac{dz'}{H(z')}\ . \tag{5.36}$$

This is the comoving distance of source at redshift z. The corresponding proper distance is:

$$d_{\rm prop} = \frac{c}{1+z}\int_0^z \frac{dz'}{H(z')}\ , \tag{5.37}$$

while the luminosity and angular diameter distances can be written as

$$d_{\rm L} = (1+z)\,S_k\left(\frac{d_{\rm com}(z)}{R_0}\right) \qquad\qquad d_{\rm A} = \frac{1}{1+z}\,S_k\left(\frac{d_{\rm com}(z)}{R_0}\right)\ . \tag{5.38}$$

Volume and galaxy number counts

The proper volume element of an hypersurface of homogeneity can be derived directly from the FRW metric:

$$\begin{aligned} dV_{\mathrm{prop}} &= a^3(t)\,\frac{r^2\,dr}{\sqrt{1-k(r/R_0)^2}}\,\sin\theta\,d\theta\,d\phi = \\ &= a^3(t)\,S_k^2(\tilde{r}/R_0)\,d\tilde{r}\,\sin\theta\,d\theta\,d\phi = \\ &= a^3(t)\,\frac{d_{\mathrm{L}}^2}{(1+z)^2}\,dd_{\mathrm{com}}\,\sin\theta\,d\theta\,d\phi\,. \end{aligned} \tag{5.39}$$

The comoving volume on our past light-cone (up to redshift z) is therefore given by:

$$V_{\mathrm{lc}} = 4\pi\,c\int_0^z S_k^2\left(\frac{d_{\mathrm{com}}(z')}{R_0}\right)\frac{dz'}{H(z')}\,. \tag{5.40}$$

To proceed any further we need to know the form of $H(z)$, which is turn depends on the functional form of $a(t)$.

5.5 Dynamics of the FRW universe

We have seen that, in full generality, a homogeneous and isotropic universe is described by the FRW metric. This metric is completely specified by three quantities: the curvature sign (k), the present-day curvature radius (R_0) of the spatial section at constant cosmic time, and the expansion parameter $a(t)$ (that we have decided to normalize so that $a(t_0) = 1$). In this section we will study how these quantities relate to the energy content of the Universe.

5.6 The Friedmann equations

If we use general relativity as our theory of gravity, we obtain an evolution equation for the expansion factor of the Universe. Combining Einstein's field equations (which equate space-time curvature as embodied by the Einstein tensor to the energy-momentum tensor of the Universe),

$$G_{\mu\nu} = \frac{8\pi G}{c^4}\,T_{\mu\nu}\;, \tag{5.41}$$

with the FRW metric requires the energy-momentum tensor of the Universe to be:

$$T_{\mu\nu} = \left(\rho + \frac{p}{c^2}\right)\,U_\mu\,U_\nu - p\,g_{\mu\nu} \tag{5.42}$$

with $U^\mu = dx^\mu/d\tau = (c, 0, 0, 0)$. In the comoving frame this gives

$$T^\nu_\mu = \begin{pmatrix} \rho\,c^2 & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix}\,. \tag{5.43}$$

This is the form of the energy-momentum tensor for a perfect fluid (no shear stresses, viscosity or heat conduction) which is, on average, at rest in the comoving coordinate system. Therefore we can interpret ρ and p as the proper mass density and pressure (intended as the flux density of x-momentum in the x-direction etc.) of the "cosmic fluid".

With this energy-momentum tensor, the time-time component of Einstein equation gives

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + 3\frac{p}{c^2}\right) , \tag{5.44}$$

while the space-space components give the single equation

$$a\ddot{a} + 2\dot{a}^2 + 2\frac{kc^2}{R_0^2} = 4\pi G\left(\rho - \frac{p}{c^2}\right)a^2 \tag{5.45}$$

and, finally, the space-time components give $0 = 0$. By eliminating $\ddot{a}$ from eqs. (5.44) and (5.45) one obtains a first-order differential equation for $a(t)$:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho(t) - \frac{kc^2}{R_0^2}\frac{1}{a^2(t)} . \tag{5.46}$$

which has been first derived by Alexander Alexandrovich Friedmann (a Russian mathematician and metereologist) in 1922 and is now known as (first) Friedmann equation (eq. (5.44) is often called the second Friedmann equation or the acceleration equation).

5.6.1 Adiabatic expansion

Equations (5.44) and (5.46) can be combined together to obtain[5]

$$\frac{d}{dt}\left(\rho\, c^2\, a^3\right) = -p\,\frac{da^3}{dt} \tag{5.47}$$

or, equivalently,

$$\frac{d}{da}\left(\rho\, c^2\, a^3\right) = -3p\, a^2 . \tag{5.48}$$

Note that this relation embodies the first law of thermodynamics

$$\delta Q = dE + p\, dV \tag{5.49}$$

where δQ is the heat flow into a region, dE is the change in internal energy, p is the pressure and dV is the change in volume of the region. If the Universe is homogeneous, then for any volume $\delta Q = 0$. In other words, the expansion process must be adiabatic and does not change the total entropy $dS = \delta Q/T = 0$:

$$\frac{dE}{dt} + p\,\frac{dV}{dt} = 0 . \tag{5.50}$$

[5] Multiply eq. (5.46) by a^2 and differentiate the result with respect to t. Then substitute $\ddot{a}$ from eq. (5.44) and finally multiply by $a\, c^2$.

5.7 The equation of state

Let us summarize where we got so far. We have derived three equations that describe how the Universe expands

$$\begin{aligned} \left(\frac{\dot a}{a}\right)^2 &= \frac{8\pi G}{3}\,\rho(t) - \frac{kc^2}{R_0^2}\frac{1}{a^2(t)}\ , && (5.51) \\ \frac{\ddot a}{a} &= -\frac{4\pi\, G}{3}\left(\rho + 3\,\frac{p}{c^2}\right)\ , && (5.52) \\ \dot\rho + 3\frac{\dot a}{a}\left(\rho + \frac{p}{c^2}\right) &= 0\ , && (5.53) \end{aligned}$$

but only two of them are independent: whichever two you pick, the third necessarily follows. These equations, however, include three unknown functions of cosmic time: $a(t)$, $\rho(t)$ and $p(t)$. We thus need another equation to close our system. It is convenient to express this additional relation in terms of an *equation of state*, $p = p(\rho)$, i.e. a mathematical relation between the pressure and energy density of the stuff that fills up the Universe.

In general, equations of state can be extremely complicated. Many cases are known where the pressure is a daunting non-linear function of the density (these are most commonly encountered in condensed-matter physics and in the astrophysics of compact objects). However, cosmology deals with very low densities (dilute gases) where the equation of state is very simple. In most cases of interest, the equation of state is linear

$$p = w\,\rho\,c^2 \tag{5.54}$$

where w is a dimensionless number. Some values of w are of particular interest.

- **Non-relativistic matter:** $w = 0$. A non-relativistic ideal gas obeys the perfect-gas law $p = n\,k_{\rm B}\,T$ (where n is the particle number density, T the gas temperature and $k_{\rm B}$ the Boltzmann constant). The energy density of a non-relativistic gas is almost entirely contributed by the rest mass of gas particles: $\epsilon = \rho\,c^2 \simeq \rho_{\rm rest}\,c^2$. The temperature T, the particle rest-mass $m_{\rm p}$ and the root-mean-square thermal velocity $\langle {\rm v}^2\rangle$ are associated by the relation $\langle {\rm v}^2\rangle = 3k_{\rm B}\,T/m_{\rm p}$. Thus, the equation of state for a non-relativistic gas can be written in the form $p = w\,\epsilon$ with $w \simeq \langle {\rm v}^2\rangle/(3c^2) \ll 1$.[6] Cosmologist and relativists often refer to all forms of "pressureless", non-relativistic matter using the term "dust" (not to be confused with dust grains present in the interstellar and intergalactic medium).

[6] For ionized hydrogen, electrons are non-relativistic as long as $T \ll 6\times10^9$ K and the protons when $T \ll 10^{13}$ K.

- **Relativistic matter and radiation:** $w = 1/3$**.** A gas of photons (or other massless particles) is fully relativistic. You might recall from your studies of statistical mechanics that the corresponding equation of state is $p = \epsilon/3 = \rho\, c^2/3$. Cosmologists and relativists often refer to all forms of relativistic matter with the generic term of "radiation".

When energy exchanges between different components (e.g. non-relativitic matter and radiation) are negligible (as we will see, barring some very early phases of the life of the Universe, this is always the case), eq. (5.47) holds for each component separately.

- For non-relativistic matter ($p \simeq 0$), this implies $\rho \propto a^{-3}$ which expresses the conservation of the number of particles;

- for ultra-relativistic components ($p = \rho\, c^2/3$) the energy-conservation equation gives $\rho \propto a^{-4}$ which embodies the conservation of the number of particles together with the cosmological redshift which reduces the energy of each particle proportionally to a^{-1}.

In general, for an equation of state $p = w\, \rho\, c^2$ with constant w, the energy density evolves as

$$\rho \propto a^{-3\,(1+w)}\ . \tag{5.55}$$

Note the particular case $w = -1$, where ρ keeps constant with the expansion of the Universe. In Section 5.9, we will discuss a form of energy which behaves this way.

5.8 Friedmann models

In general, a world model which is based on

1. the cosmological principle (and thus the FRW metric);

2. general relativity (and thus the Friedmann equation);

3. the energy-conservation equation;

4. an equation of state for the cosmological fluid;

is called a Friedmann model of the Universe.

5.9 The cosmological constant

The state of observational astronomy in 1917 (when Einstein published his first cosmological model based on general relativity) was such that:

- the Kapteyn model of the Milky Way was favoured by some (but not all) astronomers;

- there was no agreement on the origin of “spiral nebulae”.
- there was no evidence of cosmic expansion.

Therefore, it is not surprising that Einstein was interested in finding a static ($\dot{a} = 0$) solution. Actually, his hope was also that general relativity would embody Mach’s principle that distant matter determines local inertia and for this reason he was looking for a finite model with positive curvature..

Note that a static universe with a positive energy density is compatible with the first Friedmann equation if the spatial curvature is positive ($k = +1$) and the density is appropriately tuned. However, eq. (5.44) implies that $\ddot{a}$ will never vanish in such a spacetime if the pressure p is also non-negative (which is true for most forms of matter, and certainly for ordinary sources such as stars and gas). Einstein therefore proposed a modification of his field equations

$$G_{\mu\nu} - \Lambda\, g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}\ , \tag{5.56}$$

where he introduced a new fundamental constant of nature (the *cosmological constant*) to balance the attractive force of gravity. [7] With this modification, the Friedmann equations become

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho(t) - \frac{kc^2}{R_0^2}\frac{1}{a^2(t)} + \frac{\Lambda\, c^2}{3}\ , \tag{5.57}$$

and

$$\frac{\ddot{a}}{a} = -\frac{4\pi\, G}{3}\left(\rho + 3\,\frac{p}{c^2}\right) + \frac{\Lambda\, c^2}{3}\ , \tag{5.58}$$

These equations admit a static solution. Considering a universe dominated by non-relativistic matter and imposing $\dot{a} = \ddot{a} = 0$, one finds

$$\Lambda\, c^2 = 4\pi G\,\rho \qquad k = +1 \qquad R_0 = \frac{c}{\sqrt{\Lambda}}\ . \tag{5.59}$$

This solution is called “Einstein static universe”. Some years later, Willem de Sitter (but also Friedmann and Eddington noted it) pointed out that Einstein’s static universe was unstable against an overall expansion or contraction. A tiny perturbation of the solution ends in runaway exponential expansion or in a Big Crunch. Einstein agreed, and together they published a paper in 1932 proposing the Einstein-de Sitter model of the universe.

It was later discovered by Edwin Hubble (1929) that other galaxies appear to be moving away from us, i.e. that the universe is actually expanding. This discovery prompted Einstein to abandon the cosmological constant but

[7] The left-hand side of (5.56) is the most general local, coordinate-invariant, divergenceless, symmetric, two-index tensor we can construct solely from the metric and its first and second derivatives.

caused other scientists to embrace it. Hubble badly underestimated the distances to galaxies and hence overestimated the value of the Hubble constant. Hubble's value of $H_0 = 500\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ corresponds to an Hubble time of $t_\mathrm{H} = H_0^{-1} = 2$ Gyr which is less than half the age of the Earth as estimated from radioactive dating. It was soon realized that if the value of Λ is large enough to make $\ddot{a} > 0$, then $\dot{a}$ was smaller in the past than it is now and consequently the universe is older than the Hubble time.

Since those times, the cosmological constant has gone in and out of fashion. Nowadays, there are reasons to believe that Λ may still be a viable part of cosmology. A common interpretation is that Λ measures the energy density of the quantum vacuum state. Alternatively, it might embody an exotic form of energy associated with an unknown field.

5.10 The Friedmann equations in terms of observables

In order to apply the Friedmann equations to the real universe, we must have some way of tying them up to observational properties. We start by noticing that the left-hand side of eq. (5.57) coincides with the Hubble parameter $H(t)$. Therefore, the first Friedmann equation evaluated at the present time can then be written as:

$$H_0^2 = \frac{8\pi G}{3}\,\rho_0 + \frac{\Lambda\,c^2}{3} - \frac{kc^2}{R_0^2}. \tag{5.60}$$

We can rearrange the terms of this equation to get:

$$\frac{kc^2}{R_0^2} = \frac{8\pi G}{3}\,\rho_0 + \frac{\Lambda\,c^2}{3} - H_0^2 \tag{5.61}$$

which shows that the sign of k (i.e. the geometry of the three-dimensional hypersurfaces of homogeneity at constant cosmic time) depends on the present energy density of the Universe and on the value of the cosmological constant.[8] In particular, if $\Lambda = 0$, there is a *critical density* (which depends on the value of the Hubble constant)

$$\rho_\mathrm{crit}(t) = \frac{3\,H^2(t)}{8\pi\,G} \tag{5.62}$$

such that

- $k > 0$ if $\rho_0 > \rho_\mathrm{crit}$;

[8] This might not seem surprising given that Einstein's field equations relate space-time curvature to energy density. Remember, however, that here we are talking of space curvature at constant cosmic time which is a very different concept from the curvature of space-time.

- $k = 0$ if $\rho_0 = \rho_{\rm crit}$;
- $k < 0$ if $\rho_0 < \rho_{\rm crit}$.

There is then a direct correspondence between the energy density of the Universe and the curvature of its homogeneity hypersurfaces. It is convenient to parameterize the Hubble constant through the dimensionless parameter h such that $H_0 = 100\, h\, {\rm km\, s^{-1}\, Mpc^{-1}}$. In this case,

$$\rho_{\rm crit} = 1.88 \times 10^{-29}\, h^2\, {\rm g\, cm^{-3}} = 2.778 \times 10^{11}\, h^2\, M_\odot\, {\rm Mpc^{-3}}\ . \qquad (5.63)$$

Cosmologists often measure the mean energy density of the Universe in units of the critical density by introducing the *density parameter*

$$\Omega(t) = \frac{\rho(t)}{\rho_{\rm crit}(t)}\ . \qquad (5.64)$$

If different components are contributing to the energy budget (for instance non-relativistic matter and radiation), it is convenient to define a density parameter for each of them

$$\Omega_i(t) = \frac{\rho_i(t)}{\rho_{\rm crit}(t)}\ . \qquad (5.65)$$

Similarly, we can define a density parameter for the cosmological constant and one for the curvature

$$\begin{aligned} \Omega_\Lambda &= \frac{\Lambda\, c^2}{3\, H^2(t)} && (5.66) \\ \Omega_k &= -\frac{kc^2}{R_0^2 H^2(t)} && (5.67) \end{aligned}$$

so that the first Friedmann equation becomes

$$\Omega_{\rm r}(t) + \Omega_{\rm m}(t) + \Omega_\Lambda(t) + \Omega_k(t) = 1\ . \qquad (5.68)$$

This holds at any cosmic time including the present one. If combined with a simultaneous measure of $\Omega_{\rm r}(t_0)$, $\Omega_{\rm m}(t_0)$, $\Omega_\Lambda(t_0)$ and $\Omega_k(t_0)$, eq. (5.68) would then provide a challenging test for the Friedmann models.[9] On the other hand, assuming that the models are correct, if you know $\Omega_{\rm m}$ and Ω_Λ, you can derive sign of the curvature k.[10] If, in addition, you know the Hubble radius, c/H_0, you can compute the present radius of curvature R_0

$$R_0 = \frac{c}{H_0}\, |1 - \Omega_{\rm tot}|^{-1/2} \qquad (5.69)$$

[9] Note, however, that at present we cannot measure Ω_k directly.

[10] Committing a little abuse of notation, in the cosmological literature it is customary to use the symbols $\Omega_{\rm r}$, $\Omega_{\rm m}$, Ω_Λ to indicate the present-day values of the functions $\Omega_{\rm r}(t)$, $\Omega_{\rm m}(t)$, $\Omega_\Lambda(t)$. From now on we will adopt this notation. The time-dependent functions will be always written expliciting their time dependence.

where $\Omega_{\rm tot} = \Omega_{\rm r} + \Omega_{\rm m} + \Omega_\Lambda$.

As we have already stressed, the first Friedmann equation gives the evolution of the Hubble parameter. Combining it with the results of the energy-conservation equation we find

$$\frac{H^2(a)}{H_0^2} = \frac{\Omega_{\rm r}}{a^4} + \frac{\Omega_{\rm m}}{a^3} + \frac{\Omega_k}{a^2} + \Omega_\Lambda \ , \tag{5.70}$$

which can be re-written as an evolution with redshift

$$\frac{H^2(z)}{H_0^2} = (1+z)^4\,\Omega_{\rm r} + (1+z)^3\,\Omega_{\rm m} + (1+z)^2\,\Omega_k + \Omega_\Lambda \ . \tag{5.71}$$

This equation provides the fundamental element to compute how the comoving, luminosity and angular-diameter distances evolve as a function of redshift.

5.11 Constraints on the cosmological parameters

Currently, the combination of different datasets (mainly CMB temperature anisotropies, Hubble diagram of Type Ia supernovae, and galaxy clustering) gives the following constraints:

$$\begin{aligned} \Omega_{\rm m} &= 0.239 \pm 0.018 \\ \Omega_{\rm b} &= 0.0416 \pm 0.002 \\ \Omega_\Lambda &= 0.761 \pm 0.018 \\ \Omega_{\rm tot} &= 1.003 \pm 0.010 \\ \Omega_{\rm r} &< 0.024 \text{ including neutrinos} \\ H_0 &= 73 \pm 2 \ {\rm km\,s^{-1}\,Mpc^{-1}} \ . \end{aligned}$$

The age of the Universe comes out to be $t_{\rm age} = 13.76 \pm 0.15$ Gyr and the age-redshift relation can be easily computed using the integral given above.

This provides the background that we will use to interpret the observations of the IGM and model its evolution with time.

Chapter 6

Quasar absorption lines: observations

6.1 Quasar spectra

Quasars are amongst the most luminous objects in the universe and can therefore be observed out to very high redshift. Figure 6.1 shows a typical quasar spectrum in the optical waveband taken with a high-resolution instrument and long integration time (corresponding to high signal-to-noise ratio, i.e. to high precision). The following features are evident

- The spectrum of a quasar consists of a power-law continuum (typically $F_\lambda \propto \lambda^{-1}$) and broad emission lines (with a Doppler parameter of several thousand km s^{-1}). If the spectrum contains more than one emission line, it might be possible to identify the atomic transitions to which they correspond (using the ratio of their wavelengths and also their strengths). This allows us to measure the redshift of the quasar via $1 + z_{\mathrm{qso}} = \lambda_{\mathrm{obs}}/\lambda_{\mathrm{em}}$ where λ_{obs} is the observed wavelength of the emission line and λ_{em} is the corresponding rest-frame wavelength. One way this is done in practice is by fitting the observed data with template quasar spectra with varying redshift.
- On the blue side of the hydrogen Lyα (emission) line, many narrow absorption lines are visible.
- On the red side of the hydrogen Lyα (emission) line, a few narrow absorption lines are visible.

What is the origin of these spectral features? We believe that quasars are the active nuclei of distant galaxies where super-massive black holes (with a mass of $10^8 - 10^9 M_\odot$) grow by swallowing gas through an accretion disk. Models show that both the quasar continuum and the emission lines are produced by gas within the “central engine” of the active galaxy. What

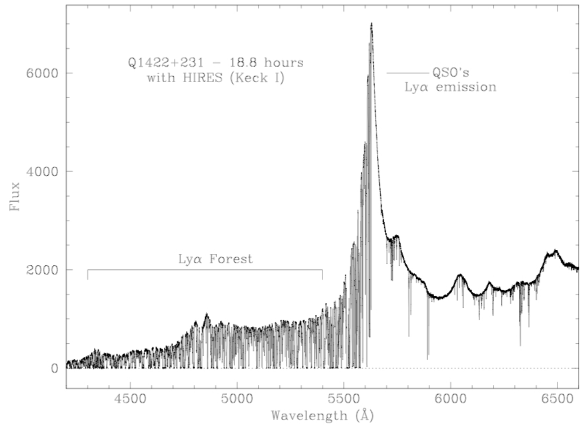


Figure 6.1: The optical spectrum of the quasar Q1422+231 (at redshift $z = 3.622$) taken with the HIRES spectrograph on the Keck telescope.

about the absorption lines, then? The light from the quasar traverses vast distances before reaching our telescopes on Earth. Any atom which happens to lie along the line of sight leaves its signature on the spectrum of the quasar in the form of absorption lines. It was soon realized (Lynds 1971) that the rich series of absorption lines blueward the quasar Lyα emission line is due to individual gas "clouds" containing neutral hydrogen. Imagine to live in one of these clouds, at redshift $z_{\rm c}$ such that $0 < z_{\rm c} < z_{\rm qso}$. For you, the quasar would have a redshift

$$1 + z_{\rm qc} = \frac{a_{\rm c}}{a_{\rm qso}} = \frac{a_{\rm c}}{a_0}\frac{a_0}{a_{\rm qso}} = \frac{1 + z_{\rm qso}}{1 + z_{\rm c}} \, . \tag{6.1}$$

Therefore, you would see the quasar spectrum redshifted and dimmed by the corresponding amount [see equation (5.29)]. In particular, neutral hydrogen atoms which are in the ground energy level will absorb Lyman-series photons from the quasar spectrum. Let us consider just the strongest line of the series: the Lyα transition. The cloud will thus remove photons from the quasar spectrum at a a wavelength of $\sim$ 1216 Å (with a Voigt profile also depending on the gas temperature and turbulent motion in the cloud). This in its rest frame, in the quasar rest-frame the absorption would correspond to a lower wavelength $1216/(1+z_{\rm qc})$. Let us now go back on Earth and observe the cloud-quasar system. The Lyα emission line will be seen at $(1+z_{\rm qso})\cdot 1216$ Å and the Lyα absorption line at $1216\cdot(1+z_{\rm qso})/(1+z_{\rm qc}) = 1216(1+z_{\rm c})$ Å. Since $z_{\rm c} < z_{\rm qso}$ (no absorption would take place otherwise), the absorption line must lie on the blue side of the emission line. The clouds that generate absorption features are generally indicated with the name of *Lyman alpha systems*. The simultaneous detection of higher order lines of the Lyman series confirmed that most of the absorption comes from hydrogen Lyα (see

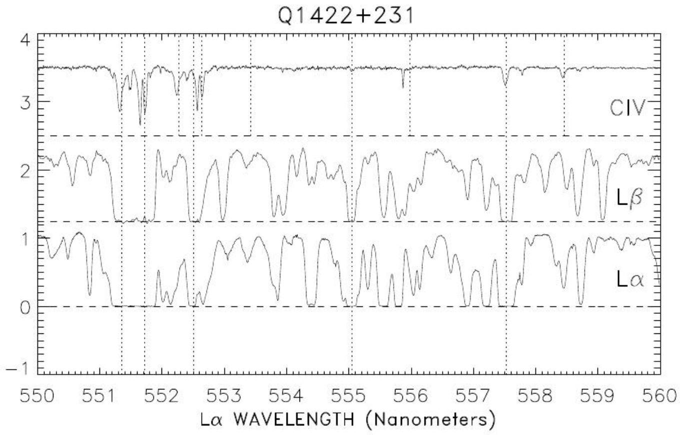


Figure 6.2: A portion of the spectrum of the quasar Q1422+231 (bottom panel) with the corresponding Lyβ (middle) and C$_{\rm IV}$ $\lambda = 1548$ Å (top) shifted in wavelength so that they lie above Lyα if they are originated at the same redshift. Dotted lines mark absorption systems that are saturated both in Lyα and Lyβ.

Figure 6.2).

What about the absorption lines which are observed on the red side of the Lyα in emission? Actually, they can be identified with the strongest resonance lines of the most abundant metals in different ionization states (as like as O$_{\rm I}$, O$_{\rm VI}$, C$_{\rm II}$, C$_{\rm III}$, C$_{\rm IV}$, Mg$_{\rm II}$, Si$_{\rm II}$, Si$_{\rm IV}$, Fe$_{\rm II}$). [1] These *metal line systems* are formed exactly as the Lyα lines but, since the rest wavelength of their transitions is redder, they end up on the red side of the Lyα in emission. Note that only a small fraction ($< 1/50$) of the Lyα absorbers can be associated (i.e. have the same redshift) with metal line transitions (see Figure 6.2). This happens because most of the single Lyα lines arise in clouds where the column density of metals is too low to produce the other absorption lines.

The study of quasar absorption line spectroscopy is the main observational tool at our disposal for detecting and studying the IGM.

6.1.1 Classification of Lyman alpha systems

Lyα systems are usually classified based on the column density of neutral hydrogen.

1. Systems with $10^{12} < N(\mathrm{H_I}) < 10^{17}$ cm^{-2} are known as the *Lyman alpha forest*. Most of the systems with $N(\mathrm{H_I}) > 10^{15}$ cm^{-2} and roughly

[1] We adopt the classical astrophysical notation for ionized elements: O$_{\rm I}$ means neutral oxygen, O$_{\rm II}$ singly-ionized oxygen, O$_{\rm III}$ doubly-ionized oxygen, and so on.

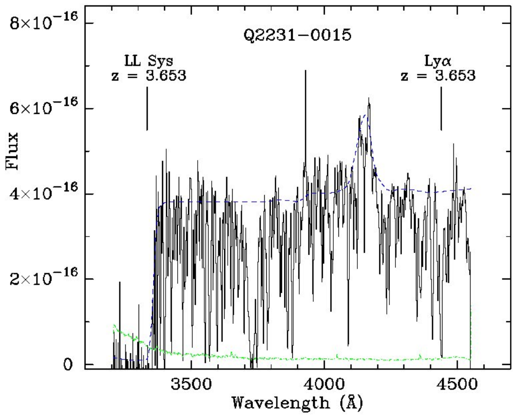


Figure 6.3: A Lyman-limit system simultaneously showing Lyman α absorption and continuum absorption beyond the Lyman limit, $\lambda < 912$ Å.

half of the systems with $N(\mathrm{H_I}) > 3 \times 10^{14}$ cm^{-2} have associated $\mathrm{C_{IV}}$ lines corresponding to a typical metallicity of $Z \sim 10^{-2} Z_\odot$ (rarely objects with $Z \sim 10^{-3} Z_\odot$ have been detected). For lower hydrogen column densities and similar metallicities, metal lines are not detectable with current instruments.

2. Systems with column densities $10^{17} < N(\mathrm{H_I}) < 10^{20}$ cm^{-2} exhibit a conspicuous discontinuity at the Lyman limit (the gas becomes optically thick to ionizing radiation) and are thus called *Lyman-limit systems*. These systems frequently are associated with metal lines of the most abundant chemical elements with several ionization stages. They also have complex velocity structure.

3. At even higher column densities ($N(\mathrm{H_I}) > 10^{20}$ cm^{-2}), the damping wings of the Lyα line become prominent. These systems are always associated with metal absorbers and are dubbed *damped Lyman alpha systems*. Their typical metallicities range between 0.02 and 0.1 $Z_\odot$.

Note that current technology does not allow the detection of systems with $N(\mathrm{H_I}) < 10^{12}$ cm^{-2}.

6.2 Observational properties

6.2.1 Wavelength and redshift ranges

Due to the presence of the atmosphere and the Galaxy, only certain spectral windows are accessible to our telescopes. Spectra with wavelengths $\lambda_{\mathrm{obs}} >$

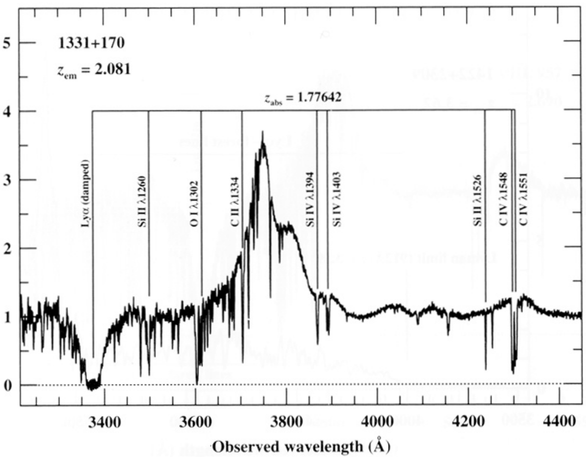


Figure 6.4: A damped-Lyman-alpha system (characterized by prominent damping wings) with its associated metal-line systems.

3200 Å are observable with ground-based telescopes, whereas spectra at shorter wavelengths must be obtained with telescopes above the atmosphere. On the red end, the optical CCDs become transparent to photons at ∼ 9000 Å. For Lyα absorbers, this correspond to the redshift range $1.6 < z_{\rm abs} < 6.4$.

Higher redshift absorbers can be detected in the near infrared but the sky background makes it difficult to detect faint quasars from the ground.

The HST can be used for spectroscopy in the range $1150 < \lambda_{\rm obs} < 3200$ Å (corresponding to $z_{\rm abs} < 1.6$ for Lyα absorbers) where the short wavelength limit arises from the physical properties of the MgFl coatings for the optics. Specialized missions like the *Hopkins Ultraviolet Telescope* (HUT) and the *Far Ultraviolet Spectroscopic Explorer* (FUSE) have been built to get spectra in the $911 < \lambda_{\rm obs} < 1150$ Å range. The Milky Way is opaque between 911 Å and 62 Å (corresponding to a photon energy of 0.2 keV, in the X-ray regime). The *Chandra X-ray observatory* can then be used to study quasar absorption lines at short wavelengths.

6.2.2 Observational challenges

Taking the electromagnetic spectrum of a faint source as like as a high-redshift quasar requires compromising between two competing needs:

1. High spectral resolution is needed to resolve the narrower lines and thus to measure the column density of the absorbing atoms and their Doppler parameter.

2. At the same time, in order to obtain high signal-to-noise spectra, one needs to collect many photons per resolution element. The total number of photons scales with the collecting area of the telescope and the

integration time of a given observation.

The conflict between these limiting factors has characterized the history of absorption-system studies.

The study of absorption systems started in the late 1960s when spectroscopy was done dispersing light of different frequencies on a photographic plate. The typical spectral resolution was of $\Delta\lambda$ =10-20 Å, not enough to resolve most of the absorption lines. A major revolution in the field happened in the 1980s when echelle spectrographs and CCD detectors allowed to take spectra with resolutions as high as $R \sim 50,000$ (corresponding to resolution elements of ~ 0.1 Å). This, however, made quasar spectroscopy very costly in terms of observing time. Taking a high-resolution spectrum with high signal-to-noise ratio required very many nights on a 4m telescope (the standard of the time). It was only with the advent of 10m telescopes (in the 1990s) that high-resolution spectra with signal-to-noise ratios above 100 became common. Recent progress has come from extending the wavelength regime into the ultraviolet band with the HST and its high-resolution spectrographs. This allowed us to study the absorber properties at low redshifts ($z < 1.6$) and the helium Lyα forest at high redshift. At variance with the past, the new limiting factor of current studies is not technology but manpower as it is difficult to keep up with the amount of data streaming of all available telescopes.

6.2.3 Observational techniques

At the dawn of quasar-absorption-systems studies only low-resolution spectrographs were available and the only observable that could be considered was the *flux decrement* (Oke & Korycansky 1982):

$$D(z) = \left\langle 1 - \frac{f_{\rm obs}[(1+z)\cdot 1216\,\text{Å}]}{f_{\rm cont}[(1+z)\cdot 1216\,\text{Å}]} \right\rangle = \langle 1 - e^{-\tau(z)} \rangle = 1 - e^{-\tau_{\rm eff}(z)} \ , \quad (6.2)$$

where $f_{\rm obs}$ is the observed flux, $f_{\rm cont}$ the flux of the unabsorbed continuum and τ is the line optical depth as a function of observed wavelength (or redshift). The average above is done over the lines of sight towards different quasars. The continuum level is usually taken to be a power law in wavelength extrapolated from the region redward of the Lyα line. This is the main source of error in the measure of D or $\tau_{\rm eff}$. Additional errors stem from the absorption contributed by metal lines which often cannot be identified as such.

With the introduction of higher-resolution instruments (in the 1980s), it became possible to distinguish between discrete absorption lines. In this case, the main observables for each line were the equivalent width, W, and the redshift z. Observational data were, however, seriously affected by line blending (which is impossible to correct at medium-low resolution) which

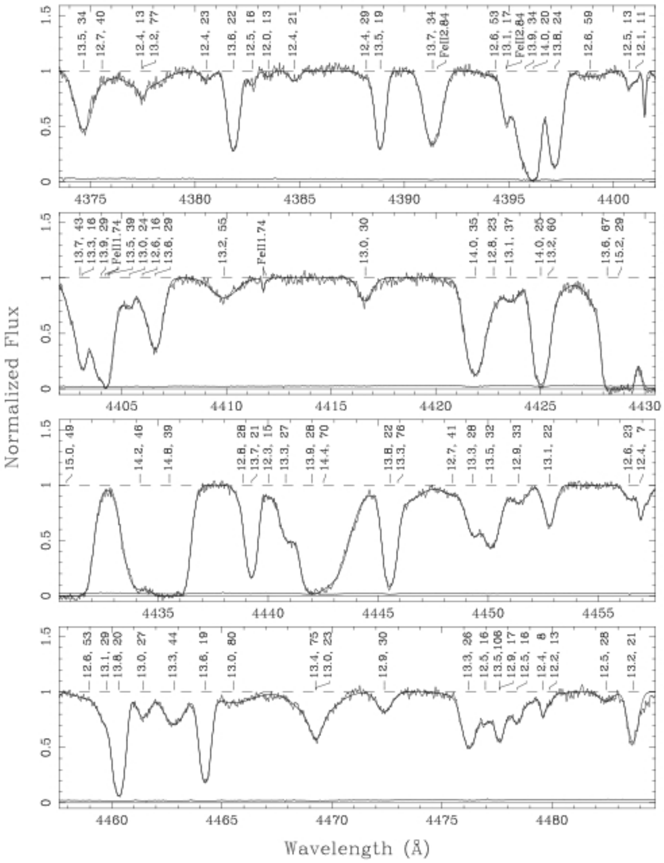


Figure 6.5: Voigt-profile fits to a quasar spectrum. The log column density (in log cm^{-2} and the velocity dispersion (in km s^{-1}) of each Lyα line are reported above a tick mark indicating the centre of the line. Metal transitions are indicated together with the redshift of the metal system. The pixel size is 2 km s^{-1}. The 1σ error of the spectrum is shown with a continuous line just above the zero level.

often made it impossible to use the curve of growth in order to derive the more meaningful physical parameters $N(\mathrm{HI})$ and b.

The field has been revolutionized with the advent of high-resolution spectrographs. It has been shown that real absorption lines are reasonably well approximated by Voigt profiles. In this case, the basic observable of each absorption feature are the Doppler parameter b, the column density $N(\mathrm{HI})$ and the redshift z. These are usually determined by Voigt-profile fitting as follows.

1. Because of some peculiarities of echelle spectrographs, the quasar continuum is estimated locally from polynomial fits to spectral regions deemed free of absorption. This tends to underestimate the continuum and it is the main drawback of the method.

2. Then the whole absorption pattern is fitted with a linear superposition of Voigt profiles. In order to deblend complex lines (made of multiple components overlapping in wavelength), additional Voigt profiles are added until the residuals from the fitted function become compatible with random fluctuations.

3. For the stronger Lyα lines, additional constraints are usually obtained by simultaneously fitting the lines from higher order transitions of the Lyman series.

4. Eventually, a catalog of absorption lines is produced listing their redshifts, equivalent widths, column densities and Doppler parameters. The catalog thus constitutes the basis for any subsequent analysis.

An open question is whether the number of independent components necessary to fit a line converges with increasing the signal-to-noise ratio of the spectrum.

Also note that density and velocity gradients in the gas responsible of the absorption can create line profiles which depart from the Voigt one. Rotating, collapsing or expanding clouds can generate a wide variety of profiles that encode information on the dynamical state of the gas.

The alternative to Voigt-profile fitting is measuring the optical depth of each line on a pixel-by-pixel basis. This keeps much more information than a Voigt-profile fit but makes comparison with models more complicated.

6.2.4 Mean absorption

Measurements of the mean absorption show a strong redshift evolution (see Figure 6.6). Parameterizing the distribution as

$$\tau_{\mathrm{eff}} = A(1+z)^{\gamma+1} \tag{6.3}$$

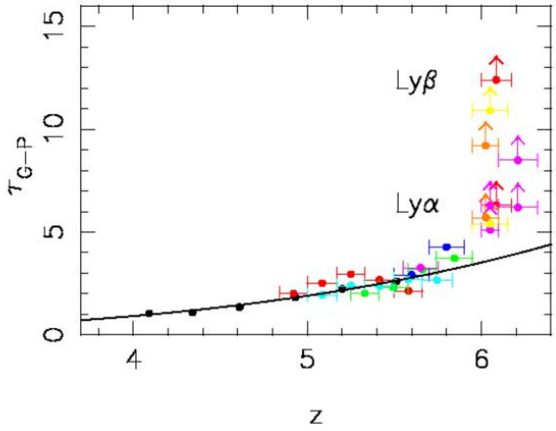


Figure 6.6: Lyα effective optical depth inferred from Lyα and Lyβ absorption.

and using 29 low-resolution spectra, Press et al. (1993) found $\gamma = 2.46 \pm 0.37$ and $A = 0.0175 - 0.0056\gamma \pm 0.0002$ for $2.5 \leq z \leq 4.3$. Recent re-analyses based on larger datasets (thousands of spectra from the Sloan Digital Sky Survey) and more sophisticated techniques for continuum fitting and line blending give steeper slopes, $\tau \simeq 0.0018\,(1+z)^{3.92}$, in the redshift range $z < z < 4$ (Bernardi et al. 2003; McDonald et al. 2006; Faucher-Giguère et al. 2009). The most recent data show an upturn at redshift $z \sim 6$ (Figure 6.6). We will see that has important implications for the history of the intergalactic medium.

6.2.5 Abundance: line counting

The observations can be fit with a function

$$\frac{d^2N}{dW\,dz} = \frac{W}{W_*}\,e^{-W/W_*}\,(1+z)^{\gamma}\,. \tag{6.4}$$

Integrating above a given equivalent width, one then has

$$\frac{dN}{dz} = N_0\,(1+z)^{\gamma}\,. \tag{6.5}$$

Typically observed values of the parameters at $z \sim 3$ are $1.5 < \gamma < 3.0$ and $W_* \simeq 0.27$ Å. The redshift evolution flattens at $z < 1.5$ (see Figure 6.7) where $\gamma \sim 0.5$ and there is some indication that it steepens for $z > 4$.

Note that the number of absorption lines (above a given equivalent width) per unit redshift can be written as

$$\frac{dN}{dz} = n_0(z)\,\sigma(z)\,(1+z)^2\,\frac{dr_{\rm com}}{dz} \tag{6.6}$$

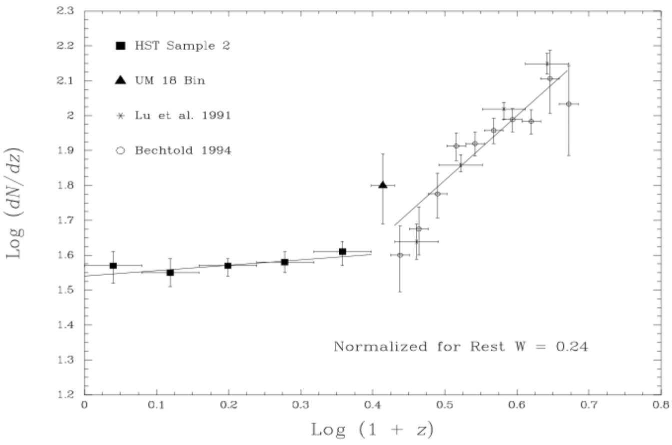


Figure 6.7: Evolution of the Lyα forest as observed from space (filled squares) and from the ground (all the other points). Two power-law fits with slopes of $\gamma = 0.5$ and 1.85 are shown.

where $n_0(z)$ is the comoving density of absorbers, $\sigma(z)$ is the proper geometric cross-section for absorption (i.e. the area of the cloud over which an absorption feature of the required equivalent width is created) and $r_{\rm com}$ is the comoving distance along the line of sight to the quasar.[2] In our summary of the FRW cosmology, we have shown that in a flat universe dominated by matter plus a cosmological constant,

$$\frac{dr_{\rm com}}{dz} = \frac{c}{H_0\,[\Omega_{\rm m}(1+z)^3 + \Omega_\Lambda]^{1/2}}\ . \tag{6.7}$$

Therefore interpreting the redshift evolution of dN/dz is not straightforward as this number encodes information on the abundance and size of the absorbers plus on the cosmological model. In order to isolate the intrinsic evolution of the absorbers by removing the cosmological redshift dependence it is often convenient to introduce the "redshift path"

$$X(z) = \int_0^z \frac{(1+z')^2}{[\Omega_{\rm m}(1+z')^3 + \Omega_\Lambda]^{1/2}}\, dz' \tag{6.8}$$

[2] Remember that comoving lengths are a factor $1+z$ longer than proper lengths. Therefore the factor $(1+z)^2$ can be seen:

- either as the conversion coefficient between the proper cross section and the comoving one;
- or as the result of the product between the proper density, $(1+z)^3\, n_0(z)$ and the proper length $r_{\rm com}/(1+z)$.

In perfect analogy with the radiative transfer case, the product $[n_0(z)\,\sigma(z)\,(1+z)^2]^{-1}$ gives the comoving mean free path to absorption.

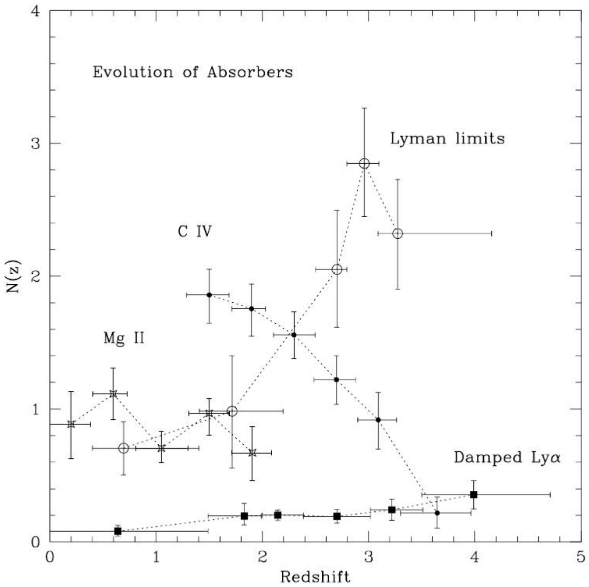


Figure 6.8: Number of absorbers per unit redshift for LLs, DLAs and some metal absorbers.

so that

$$\frac{dN}{dX} = \frac{c}{H_0} n_0(z) \, \sigma(z) \; . \tag{6.9}$$

Note, however, that to compute dN/dX out of some data one has to assume a cosmology while dN/dz is calculated only using observed quantities.

The low exponent, $\gamma \sim 0.5$, for the evolution of the Lyα forest at $z < 1$ suggests very little evolution in the population of absorbers. On the other hand, the rapid upturn at higher redshifts points towards an evolving population of absorbers.

The degree of evolution of the Lyα absorbers depends on their column density (see Figure 6.8). In particular, the abundance of damped systems shows very little evolution from $z = 0$ to $z = 5$ thus suggesting that they are of a different nature than the forest.

6.2.6 Line-width distribution

The distribution of Doppler parameters at $z \sim 3$ is well approximated by a Gaussian with a mean of 30 km s^{-1} and rms value of 8 km s^{-1}, truncated below a cutoff of 15 km s^{-1} (see Figure 6.9).

The same functional form describes the distribution of Doppler parameters at other redshifts. However it appears that there is a trend towards lower values of b at higher redshifts. For instance, the median Doppler parameter passes from 41 km s^{-1} at $z \sim 2.3$ to 31 km s^{-1} at $z \sim 3.7$.

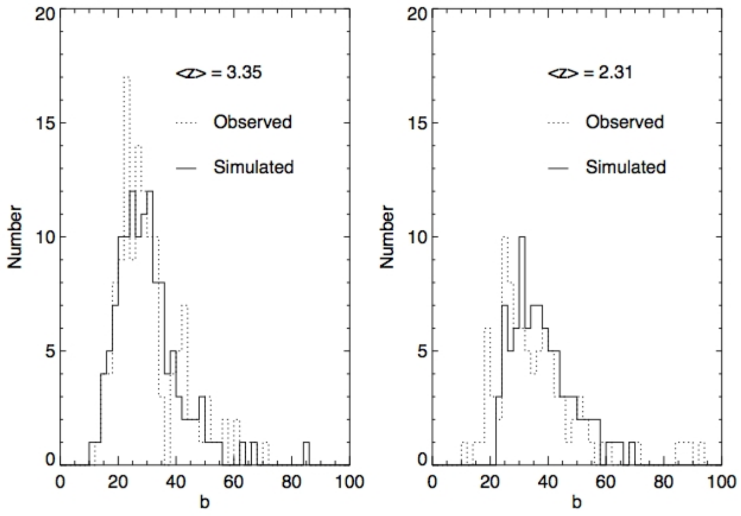


Figure 6.9: The distribution of b-values of the Lyα forest derved from Voigt-profile fitting.

Note that the b-distributions must be corrected slightly for blending using simulated spectra.

6.2.7 Column density distribution

The number of absorbers per unit $\mathrm{H_I}$ column density can be parameterized as

$$\frac{dN}{dN(\mathrm{H_I})} \propto N(\mathrm{H_I})^{-\beta} \, , \tag{6.10}$$

with $\beta \sim 1.5$. This law approximately extends over ten orders of magnitude in column density from 10^{12} cm^{-2} to 10^{22} cm^{-2} (see Fig. 6.10). Observational results are often reported in terms of the function

$$f = \frac{\Delta N}{\Delta N(\mathrm{H_I}) \sum_{i=1}^{N_{\mathrm{QSO}}} \Delta X_i} \, , \tag{6.11}$$

where ΔN is the observed number of systems with column density between $N(\mathrm{H_I})$ and $N(\mathrm{H_I}) + \Delta N(\mathrm{H_I})$ and ΔX the sum of the observed "redshift paths" towards each quasar.

To higher accuracy, there is strong evidence for departures from a pure power-law scaling. A steepening of the counts appears to be present at $N(\mathrm{H_I}) \sim 10^{14}$ cm^{-2} (where $\beta \sim 1.8$) and a flattening ($\beta \sim 1.3$) is manifest for the damped systems. This means that damped systems are more abundant than expected by extrapolating the power-law fit of the lower column densities.

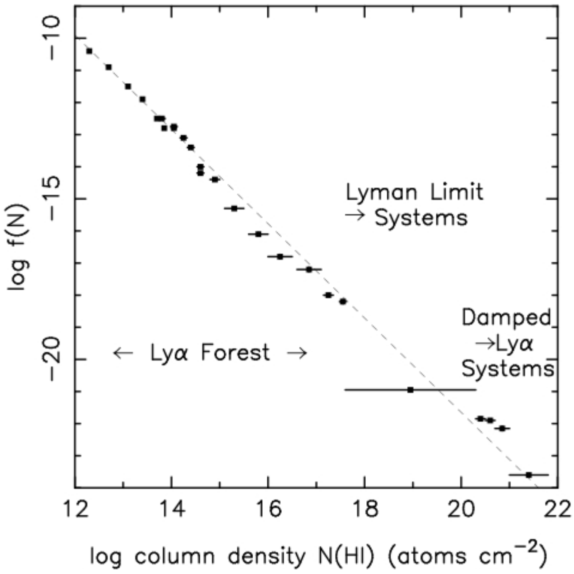


Figure 6.10: The column density distribution of neutral hydrogen in quasar absorption systems. The dashed line is a power law with exponent -1.46 which fits the data reasonably well over 10 orders of magnitude in column density.

6.2.8 Clustering

A number of statistics have been used to demonstrate that the distribution of absorbers in velocity space is not random (i.e. Poissonian) and shows some degree of clustering. The most commonly used one is the two-point correlation function along the line of sight, $\xi(\Delta v)$. This is defined as the excess probability (with respect to random) of finding a pair of clouds separated by a velocity interval Δv

$$\Delta p = n_0\,\sigma\,\Delta v\,[1+\xi(\Delta v)]\;. \tag{6.12}$$

Observational data indicate that there is weak small-scale clustering ($\xi \sim 1$ at ~ 100 km s^{-1}) in the forest at $z \sim 3$ (results might be affected by line blending, though). Anyway, the clustering amplitude seems to increase with $N(\mathrm{H_I})$. On the other hand metal absorption systems are found to be strongly clustered at velocity separations of a few hundred km s^{-1} (see Figure 6.11).

The presence of voids (large regions with no absorption) has also been used as a measure of clustering. Results show that Lyα absorption does not present void regions as large as those apparent in the galaxy population. However, some individual large gaps extending for a few tens of Mpc have been found.

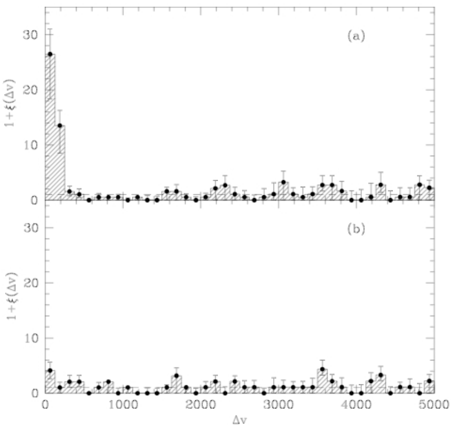


Figure 6.11: Two-point correlation function of quasar absorption systems at $z \sim 2.5$. The top panel refers to $\mathrm{C_{IV}}$ absorbers and the bottom one to Lyα absorbers.

6.2.9 Characteristic sizes

One of the main shortcomings of high-redshift quasar spectroscopy is the lack of multi-dimensional information about the absorbers. This fact makes it hard to understand their geometry and to disentangle velocity effects. The only information on the size and geometry of the absorbers can be retrieved using the observation of common absorption systems along multiple, closely separated lines of sight. Two cases must be distinguished.

Gravitationally lensed QSOs. One can use the multiple images of a gravitationally lensed quasar to probe different lines of sight. The typical angular separation between the multiple images is of a few arcsec which, at the redshift of the absorbers, correspond to less than 100 kpc. Generally all the lines seen in one image are seen in the other with very strongly correlated equivalent widths.

Close QSO pairs. Close quasar pairs have typical separations on the sky ranging from 10 arcsec to a few arcmin. Common absorption lines are seen only for the quasar pairs with the smallest separations.

This indicates that Lyα absorbers have characteristic sizes of 200-500 h^{-1} kpc.

6.3 Counterparts in emission

Deep imaging surveys in the optical waveband have been undertaken in order to identify possible counterparts in emission to the absorbers. With

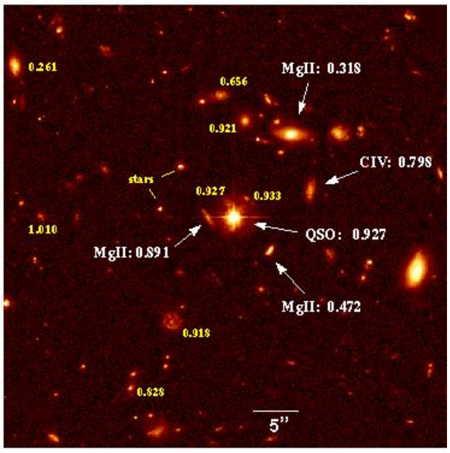


Figure 6.12: Deep HST image of the field around the quasar 3C 336 (at $z = 0.927$) obtained with a 24,000 s exposure. The labels indicate the redshift of all galaxies for which it has been measured. Bold numbers mark galaxies that give rise to metal-line absorption systems in the quasar spectrum.

few exceptions, galaxies have been found at the redshift of intervening $\mathrm{Mg_{II}}$ absorbers thus suggesting that strong metal lines are generated by interstellar clouds in intervening galaxies. Overall it appears that low-redshift $\mathrm{Mg_{II}}$ absorbers lie in normal galaxies (some in the halos of disk galaxies). Some galaxies do not produce absorption in background quasars at all. Similar conclusions can be drawn for other metal absorbers. What changes is the characteristic cross section for absorption around a galaxy (see e.g. Figure 6.13).

Deep galaxy redshift surveys in the optical typically find galaxies only up to $z \sim 1.5 - 2$ (color selection – like the Lyman-break technique – is normally employed to target special classes of galaxies at higher redshift) so galactic counterparts of Lyα absorbers have to be searched differently. Infrared surveys identified counterparts of damped Lyα absorbers via direct imaging and Hα emission. In brief, they appear to be associated with small, star-forming galaxies that constitute the bulk of the galaxy population at high redshift.

While high-column-density absorbers appear to be associated with galaxies, no such a correspondence is seen for the Lyα forest.

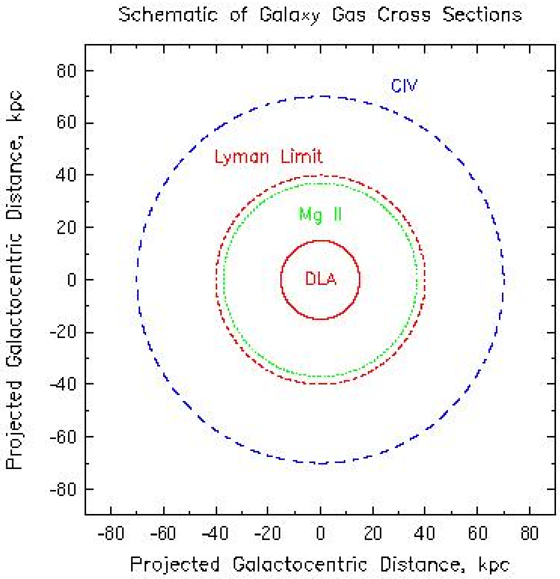


Figure 6.13: The inferred sky-projected cross sections of damped Lyα absorbers (DLAs), $\mathrm{Mg_{II}}$ absorbers, Lyman limit systems, and $\mathrm{C_{IV}}$ systems are shown (in kpc).

Chapter 7

Ionization and recombination

7.1 The Gunn-Peterson effect

In 1965 Gunn and Peterson pointed out that any generally distributed neutral hydrogen would produce a broad depression in the spectrum of high-redshift quasars at wavelengths shortward of 1216 Å. Since such a depression is not seen we conclude that the intergalactic medium must be mostly ionized.

The argument proceeds as follows. Let us compute the optical depth for Lyα absorption due to a smoothly distributed "sea" of neutral hydrogen in the expanding universe. At the observed frequency ν, this reads:

$$\tau(\nu) = \int_0^{z_\mathrm{s}} \sigma_{\mathrm{Ly}_\alpha}[\nu(1+z)]\, n_{\mathrm{H_I}}(z)\, \frac{dr_\mathrm{prop}}{dz}(z)\, dz \ , \tag{7.1}$$

where z_s denotes the redshift of the light source against which absorption is detected and

$$\sigma_{\mathrm{Ly}_\alpha}(\nu) = \frac{\pi e^2}{m_\mathrm{e} c}\, f\, \phi(\nu) \ , \tag{7.2}$$

is the cross section for Lyα absorption (neglecting stimulated emission) with $f = 0.4162$ the oscillator strength. As the cross-section is sharply peaked around the central frequency and its broadness is negligible in redshift units, we can replace the line profile ϕ with the Dirac delta function $\delta_\mathrm{D}[\nu(1+z) - \nu_{\mathrm{Ly}_\alpha}]$ and obtain (for $\nu > \nu_{\mathrm{Ly}_\alpha}$):

$$\begin{aligned} \tau(\nu) &= \frac{\pi e^2}{m_\mathrm{e} c}\, \frac{f}{\nu}\, n_{\mathrm{H_I}}(\tilde{z})\, \frac{dr_\mathrm{prop}}{dz}(\tilde{z}) = &(7.3)\\ &= \frac{\pi e^2}{m_\mathrm{e} c}\, \frac{f}{\nu_{\mathrm{Ly}_\alpha}}\, (1+\tilde{z})\, n_{\mathrm{H_I}}(\tilde{z})\, \frac{dr_\mathrm{prop}}{dz}(\tilde{z}) \ , &(7.4) \end{aligned}$$

with $1 + \tilde{z} = \nu_{\mathrm{Ly}_\alpha}/\nu$ (i.e. only one specific redshift contributes for each observed frequency).

In a flat universe dominated by matter and cosmological constant,

$$\frac{dr_{\rm prop}}{dz}(z) = \frac{c}{H_0}\,\frac{1}{1+z}\,\frac{1}{\sqrt{\Omega_{\rm M}(1+z)^3+\Omega_\Lambda}}\ , \tag{7.5}$$

so that

$$\tau(\nu) = \frac{\pi e^2}{m_{\rm e}c}\,\frac{f}{\nu_{{\rm Ly}_\alpha}}\,\frac{c}{H_0}\,\frac{n_{\rm H_I}(\tilde{z})}{\sqrt{\Omega_{\rm M}(1+\tilde{z})^3+\Omega_\Lambda}}\ . \tag{7.6}$$

Introducing the hydrogen neutral fraction,

$$f_{\rm neut}(z) = \frac{n_{\rm H_I}(z)}{n_{\rm H}(z)}\ , \tag{7.7}$$

and keeping into account the expansion of the universe,

$$n_{\rm H}(z) = n_{\rm H}(0)\,(1+z)^3\ , \tag{7.8}$$

one obtains

$$\tau(\nu) = \frac{\pi e^2}{m_{\rm e}c}\,\frac{f}{\nu_{{\rm Ly}_\alpha}}\,\frac{c}{H_0}\,n_{\rm H}(0)\,\frac{f_{\rm neut}(\tilde{z})\,(1+\tilde{z})^3}{\sqrt{\Omega_{\rm M}(1+\tilde{z})^3+\Omega_\Lambda}} \tag{7.9}$$

with

$$\begin{aligned} n_{\rm H}(0) = 0.76\,\frac{\Omega_{\rm b}\rho_{\rm crit}(0)}{m_{\rm p}} &\simeq 0.188\left(\frac{\Omega_{\rm b}h^2}{0.022}\right)\,{\rm m}^{-3} = \\ &= 1.88\times10^{-7}\left(\frac{\Omega_{\rm b}h^2}{0.022}\right)\,{\rm cm}^{-3} \end{aligned} \tag{7.10}$$

where 0.76 is the mass fraction in hydrogen atoms from primordial nucleosynthesis. Replacing all the constants (remember that $e^2/(m_{\rm e}c^2) \simeq 2.818\times10^{-13}$ cm is the classical electron radius), one eventually gets

$$\frac{\pi e^2}{m_{\rm e}c}\,\frac{f}{\nu_{{\rm Ly}_\alpha}} = 4.472\times10^{-18}\ {\rm cm}^2\ , \tag{7.11}$$

and

$$\tau(\nu) = 7777\,h^{-1}\left(\frac{\Omega_{\rm b}h^2}{0.022}\right)\,\frac{f_{\rm neut}(z)\,(1+z)^3}{\sqrt{\Omega_{\rm M}(1+z)^3+\Omega_\Lambda}}\ . \tag{7.12}$$

The equation applies to all parts of the source spectrum to the blue of the Lyα emission line. If $f_{\rm neut} \sim 1$ then $\tau \gg 1$ at all observable frequencies (i.e. redshifts), and an absorption trough should be detected in the level of the rest frame UV continuum of quasars. This is called the Gunn-Peterson effect. Current upper limits at $z \simeq 5$ are $\tau < 0.1$ and this implies $f_{\rm neut}(z = 5) < 4.3\times10^{-7}\,h$. Even assuming that 99% of cosmic hydrogen is in the Lyα forest (or in galaxies), with only 1% in a smoothly distributed component, still $f_{\rm neut}(z = 5) < 4.3\times10^{-5}\,h$. In summary: since we see the quasar continuum between the discrete absorption lines in a quasar spectrum the smooth component of the IGM must be highly ionized.

7.2 Bound-free transitions

The lack of Gunn-Peterson troughs in quasar spectra at $z < 5$ shows that ionization processes are very important in the IGM. For this reason, we will spend some time to describe how we can model them. Let us start from the *photoionization* process, where the absorption of a photon results in an electron being liberated from the atom. Since the electron passes from an initial bound state to a final free state, this is also known as a bound-free transition. Contrary to their bound-bound counterparts, bound-free transitions are not sharply defined in energy since the ionized electron can have anything from zero energy (if it was barely ionized) to a large energy (if it was ejected from the atom with a large velocity). There is, however, a minimum photon energy required to ionize the atom (the *ionization potential*, χ). Therefore the characteristic absorption coefficient for bound-free transitions is an *edge*: no absorption below some energy, then a sharp onset in the absorption above that critical energy.[1]

The differential quantum mechanical transition rate (per unit direction and per unit modulus of the momentum, $\mathbf{p}$, of the emerging electron) for ionizing an atom with radiation of specific intensity $I(\omega)$ is

$$\frac{d^2w}{dp\,d\Omega} = \frac{4\pi^2 e^2}{m_\mathrm{e}^2 c}\,\frac{I(\omega)}{\omega^2}\,|\langle f|\exp{(i\mathbf{k}\cdot\mathbf{r})}\,\mathbf{u}\cdot\nabla\,|i\rangle|^2\,\frac{d^2n}{dp\,d\Omega} \tag{7.13}$$

where $d^2n/(dp\,d\Omega)$ is the density of available states in the continuum and $\hbar\,d\omega = p\,dp/m_\mathrm{e}$ to ensure energy conservation. Note that here i denotes a bound state and f is a continuum state for the electron. Similarly to the bound-bound case, we can express the transition rate in terms of a total bound-free cross section, σ_bf, which is obtained by summing up over all possible velocities of the emerging electron. For hydrogenoid atoms with atomic number Z, ionized from the level of quantum numbers (n,ℓ), the cross section can be written as:

$$\sigma_\mathrm{bf}(n,\ell) = \frac{512\,\pi^7 m_\mathrm{e} e^{10} Z^4}{3^{3/2} c\,h^6 n^5}\,\frac{g(\omega,n,\ell,Z)}{\omega^3} \tag{7.14}$$

where g is the bound-free Gaunt factor (which is unity within 20% near the ionization threshold). Note that σ_bf is zero for $\omega < \omega_n = \chi_n/\hbar = \alpha^2 m_\mathrm{e} c^2 Z^2/(2\hbar n^2)$ with χ_n the ionization potential from the bound level n.

The absorption coefficient of the radiative transfer equation due to ionization is

$$\alpha_\nu = \sum_n n_n\,\sigma_\mathrm{bf}(n)\ , \tag{7.15}$$

where n_n is the number density of atoms at the absorbing level. The absorption coefficient thus presents a series of absorption edges corresponding to

[1] This is the reason why Lyman-limit systems present an absorption edge in quasar spectra.

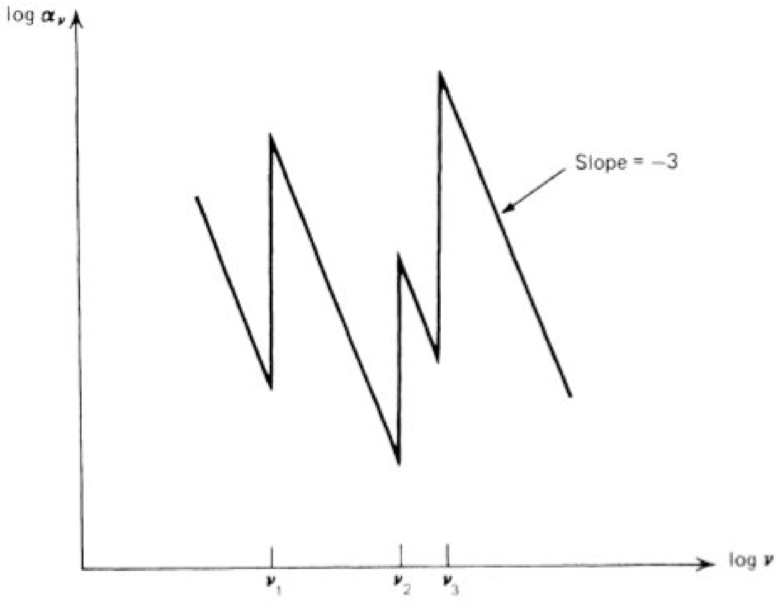


Figure 7.1: Schematic illustration of the frequency dependence of the absorption coefficient due to bound-free transitions. The sharp rises (absorption edges) occur at the photoionization threshold of a particular atomic level.

the ionization potentials from the different levels (see, however, Section 7.4 for a simplification valid for the IGM). The relative strength of the edges depends on the number of atoms in each level and on the quantum-mechanical cross sections. On the other hand their location in frequency only depends on atomic physics.

7.2.1 Other sources of ionization

Beyond photoionization there are other physical processes that can ionize atoms and ions.

1. Collisional ionization is the ionization of an atom induced by the collision with an energetic particle (typically an electron). On energetic grounds, it is normally the outermost electron which is removed. The rate of collisional ionization per unit volume, per unit time from the energy level n may be written as $n_{\rm e}\, n_{\rm i}\, q_n({\rm X}, T)$ where the coefficients $q_n({\rm X}, T)$ are available in tabulated form for the most important atomic species, X. Collisional ionizations are normally unimportant in the IGM (densities are too low).

2. Autoionization (also known as Auger effect) is a process in which an

atom with a vacant electron in an inner shell spontaneously readjusts itself to a more stable state by ejecting one or more electrons instead of radiating a photon. In the IGM this process is important to determine the relative abundances of highly ionized metals.

7.3 Radiative recombination

The inverse process to photoionization is *radiative recombination*, in which an electron is captured by a ion into a bound state with emission of a photon. The number of recombinations per unit time per unit volume can be written as

$$n_+ \, n_\mathrm{e} \, \sigma_\mathrm{fb} \, f(v) \, v \, dv \tag{7.16}$$

where n_+ is the ion density, n_e the electron density, v the electron speed and $f(v)$ its probability distribution. There are connections between the cross sections for photoionization and radiative recombination, analogous to the relations between the Einstein coefficients. Detailed balancing gives the so-called Milne relation:

$$\frac{\sigma_\mathrm{bf}}{\sigma_\mathrm{fb}} = \frac{m_\mathrm{e}^2 c^2 v^2}{\nu^2 h^2} \, \frac{g_\mathrm{e} \, g_+}{2 \, g_\mathrm{n}} \tag{7.17}$$

with $h\nu = (1/2) m_\mathrm{e} v^2 + \chi$ and where the g coefficients are the quantum degeneracy factors of electrons, ions and of the recombined atom.

The recombination coefficient at a particular quantum level n is defined as

$$\alpha_n = \langle v \, \sigma_\mathrm{fb}(n) \rangle = \int v \, f(v) \, \sigma_\mathrm{fb} \, dv \tag{7.18}$$

and requires knowledge of the velocity distribution of the electrons. The recombination rate at the level n is therefore $n_+ \, n_\mathrm{e} \, \alpha_n$. If the velocity distribution is thermal, then α_n is only a function of the electron temperature T: $\alpha_n(T)$. These coefficients are available in tabulated form for the most important elements. The total recombination coefficient for the ion X is usually indicated with the symbol

$$\alpha_\mathrm{A}(\mathrm{X}, T) = \sum_n \alpha_n(\mathrm{X}, T) \; . \tag{7.19}$$

For hydrogen at $T \sim 10^4$ K, $\alpha_\mathrm{A} = 4.18 \times 10^{-13}$ cm^3 s^{-1} and roughly scales as $T^{-1/2}$.

In a pure hydrogen gas cloud at temperature T with no sources of ionization, the number density of ionized atoms n_p follows

$$\frac{1}{n_\mathrm{p}} \frac{dn_\mathrm{p}}{dt} = -n_\mathrm{e} \, \alpha_\mathrm{A}(\mathrm{H_I}, T) \tag{7.20}$$

so that we can define the *recombination timescale* as

$$t_{\rm rec} = \frac{1}{n_{\rm e}\alpha_{\rm A}({\rm H_I}, T)} \,, \tag{7.21}$$

which, for $n_{\rm e} \sim 10^{-5}$ cm^{-3} (corresponding to the mean proper cosmic density of hydrogen at $z \sim 3$), gives $t_{\rm rec} \sim 7.6 \times 10^9$ yr at $T = 10^4$ K. Out of ionization equilibrium, the recombination timescale is generally longer compared with the ionization timescale and with the electron thermalization timescale (see below for the precise definition of these quantities).

7.3.1 Other recombination processes

There are other physical processes that lead to the recombination of ions and electrons beyond radiative recombination.

1. Three-body recombination is the reverse process of collisional ionization, and it is unimportant in low-density astrophysical plasmas because its rate is proportional to the square of the density.

2. Dielectronic recombination is the inverse process of autoionization, and it takes place when the captured electron excites an inner core electron. The excited atom then relaxes via a two-step process: one of the valence electrons radiatively de-excites, then the atom radiatively cascades like in radiative recombination. Dielectronic recombination is important for metals in two temperature regimes: at relatively low temperature ($T \sim 1,000 - 3,000$ K) and at very high temperatures ($T > 20,000$ K).

7.4 The nebular approximation

When an electron recombines with a proton, it can end up in a highly excited bound state, followed by a radiative cascade into the ground level of the hydrogen atom. The lifetime of the excited levels (which depends on the Einstein coefficients A_{ij}) is typically of the order of 10^{-4} s and reaches 0.12 seconds for the metastable $2s\,{}^2S$ state. These timescales are negligibly small with the typical photoionization timescale (typically, at least 10^{12} s, see below). Thus, to extremely good approximation, we may consider that all hydrogen atoms are in the ground level at the moment of ionization. This is known as the *nebular approximation* and greatly simplifies calculations of physical conditions in the IGM. It also implies that, when ionization equilibrium is established, photoionization from the ground level is balanced by recombinations to all levels as each recombination to any excited level is followed very quickly by radiative transitions downward, leading ultimately to the ground level.

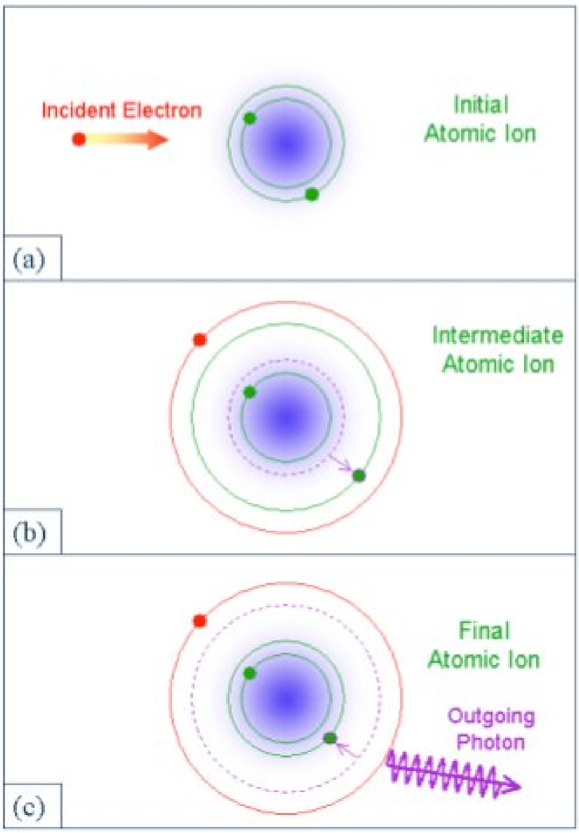


Figure 7.2: Cartoon showing the process of dielectronic recombination. This process is important in determining the elemental abundances of cosmic gas clouds that are photoionized by very energetic ultraviolet light. Only recently have accurate dielectronic-recombination rates from theory and laboratory experiments become available.

Similar reasoning applies to atomic species different from hydrogen. From now on we will indicate the photoionization cross-section from the ground level with the symbol $a_\nu(\mathrm{X})$ where X denotes a given ion or atom. This quantity is available in tabulated form. For hydrogen $a_\nu(\mathrm{H_I}) = 6.3 \times 10^{-18}$ cm^2 at the ionization threshold and approximately decreases as ν^{-3} (see Figure 7.3).

7.4.1 Ionization and recombination rates

The discussion above makes it possible to compute the ionization structure of a pure hydrogen cloud. The photoionization rate per unit volume (i.e. the number of ionization events per unit volume per unit time) is:

$$n_{\mathrm{H_I}} \int_{\nu_0}^{\infty} 4\pi \,\frac{J_\nu}{h\nu}\, a_\nu(\mathrm{H_I})\, d\nu \tag{7.22}$$

(where $\nu_0 = \chi/h$ is the frequency corresponding to the hydrogen ionization potential of 13.6 eV) and we can define a *photoionization timescale* as

$$t_{\mathrm{ion}} = \Gamma_{\mathrm{ion}}^{-1} = \left(\int_{\nu_0}^{\infty} 4\pi \,\frac{J_\nu}{h\nu}\, a_\nu(\mathrm{H_I})\, d\nu\right)^{-1} \tag{7.23}$$

(Γ_{ion} is the photoionization rate per unit $\mathrm{H_I}$ atom, we will see in the following classes that typical values for the IGM range between $10^{-14} < \Gamma_{\mathrm{ion}} < 10^{-12}$ s^{-1}). Note that both quantities are fully determined by the spectral shape and amplitude of the radiation intensity at frequencies above ν_0. On the other hand, in thermal equilibrium, the recombination rate is:

$$n_{\mathrm{p}}\, n_{\mathrm{e}}\, \alpha_{\mathrm{A}}(\mathrm{H_I}, T)\ . \tag{7.24}$$

Putting everything together, one obtains the equation for the ionization balance:

$$\frac{dn_{\mathrm{p}}}{dt} = n_{\mathrm{H_I}} \int_{\nu_0}^{\infty} 4\pi \,\frac{J_\nu}{h\nu}\, a_\nu(\mathrm{H_I})\, d\nu - n_{\mathrm{p}}\, n_{\mathrm{e}}\, \alpha_{\mathrm{A}}(\mathrm{H_I}, T)\ . \tag{7.25}$$

Note that this equation couples the radiation field at frequencies $\nu > \nu_0$ with the density of the ionized species. The ionization equilibrium is reached when the right-hand-side equals zero. In this case, $t_{\mathrm{ion}} = t_{\mathrm{rec}}$.

7.4.2 Optically thick clouds

Radiative recombinations to the ground level produce photons with energies $E > 13.6$ eV that can potentially ionize hydrogen atoms. The emission coefficient for this radiation is

$$j_\nu = \frac{2h\nu^3}{c^2}\left(\frac{h^2}{2\pi m_{\mathrm{e}} k_{\mathrm{B}} T}\right)^{3/2} a_\nu\, \exp\left[-\frac{h(\nu-\nu_0)}{k_{\mathrm{B}} T}\right] n_{\mathrm{p}}\, n_{\mathrm{e}}\ , \tag{7.26}$$

(for $\nu > \nu_0$ and zero otherwise) which is strongly peaked around $\nu = \nu_0$. Therefore, diffuse radiation generated by recombinations within the cloud contributes to the specific intensity of radiation J_ν and we can distinguish it from that coming from external sources by writing $J_\nu = J_\nu^{(\mathrm{ext})} + J_\nu^{(\mathrm{diff})}$. For an optically thin gas cloud (at $\nu = \nu_0$), a good first-order approximation is to take $J_\nu^{(\mathrm{diff})} = 0$ as the diffuse photons will likely escape the cloud. On the other hand, in an optically thick cloud, diffuse photons will not be able to escape and will lead to further ionization. In this case, every photon of the diffuse radiation field is absorbed elsewhere in the gas cloud:

$$4\pi \int_V \frac{j_\nu}{h\nu}\, dV = \int_V n_{H_I} \frac{a_\nu J_\nu^{(\mathrm{diff})}}{h\nu}\, dV \ , \tag{7.27}$$

where the integration is over the volume of the cloud. In the most extreme cases (where the medium is very optically thick), the photons will be absorbed very close to the point where they have been generated and it makes sense to assume that

$$J_\nu^{(\mathrm{diff})} = \frac{j_\nu}{n_{H_I}\, a_\nu} \ , \tag{7.28}$$

which automatically satisfies eq. (7.27). This is known as the "on-the-spot approximation" and holds true when the mean free path of the diffuse photons is very small. Since the total number of photons generated by recombinations to the ground level is $n_\mathrm{p}\, n_\mathrm{e}\, \alpha_1(\mathrm{H_I}, T)$ (at thermal equilibrium), we can thus write:

$$\frac{dn_\mathrm{p}}{dt} = n_{\mathrm{H_I}} \int_{\nu_0}^{\infty} 4\pi \frac{J_\nu^{(\mathrm{ext})}}{h\nu}\, a_\nu(\mathrm{H_I})\, d\nu - n_\mathrm{p}\, n_\mathrm{e}\, \alpha_\mathrm{B}(\mathrm{H_I}, T) \ , \tag{7.29}$$

where

$$\alpha_\mathrm{B}(\mathrm{H_I}, T) = \alpha_\mathrm{A}(\mathrm{H_I}, T) - \alpha_1(\mathrm{H_I}, T) = \sum_{n=2}^{\infty} \alpha_n(\mathrm{H_I}, T) \ . \tag{7.30}$$

The physical meaning is that in optically thick clouds, the ionizations caused by external sources are balanced by recombinations to excited levels of H, while recombinations to the ground level generate ionizing photons that are quickly re-absorbed and have no effect on the overall ionization balance. Thus, using the on-the-spot approximation, the only difference between the optically thick and thin cases lies in the value of the recombination coefficient ($\alpha_\mathrm{B}(\mathrm{H_I}, T = 10^4\ \mathrm{K}) = 2.59 \times 10^{-13}\ \mathrm{cm^{-3}\ s^{-1}}$). Note, however, that this is only an approximation, the exact result can only be found by numerically solving the radiative transfer problem within the cloud.

7.4.3 Helium ionization

So far we have considered pure hydrogen. Helium is the second most abundant element after hydrogen, with typical abundances by number of

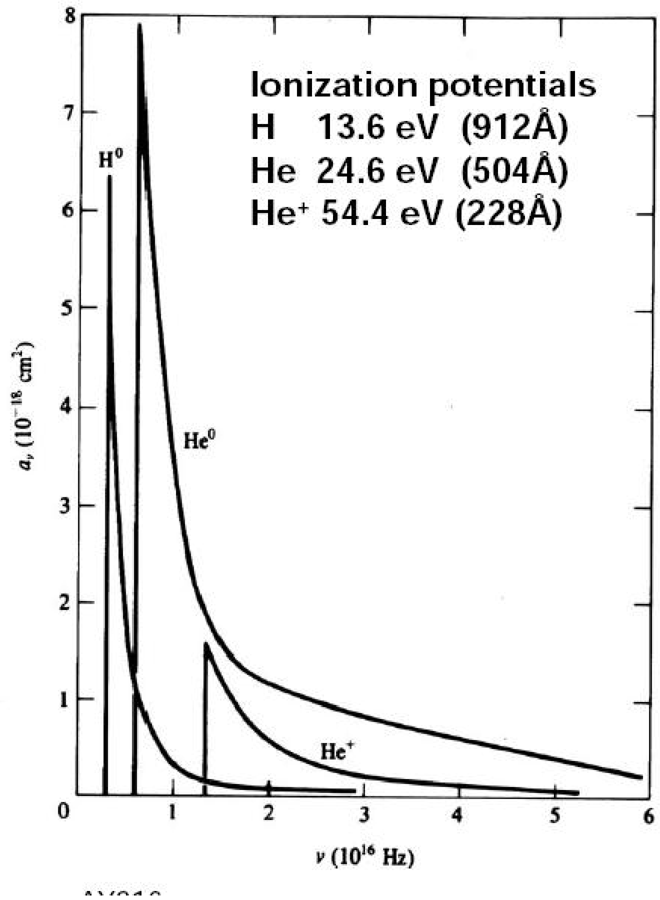


Figure 7.3: Ionization cross sections from the ground levels of neutral hydrogen, neutral helium and singly-ionized helium.

He/H$\sim$ 0.1. The atomic physics of He is made more complex by its two electrons and three possible ionic forms. All equations above can be easily generalized to Helium by simply replacing the appropriate coefficients.

The ionization potentials for $\mathrm{He_I}$ and $\mathrm{He_{II}}$ are, respectively, 24.6 eV and 54.4 eV. All the corresponding cross sections decrease as ν^{-3} for frequencies above the threshold. However, the cross section for $\mathrm{He_I}$ at its threshold is nearly 10 times larger than that for $\mathrm{H_I}$ at the same frequency (see Figure 7.3). This compensates for the different abundance of the elements and implies that photons with energy $\sim$ 24.6 eV will be partially used to ionize $\mathrm{He_I}$ instead of $\mathrm{H_I}$. This phenomenon has a significant impact on the ionization structure of a gas cloud which will be strongly dependent on the details of the radiation spectrum.

Chapter 8

Thermal balance

We have now developed the tools to derive the ionization structure of intergalactic gas assuming a given temperature. However, this is not a self-consistent approach. The temperature of the gas is fixed by the balance among a series of heating and cooling processes. In this Chapter we wil discuss the most important ones.

8.1 Adiabatic cooling and heating of an ideal gas

The entropy of an ideal gas made of N monoatomic particles contained in a volume V can be written as

$$S = k_\mathrm{B} N \left[\ln \frac{V}{N} + \frac{3}{2} \ln \frac{U}{N} + X \right] \tag{8.1}$$

where $U = (3/2) N k_\mathrm{B} T$ denotes the internal energy of the gas at temperature T and X is a constant (known as the Sackur-Tetrode constant). Let us now define the entropy per unit mass $s = S/(mN)$ (with m the mass of the gas particles) and compute its differential in terms of the variations in number density $n = N/V$ and temperature T. After some simple algebra, we obtain:

$$ds = d\left(\frac{S}{mN}\right) = \frac{k_\mathrm{B}}{m}\left(-\frac{dn}{n} + \frac{3}{2}\frac{dT}{T}\right) \ . \tag{8.2}$$

The energy exchanged by the gas with the ambient per unit volume per unit time can be written as $\rho\, T\, ds/dt$, where $d/dt \equiv \partial/\partial t + \mathbf{v}\cdot\nabla$ denotes the total (Lagrangian) derivative following each fluid element along its trajectory with velocity $\mathbf{v}$. Taking into account that $\rho = m\, n$, we find that

$$\rho\, T\, \frac{ds}{dt} = \frac{3}{2}\, n\, k_\mathrm{B} \frac{dT}{dt} - k_\mathrm{B} T \frac{dn}{dt} \ . \tag{8.3}$$

By definition, the left-hand side vanishes for adiabatic processes so that:

$$n\, \frac{d}{dt}\left(\frac{3}{2} k_\mathrm{B} T\right) - k_\mathrm{B} T \frac{dn}{dt} = 0 \ . \tag{8.4}$$

In an expanding universe, it is convenient to express the number density of particles as

$$n = \frac{n_0}{a^3}\,(1+\delta) \tag{8.5}$$

with n_0 the comoving number density, and δ the density contrast (accounting for spatial density fluctuations). If the total number of particles does not change with time (i.e. $n_0 =$ const.), then equation (8.4) gives

$$n\,\frac{d}{dt}\left(\frac{3}{2}k_{\rm B}T\right) - nk_{\rm B}T\left(-3\frac{\dot{a}}{a} + \frac{\dot{\delta}}{1+\delta}\right) = 0\ , \tag{8.6}$$

where $\dot{x} \equiv dx/dt$. If we just consider the expansion of the universe with no spatial fluctuations (i.e. impose $\delta = 0$), we obtain

$$\frac{d\ln T}{dt} = -2\,\frac{d\ln a}{dt} \tag{8.7}$$

which gives $T \propto a^{-2}$. Remember that the temperature of the cosmic microwave background scales as $T \propto a^{-1}$ as the energy of each photon is redshifted by the cosmic expansion factor. Therefore, when baryonic matter are radiation are decoupled they follow different thermal histories.

On the other hand, if we consider the growth of spatial perturbations (and keep a fixed), we get

$$\frac{d\ln T}{dt} = \frac{2}{3}\,\frac{d\ln(1+\delta)}{dt}\ , \tag{8.8}$$

or $T \propto (1+\delta)^{2/3}$. The collapse (expansion) of adiabatic perturbations makes the gas hotter (colder). This phenomenon is known as *adiabatic heating (cooling).*

8.2 Cooling and heating functions

A number of physical processes can inject or subtract energy from a gas. The *heating function*, $\mathcal{H}$, gives the total energy gained per unit volume per unit time. Similarly, the *cooling function*, Λ, quantifies the energy lost per unit volume and unit time. The kinetic temperature of the gas in a steady state is determined by the condition that $\mathcal{H} = \Lambda$. More generally, the heating and cooling functions are related to the change of temperature with time:

$$n\,\frac{d}{dt}\left(\frac{3}{2}k_{\rm B}T\right) - nk_{\rm B}T\left(-3\frac{\dot{a}}{a} + \frac{\dot{\delta}}{1+\delta} + \frac{\dot{n}_0}{n_0}\right) = \mathcal{H} - \Lambda\ , \tag{8.9}$$

(we are assuming here that all the components in the gas – i.e. electrons, ions, molecules – have the same temperature; we will see in Section 8.2.2

that this is a good approximation). This equation simply states that the net thermal input per unit volume and time, $\mathcal{H} - \Lambda$, equals the rate of increase of thermal energy, plus the work done by the gas.

Ionization and recombination processes change the comoving number density of cosmic gas. Therefore an extra term proportional to $\dot{n}_0$ appears in the equation above with respect to equation (8.6). In fact, n_0 depends on the number density and ionization status of the different ionic species. For instance, in a cloud containing only hydrogen, $n_0 = n_{\rm H_I} + n_{\rm p} + n_{\rm e} = (2 - f_{\rm neut})\, n_{\rm H}$.

Equation (8.9) does not consider thermal conduction (the flow of internal energy from a region of higher temperature to one of lower temperature by the interaction of adjacent gas particles), which is negligible in the IGM at low temperatures ($T <$ a few $\times\, 10^4$ K), as the conductivity of an ideal plasma is $\kappa = 5 \times 10^{-7}\, T^{5/2}$ erg s^{-1} cm^{-1} K^{-1} (the so-called Spitzer conductivity). For much higher temperatures(for instance those found in the intracluster medium) and no magnetic fields, the long mean free path of particles gives a high thermal conductivity κ and an additional term $\nabla \cdot (\kappa\, \nabla T)$ must be added on the right hand side of the equation above. Since heat-conducting electrons spiral around magnetic-field lines, the presence of a magnetic field is sufficient to markedly reduce the conductivity transverse to the field lines. In the presence of tangled magnetic fields, heat conduction is further reduced as: 1) electrons travelling along tangled magnetic field lines must cover longer distances between hot and cold regions of space; 2) electrons, while they are traveling along the field lines, become trapped and untrapped between magnetic mirrors, regions of strong magnetic field.

8.2.1 Energy input by photoionization

A primary mechanism for heating the IGM is photoionization. It is simplest to begin by considering a gas cloud made of pure H. The kinetic energy of each newly created photoelectron is given by the energy of the ionizing photon as

$$\frac{1}{2}\, m_{\rm e}\, v_{\rm e}^2 = h(\nu - \nu_0)\ . \tag{8.10}$$

At any specific point in the gas distributuion, the energy input per unit volume per unit time is

$$\mathcal{H}_{\rm pi}({\rm H}) = n_{\rm H_I}\, 4\pi \int_{\nu_0}^{\infty} \frac{J_\nu}{h\nu}\, h(\nu - \nu_0)\, a_\nu({\rm H_I})\, d\nu\ . \tag{8.11}$$

The mean kinetic energy of photoelectrons right after ionization is then

$$\frac{\displaystyle\int_{\nu_0}^{\infty} \frac{J_\nu}{h\nu}\, h(\nu - \nu_0)\, a_\nu({\rm H_I})\, d\nu}{\displaystyle\int_{\nu_0}^{\infty} \frac{J_\nu}{h\nu}\, a_\nu({\rm H_I})\, d\nu} = \frac{3}{2}\, k_{\rm B} T_i\ . \tag{8.12}$$

The quantity $(3/2)k_{\rm B}T_i$ represents the mean energy of the newly created photoelectrons expressed in terms of an initial temperature. Note that T_i does not depend on the normalization of the intensity of radiation but only on its spectral distribution. For any J_ν the integration can be carried on numerically. Radiative transfer effects slightly complicate the picture. The higher energy photons penetrate more into the gas (remember that $a_\nu \propto \nu^{-3}$) and the mean energy of photoelectrons produced at larger optical depths from the source of the radiation field is higher.

The photoheating rate for other elements than hydrogen can be described with equations analogous to (8.11).

8.2.2 Thermalization

Photoelectrons are rapidly thermalized, and on a short time scale the velocity distribution of atoms, electrons, and molecules are closely Maxwellian with a single temperature applicable to all these components. Typical deviations from a Maxwellian distribution are of a part every 10^6. This circumstance results mainly from the enormous predominance of H and He relative to the other elements. Collisions at energies of 10 eV or less among these atoms or between their ions and the free electrons are almost perfectly elastic and thus translational kinetic energy is exchanged back and forth many times before an anelastic collision can occur with an heavier atom or a molecule. This is just the condition for the establishment of a Maxwellian distribution and for equipartition of translational kinetic energy between different particles. For time reasons we will not derive the thermalization timescale from first principles (Coulomb scattering and collisions between neutral atoms), the interested reader can study Chapter 2 in the famous book by Spitzer ("Physical processes in the interstellar medium"). We will simply mention here that the timescale for electron-electron collisions is of the order of $10^5/n_{\rm e}$ s (when the electron density is expressed in cm^{-3}) for the energetic particles released by photoionization. For electron-ion collisions, the corresponding timescale is 22 $A_iT_{\rm e}^{3/2}/(n_{\rm e}Z_i^2)$ s, where a_i and Z_i are the ion mass and charge in atomic units. These timescales are much shorter than the ionization and recombination timescales and this validates our previous results.

8.2.3 Other sources of energy input

Photoionization is the most important heating process in the IGM. However, other processes might give relevant contributions especially at high redshift ($z > 2$). For instance,

- Compton heating of electrons by photons of the hard X-ray background;

- Photoelectric emission from dust grains hit by hard background photons.

It is still an open question how important these processes are to explain the thermal properties of the IGM.

8.2.4 Compton cooling/heating against the CMB

Compton scattering of CMB photons off free electrons couples the kinetic temperature of cosmic gas with the temperature of the photon background. Note that the spectrum of the CMB remains close to a blackbody because the heat capacity of radiation is very much larger than that of matter (i.e. there are vastly more photons than baryons).

Due to the rapid cosmic expansion, Compton scattering keeps $T_{\rm gas} = T_{\rm CMB} = 2.726\,(1+z)$ K only down to a redshift $z > 150\,(\Omega_{\rm b}h^2/0.022)^{2/5}$. Subsequently, for a smooth background, the gas expands adiabatically $T_{\rm gas} \propto (1+z)^2$ while $T_{\rm CMB} \propto (1+z)$.

However, in the presence of density fluctuations, Compton scattering can provide an important cooling mechanism down to $z > 2-3$ while, at lower redshifts, adiabatic cooling dominates. The Compton cooling rate can be written as:

$$\Lambda_{\rm com} = 5.406 \times 10^{-36} {\rm erg~cm^{-3}~s^{-1}~K^{-1}}\,(1+z)^4\,(T_{\rm gas} - T_{\rm CMB})\,\frac{n_{\rm e}}{n_{\rm tot}}\ . \quad (8.13)$$

8.2.5 Recombination cooling

The energy lost by the thermal electron plasma (per unit volume per unit time) when electrons recombine with protons to form neutral hydrogen is:

$$\Lambda_{\rm r}({\rm H}) = n_{\rm e}\,n_{\rm p}\,k_{\rm B}T\,\beta_{\rm A}({\rm H},T)\ , \quad (8.14)$$

where

$$\beta_{\rm A}({\rm H},T) = \sum_{n=1}^{\infty} \beta_n({\rm H},T) = \sum_{n=1}^{\infty}\sum_{\ell=0}^{n-1} \beta_{n\ell}({\rm H},T) \quad (8.15)$$

with

$$\beta_{n\ell}({\rm H},T) = \frac{1}{k_{\rm B}T}\int_0^\infty v_{\rm e}\sigma_{n\ell}({\rm H},T)\,\frac{1}{2}m_{\rm e}v_{\rm e}^2\,f(v_{\rm e})\,dv_{\rm e}\ . \quad (8.16)$$

This is effectively a kinetic-energy averaged recombination coefficient. Note that since the recombination cross sections are approximately proportional to $v_{\rm e}^{-2}$, the electrons of lower kinetic energy are preferentially captured, and the mean energy of the captured electrons is somewhat less than $(3/2)k_{\rm B}T$. If recombinations were the only cooling mechanism available, the resulting electron temperature would actually be slightly hotter than the "radiation temperature" after recombination cooling. This is because the slower electrons are preferentially recombined out the free electron plasma, skewing the velocity distribution of the remaining free electrons towards higher energies and hence higher temperatures.

8.2.6 Continuum radiation and free-free cooling

Free-free, free-bound and two-photon processes emit a continuum spectrum of radiation and contribute to the gas cooling. By far, the dominant process is free-free radiation: thermal electrons can scatter off ions and emit bremsstrahlung radiation. The free-free cooling rate for an ion with nuclear charge Z is

$$\Lambda_{\rm ff}(Z) = \frac{32\pi e^6 Z^2}{3^{3/2} h m_{\rm e} c^3} \left(\frac{2\pi k_{\rm B} T}{m_{\rm e}}\right)^{1/2} g_{\rm ff}\, n_{\rm e}\, n_+ \ , \tag{8.17}$$

where n_+ is the number density of ions with nuclear charge Z, and $g_{\rm ff}$ is the free-free Gaunt factor, which is a slowly varying function of density and temperature. For UV-to-NIR wavelengths and typical conditions of the IGM, $g_{\rm ff}$ ranges between 1.0 and 1.5. Substituting the values of all the physical constants, the free-free cooling rate becomes (in cgs units):

$$\Lambda_{\rm ff}(Z) = 1.42 \times 10^{-27}\, Z^2\, T^{1/2}\, g_{\rm ff}\, n_{\rm e}\, n_+ \ . \tag{8.18}$$

Overall, the free-free cooling is fairly inefficient in the IGM but it becomes important at high temperatures.

8.2.7 Collisionally excited line emission

Metal ions like $\rm O_{II}$, $\rm O_{III}$, $\rm N_{II}$, and a few others, while relatively underabundant compared to H or He, turn out to be the most important coolants in the IGM. In the ground-state, the fine-structure levels of these ions have typical excitation potentials of a few eV. The thermal energies of the electrons are also in the same ball park for typical temperatures of 10^4 K. This makes electron-ion impact excitations of the metal ions very efficient. By contrast, the first excited levels of H and He are $\sim$ 10 eV above the ground state, so that collisional excitation of these elements is very inefficient at typical densities and temperatures of the IGM. Proton-ion and ion-ion impact excitation are inefficient because the Coulomb repulsion between the ions is too large. However, some important collisional processes do occur between neutral atoms and ions (e.g. charge-exchange reactions between O and H that happen to have nearly identical ionization potentials) that can contribute to the cooling. Electron-ion impact excitation of metal ions followed by radiative line emission is the dominant cooling mechanism in ionized gas with metallicities greater than a few percent of the solar value. The abundance of metals relative to H in the IGM thus plays a crucial role in determining its thermal structure.

8.2.8 The cooling function

For a cosmic plasma, we can define a cooling timescale as the ratio

$$t_{\mathrm{c}} = \frac{u_{\mathrm{gas}}}{n_{\mathrm{e}}\, n_{\mathrm{tot}}\, \Lambda(T)} \,, \tag{8.19}$$

where Λ is obtained summing up over all the relevant processes, $u_{\mathrm{gas}} = (3/2) n_{\mathrm{tot}} k_{\mathrm{B}} T$ is the energy density of the plasma, and n_{tot} the total number density.

In the absence of any significant radiation field, ionization fractions and level populations can be computed assuming collisional ionization equilibrium (CIE). Under this condition, the only terms that significantly contribute to the cooling function are: collisional line radiation, continuum radiation, and recombination cooling. The resulting cooling function obtained by Sutherland and Dopita (1993) is shown in Figure (8.1) for different chemical compositions of the plasma. Its main features are as follows.

1. When the temperature of the gas is larger than $10^6 - 10^7$ K, the cooling function is dominated by free-free radiation, and the cooling function increases slowly;

2. For temperatures between 10^4 and 10^6 K, the energy loss is dominated by atomic line cooling. The peak at temperatures slightly above 10^4 K is due to Lyα emission from atomic hydrogen. At very low metallicities a second peak arises near 10^5 K due to recombination of atomic He. Metals give rise to a higher peak at $\sim 10^5$ K and slightly above, due to line emission from the heavier atoms.

3. For lower temperatures, cooling becomes extremely inefficient and cooling times start approaching the age of the universe. Basically, gas at $T \sim 10^4$ K stays at constant temperature for very long time (see, however, the discussion in Section 8.2.9).

The cooling properties of the gas are very different in the presence of a strong radiation field, when level populations and ionization fractions are far from the CIE value. One particular example is provided in the bottom panel of Figure 8.1. Note that the cooling function substantially differ from the CIE case.

We often characterize the cooling coefficient with a single parameter, the temperature. These cooling functions are determined by assuming either that the ionization state at a given temperature is characterized by collisional equilibrium or that all gas follows a particular ionization history. Since the cooling of a plasma depends on the ionization history of the constituent ions (and thus on the thermal history of the plasma itself), there can actually be a range in the value of Λ at a given T, depending on the details of the ionization evolution (compare top and bottom panels in Figure 8.2). Moreover, the

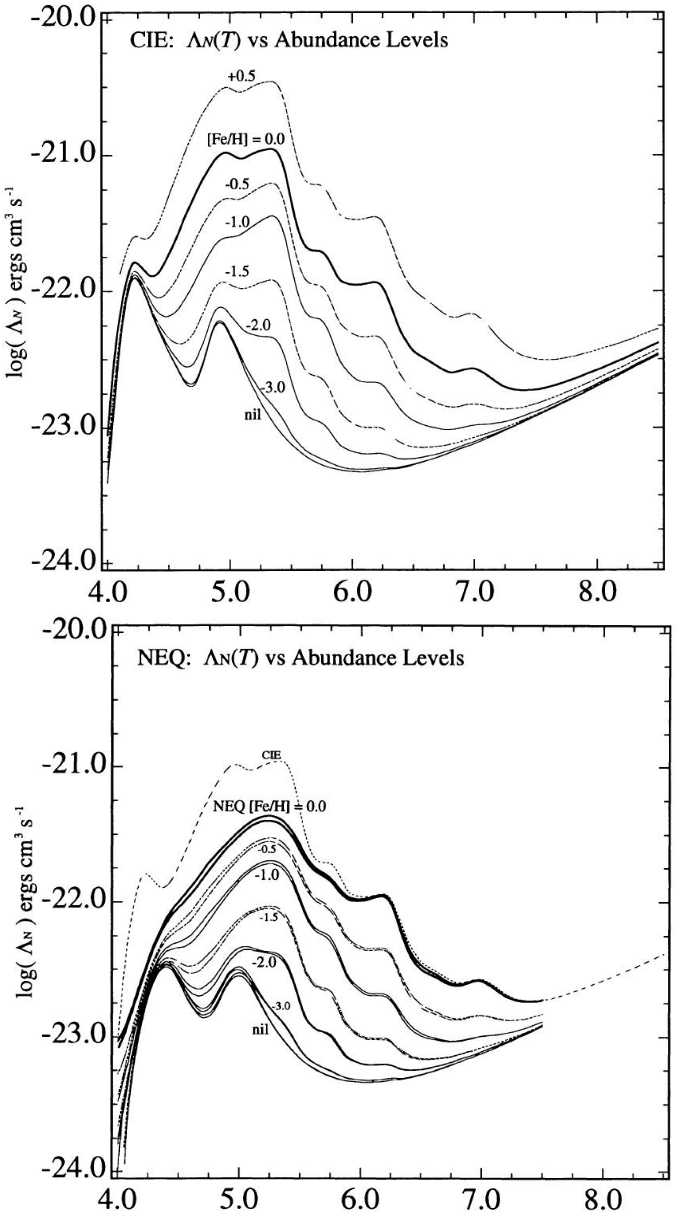


Figure 8.1: Cooling functions of low-density cosmic gas in CIE (top) and in the presence of a particular radiation field (bottom). Different lines refer to different metal abundances as indicated in the labels.

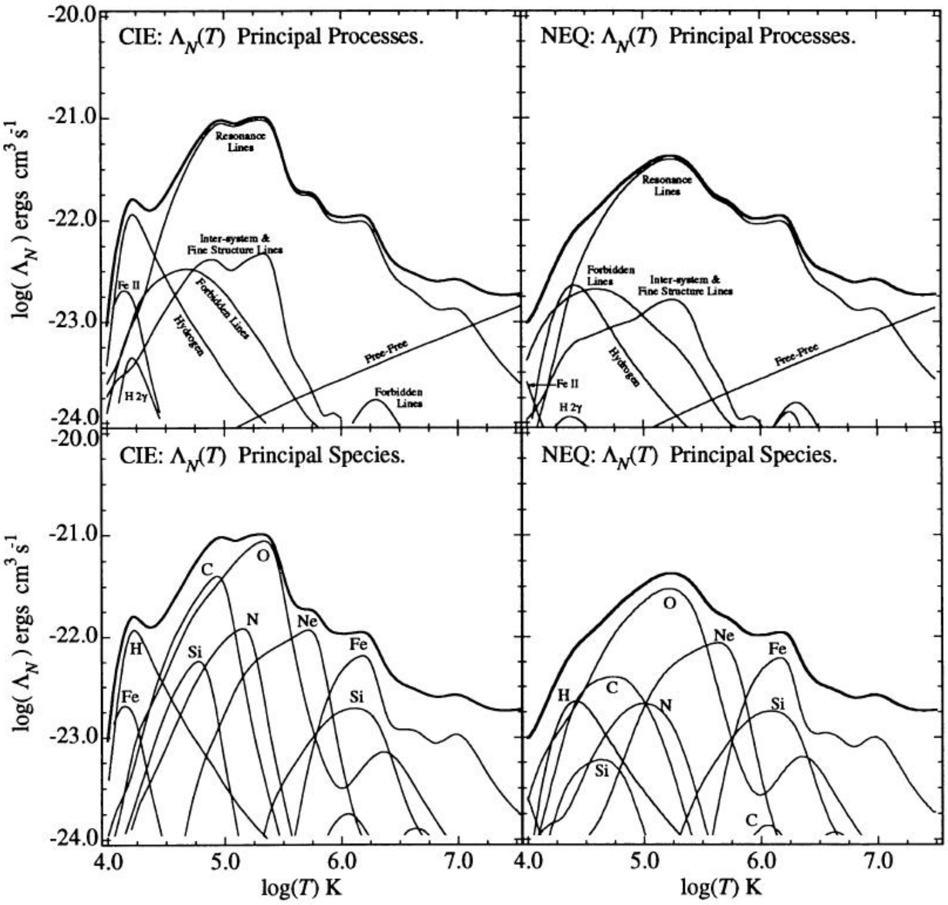


Figure 8.2: Processes and chemical elements contributing to the cooling functions in the previous figure (for solar metallicity).

cooling function will also depend on geometric factors (e.g. the shape of the gas cloud) that will affect (through the radiative transfer equation) the degree of radiation coupling in the ionization balance calculations.

All these effects are generally accounted for in the most accurate models of the IGM. In particular, cooling functions are computed following the specific ionization and thermal history of each fluid element.

8.2.9 Molecular cooling

In the absence of metals, collisional excitation of roto-vibrational molecular transitions followed by radiative decay is effective in cooling metal-free gas below a temperature of $\sim 10^4$ K. At these low temperatures, simple molecules (such as H_2, H_2^+, HD, HeH^+, LiH) can be produced in the gas phase (without the catalyzing effect of dust grains). Collisions with other molecules or with H and He atoms thus excite molecular line transitions and provide an additional channel which dominates the cooling of metal-free gas at $T < 10^4$ K. The main coolant is molecular hydrogen, H_2, which is anyway very inefficient (see Figure 8.3). Cooling at these temperatures is thus very slow.

Note that:

1. Molecules are very fragile and can be easily dissociated by ambient radiation. For instance, the H_2 molecule can be destroyed by the so-called Solomon process. In this case, UV photons in particular transition lines in the Lyman ($h\nu > 11.2$ eV, $\lambda < 1108$ Å) and Werner ($h\nu > 12.3$ eV, $\lambda < 1008$ Å) bands can excite electronic states of the H_2 molecule. Radiative decay from the excited states leads to molecular dissociation in 15% of the cases. It has been estimated that a radiation intensity of $J_\nu > 10^{-23}$ (in cgs units) in the Lyman-Werner bands is sufficient to make molecular cooling inefficient.[1]

2. As soon as the IGM is polluted with material expelled from stars, metals become the main coolants.

[1] Photons with $\lambda < 912$ Å are more likely to ionize $\mathrm{H_I}$, only photon energies between 11.2 and 13.6 eV are of interest here.

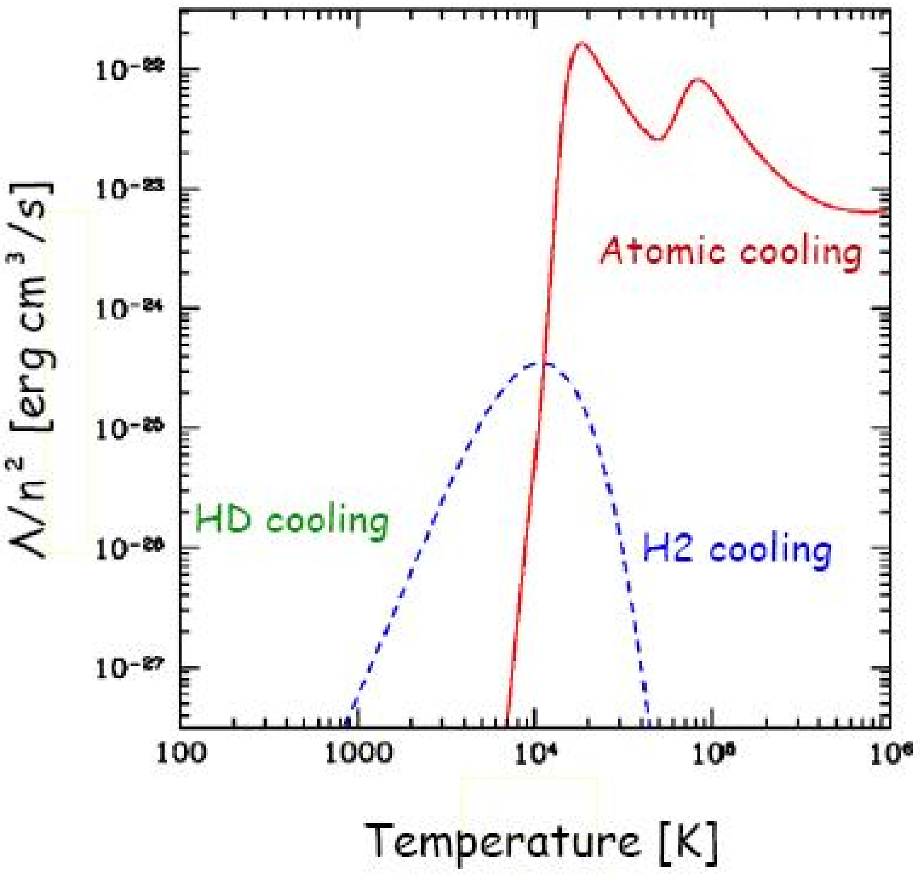


Figure 8.3: Cooling function of primordial gas in the presence of molecules.

Chapter 9

The extragalactic UV background

The absence of a Gunn-Peterson trough in quasar spectra at $z < 5$ indicates that intergalactic hydrogen is highly ionized. A key factor in determining the ionization state of the baryons is the intensity of the cosmic UV background. What are the main sources of UV photons in intergalactic space? Do they produce enough photons to maintain the intergalactic diffuse gas in a highly ionized state? This is the subject of today's class.

9.1 Models

A radiation background arises as the integrated emission from all sources along a given line of sight. Due to the expansion of the universe, more distant sources contribute to the background with light emitted at earlier epochs and higher frequencies. Mathematically, the mean specific intensity of the UV background as seen at frequency $\nu_{\rm obs}$ by an observer at redshift $z_{\rm obs}$ can be written as

$$J_{\nu_{\rm obs}}(z_{\rm obs}) = \frac{1}{4\pi}\int_{z_{\rm obs}}^{\infty}\left(\frac{1+z_{\rm obs}}{1+z}\right)^3 \epsilon_\nu(z)\,\exp\left[-\tau_{\rm eff}(\nu_{\rm obs}, z_{\rm obs}, z)\right]\frac{dl}{dz}\,dz \tag{9.1}$$

where $\nu = \nu_{\rm obs}(1+z)/(1+z_{\rm obs})$, ϵ_ν is the mean (proper) volume emissivity of UV radiation, $\tau_{\rm eff}$ is the effective optical depth at $\nu_{\rm obs}$ of the IGM between redshifts $z_{\rm obs}$ and z, and dl/dz is the proper line element. Equation (9.1) gives the formal solution of the radiative-transfer equation in an expanding universe.

The function $\epsilon_\nu(z)$ gives the energy emitted in radiation of frequency ν per unit time and unit proper volume by all possible cosmic sources at redshift z. As we will see in greater detail below, ϵ_ν includes two terms: a contribution from the IGM itself (due to recombination radiation that escapes the clouds) and one from direct sources of UV radiation (galaxies and quasars). For the latter case, the proper emissivity can be estimated by taking into account the evolution in luminosity, number density and spectrum of all sources in the universe. Needless to say, it is a challenge to accurately determine this function.

The effective optical depth is defined as $\exp(-\tau_{\rm eff}) = \langle \exp(-\tau) \rangle$ where the mean is taken over all the lines of sight from the redshift of interest. This term accounts for the radiative transfer through the clumpy IGM. As a first approximation one can consider the IGM as a random distribution of discrete clouds. In this case it can be shown that

$$\tau_{\rm eff}(\nu_{\rm obs}, z_{\rm obs}, z) = \int_{z_{\rm obs}}^{z} dz' \int_0^\infty dN_{\rm H_I}\, f(N_{\rm H_I}, z') \left[1 - \exp(-\tau(\nu'))\right] \;, \quad (9.2)$$

where $\tau(\nu')$ is the optical depth of an individual cloud for radiation with frequency $\nu' = \nu_{\rm obs}(1+z)/(1+z_{\rm obs})$ while $f(N_{\rm H_I}, z) = \partial^2 N/\partial N_{\rm H_I} \partial z$ indicates the redshift and column-density distribution of the absorbers. Analytical fits to the observed $f(N_{\rm H_I}, z)$ of quasar absorption lines can then be used to compute $\tau_{\rm eff}$. This, however, requires knowledge of the frequency-dependent optical depth of each cloud. Radiative transfer calculations for single clouds indicate that, for photon energies between 13.6 and 54.4 eV (wavelenghts between 228 Å and 912 Å), $\tau \simeq N_{\rm H_I}\, a_\nu({\rm H_I})$. This is because Helium is almost completely ionized and its first ionization threshold at 504 Å gives negligible contributions to the opacity. For $\lambda < 228$ Å, instead, also the ionization of He$_{\rm II}$ becomes important and $\tau \simeq N_{\rm H_I}\, a_\nu({\rm H_I}) + N_{\rm He_{II}}\, a_\nu({\rm He_{II}})$ with

$$N_{\rm He_{II}} \simeq 1.8 N_{\rm H_I} \frac{J_{\nu_{\rm H_I}}}{J_{\nu_{\rm He_{II}}}} \quad (9.3)$$

which holds for optically thin clouds. The *hardness ratio* (or, better, softness ratio) $J_{\nu_{\rm H_I}}/J_{\nu_{\rm He_{II}}}$ compares the intensity of the background at the ionization edges for H$_{\rm I}$ (13.6 eV) and He$_{\rm II}$ (54.4 eV). Typically this ratio assumes a value close to 50 for quasar generated backgrounds and much larger values for galaxy generated backgrounds (which are much softer).

Note that the redshift integration in equation (9.1) formally extends to infinity but will not receive any contributions from the redshifts z where $\tau_{\rm eff}(\nu_{\rm obs}, z_{\rm obs}, z) \gg 1$. In Figure 9.1, we show the redshift separation $\Delta z = z - z_{\rm obs}$ corresponding to $\tau_{\rm eff} = 1$ as a function of $z_{\rm obs}$. This gives the mean free path of a photon in redshift units. For radiation at 912 Å, Δz decreases from 1.8 at $z_{\rm obs} = 0$ to 0.08 at $z_{\rm obs} = 5$. Similarly, at 600 Å (where the IGM absorption is lower due to the strong frequency dependence of $a_\nu({\rm H_I})$), Δz equals 4 at $z_{\rm obs} = 0$ and 0.2 at $z_{\rm obs} = 5$ (neglecting helium absorption). In other words, the UV background becomes more dominated by local sources as the redshift increases. This happens because the IGM is more optically thick at high redshift and absorbs all the radiation coming from distant

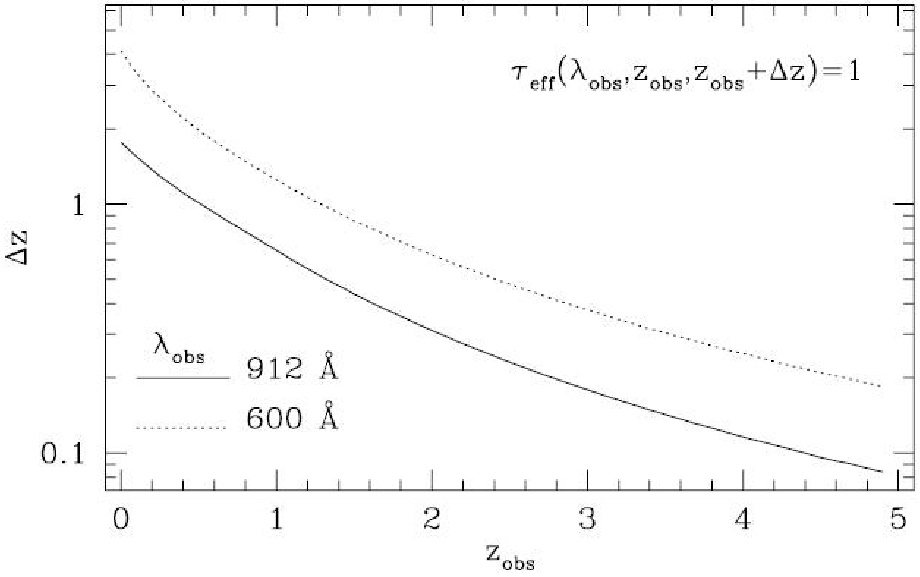


Figure 9.1: Redshift separation $\Delta z = z - z_{\rm obs}$ corresponding to $\tau_{\rm eff}(\nu_{\rm obs}, z_{\rm obs}, z) = 1$ as a function of $z_{\rm obs}$. Note that only hydrogen absorption is considered.

sources. As a consequence of this, spatial fluctuations in the metagalactic hydrogen ionization rate at $z < 4$ are expected to be small. At these redshifts, the mean free path for ionizing photons is substantially larger than the mean separation between ionizing sources (galaxies and quasars); therefore a spatially uniform ionizing background is expected to be a reasonable approximation. However, towards higher redshifts, spatial fluctuations in the ionizing background are gradually amplified. The amplification is attributable to diminishing source numbers, a smaller mean free path and the inhomogeneous distribution of the ionizing sources themselves.

9.1.1 Cosmic sources of UV photons

Many astronomical sources are capable of emitting UV light with photon energies in the range between 10 and 100 eV.

Quasars

Bright quasars are copious sources of ionizing photons. Their UV spectra show power-law continua, $F_\nu \propto \nu^\alpha$, with α ranging between ~ -0.5 and ~ -1.5 (see Figure 9.2).

The number density of bright quasars presents a marked redshift evolution (see Figure 9.3). These sources are extremely rare nowadays but were

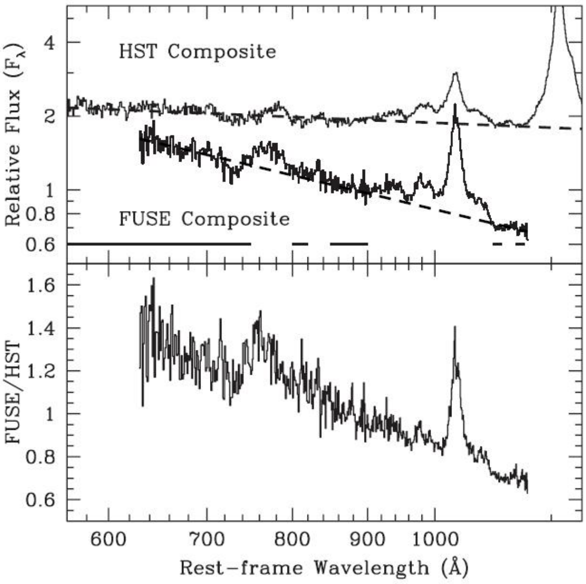


Figure 9.2: *Top:* Composite QSO spectra (solid lines) from HST and FUSE data with power-law continuum fits (dashed lines). The best-fitting slope is $\alpha = -1.76^{+0.12}_{-0.12}$ for HST and $\alpha = -0.56^{+0.38}_{-0.28}$ for FUSE. Note that the FUSE sample is dominated by low-redshift, low-luminosity AGNs that have a tendency towards hotter accretion disks and harder UV spectra. *Bottom:* Ratio of FUSE to HST composite spectra.

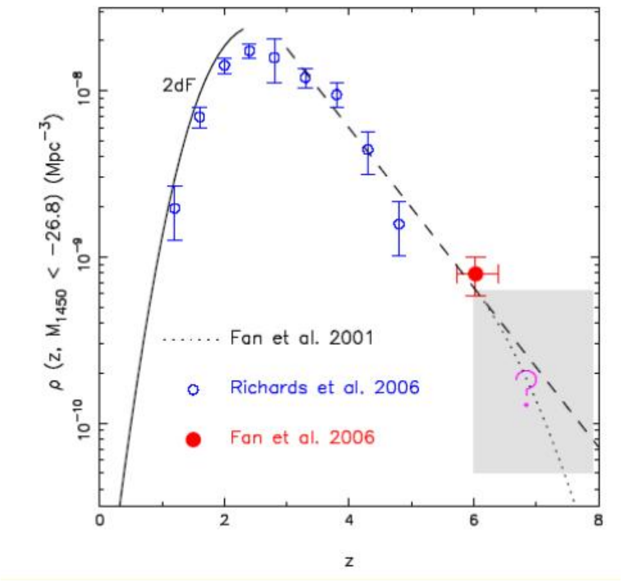


Figure 9.3: Observed evolution of the number density of bright, optically selected quasars.

much more abundant in the past. Their abundance peaks at $z \sim 2-3$ (the "cosmic quasar era") and declines by a factor of 20 between redshift 3 and 6. This is interpreted within galaxy formation models as follows. Building supermassive black holes via mass accretion and merging of smaller units takes time. Therefore these objects must be extremely rare in the young universe (i.e. at very high redshift). On the other hand, the quasar phenomenon requires efficient gas accretion onto the black holes. As massive galaxies form their stars out of their gas reservoir, they might, sooner or later, run out of fuel. The peak of quasar activity therefore corresponds to an epoch which is sufficiently late to allow for the formation of the supermassive black holes but sufficiently early to prevent galaxies to run out of gas for quasar accretion.

The cosmic UV emissivity due to quasars can be estimated by combining observations of their optical luminosity function (Figure 9.4) with spectral templates for quasar emission (in order to link the optical and the UV photon output). The resulting function $\epsilon_\nu^{\rm QSO}(z)$ at $h\nu$ = 13.6, 24.6, and 54.4 eV obtained in the pioneering paper by Haardt & Madau (1996) is shown in Figure 9.5. Note that most of the UV photons are injected during the cosmic quasar era.

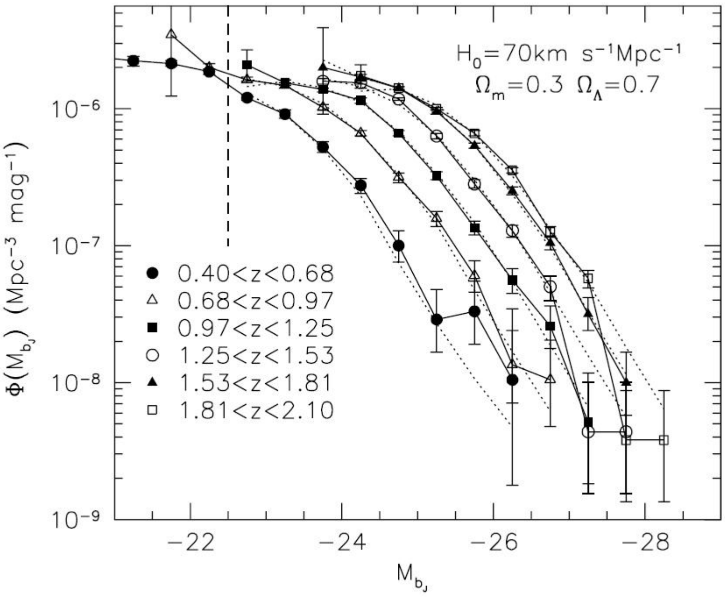


Figure 9.4: Observed evolution of the luminosity function for optically selected quasars from the 2dF redshift survey.

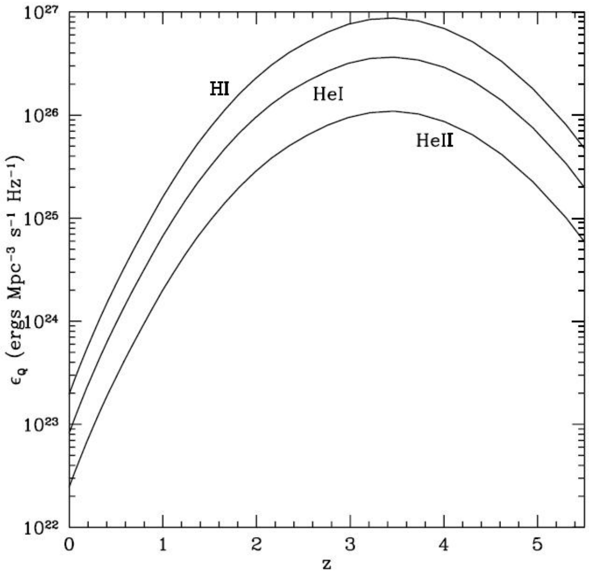


Figure 9.5: Quasar proper volume emissivity at the $\mathrm{H_I}$, $\mathrm{He_I}$ and $\mathrm{He_{II}}$ ionization edges obtained by Haardt & Madau (1996) combining the observed evolution of the quasar luminosity function and their observed spectral energy distribution.

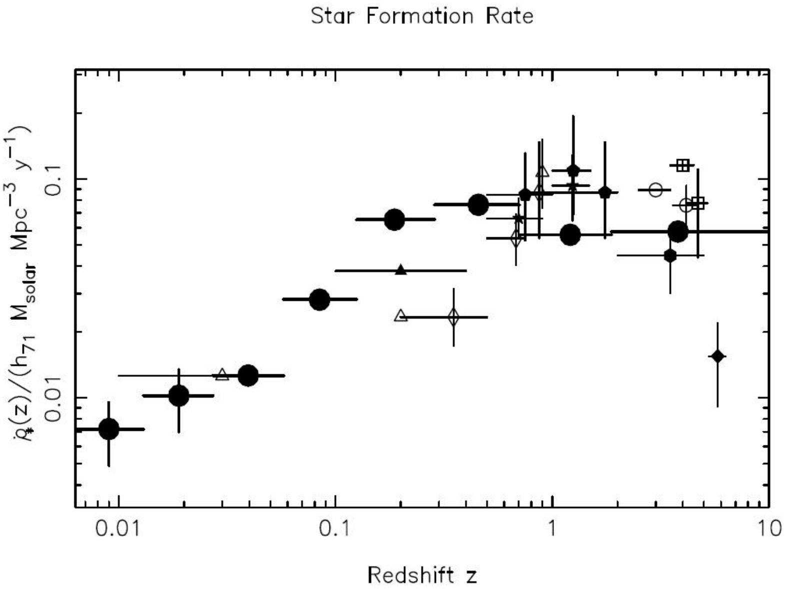


Figure 9.6: Cosmic star-formation history as determined by a number of different observables.

Massive stars

The hydrogen ionization threshold corresponds to a temperature of 15800 K. Therefore stars like the Sun ($T_{\rm eff} \sim 5800$ K) are too cold to produce significant amounts of ionizing photons. On the other hand, O ($T_{\rm eff} > 30,000$ K, $10 < M < 100\,M_\odot$) and the hottest B ($12,000 < T_{\rm eff} < 30,000$, $3 < M < 20\,M_\odot$) stars certainly contribute to the extragalactic UV background. Since these massive stars are extremely short lived for cosmological standards (O stars have lifetimes of only a few million years, B stars of a few tens of million years) they will only be found in actively star-forming galaxies.

In order to estimate the contribution of galaxies to the UV background it is thus necessary to determine the cosmic history of star formation. A large number of research groups have obtained consistent answers by using a variety of methods which suffer of different sytematics. These results are summarized in Figure 9.6 where the mean cosmic star-formation rate, $\dot{\rho}_*$ (the mass of newly formed stars per unit comoving volume per unit time), is plotted against redshift. Note that nowadays the star-formation activity is rather low with respect to the past.

It is then important to know what fraction of the newly formed stars is in the O and B classes. This is quantified by the initial mass function (IMF) of the stars, $\phi(M)$, which is often approximated with a single slope power-law $\phi(M) \propto M^{-s}$ for $M_{\rm min} < M < M_{\rm max}$. The value $s = 2.35$ corresponds to the classical Salpeter IMF. The quantity

$$\frac{\phi(M)\,dM}{\int_{M_{\rm min}}^{M_{\rm max}} \phi(M)\,dM} \tag{9.4}$$

gives the fraction of stars in the mass interval $(M, M + dM)$ while

$$\frac{\phi(M)\,M\,dM}{\int_{M_{\rm min}}^{M_{\rm max}} \phi(M)\,M\,dM} \tag{9.5}$$

gives the mass fraction locked in the same objects. Note that, in principle, $\phi(M)$ can change with redshift or with the properties of the star forming objects. This introduces extra degrees of freedom in the models thus making them more uncertain.

The spectral energy distribution of the light emitted by young stars can be computed using sophisticated stellar population models. A sample output is shown in Figure 9.7.

Finally, since star-forming regions are always embedded in the interstellar medium of a galaxy (gas and dust), it is important to know how transparent this medium is for the UV photons. In other words, we need to estimate how much radiation will be able to reach the IGM streaming out of the galaxy. This is generally parameterized by the so-called escape fraction of UV photons, $f_{\rm esc}$. For starbursting galaxies at very low redshift, current estimates give $f_{\rm esc} < 0.06$ but the value at high redshift is still very uncertain (with published values ranging from $f_{\rm esc} < 0.04$ to $f_{\rm esc} > 0.5$). Recent numerical simulations of galaxy formation suggest that $f_{\rm esc}$ might indeed increase with redshift.

Reprocessed radiation

Beyond acting as a sink of UV radiation via photoionization, the IGM will also contribute to the extragalactic UV background via radiative recombinations and redshifted line radiation (line radiation is smeared into a continuum by the redshift effect). For instance, the following processes contribute to the diffuse background field:

1. Recombinations to the ground state of $\rm H_I$ and $\rm He_I$;

2. $\rm He_{II}$ Lyα emission at 40.8 eV;

3. $\rm He_{II}$ two-photon continuum emission;

4. $\rm He_{II}$ Balmer continuum emission at $h\nu \geq 13.6$ eV.

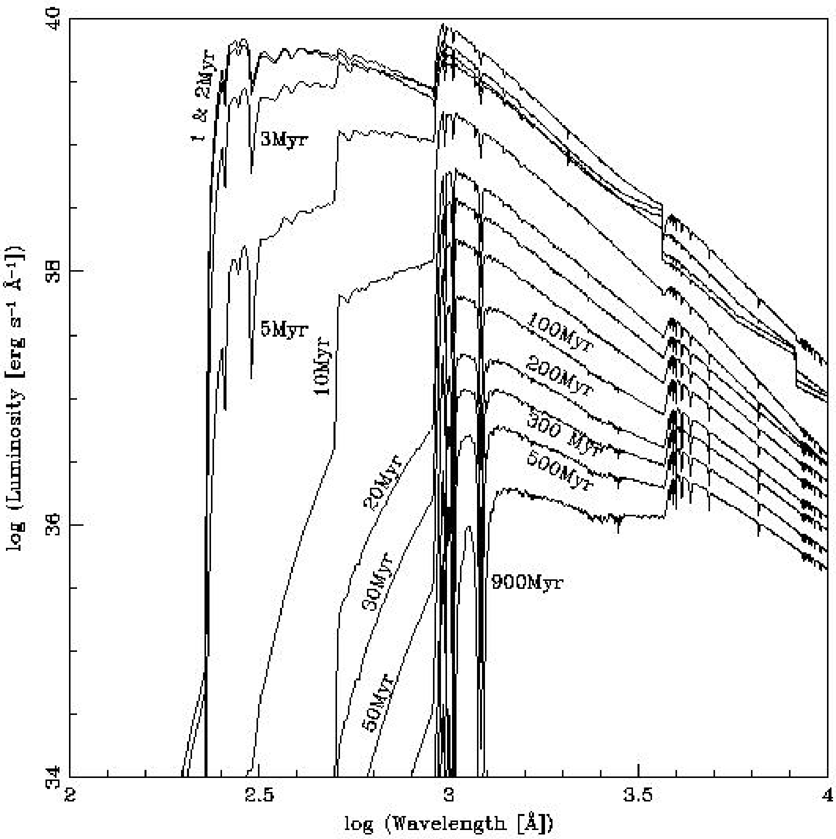


Figure 9.7: Model for the spectral energy distribution of the light emitted by a starburst which converts $10^6 M_\odot$ of gas into stars. Different curves correspond to different epochs from the burst as indicated by the labels. A metallicity of $Z = 0.001$ and a Salpeter IMF with a minimum stellar mass of 1 $M_\odot$ and a maximum of 100 $M_\odot$ are assumed. Nebular emission due to the diffuse, ionized gas surrounding the starbursting region is included.

Radiative decay of exotic particles

A more speculative potential source of diffuse extragalactic UV emission is the radiative decay of exotic particles of cosmological origin. In the early 1990s it was proposed that massive decaying neutrinos could simultaneously explain the dark-matter problem and the ionization structure of the IGM. This hypothesis is now ruled out by observations but we cannot exclude the existence of some other unknown decaying particle which contributes to the extragalactic UV background.

9.1.2 Results

The cookbook recipe to estimate the mean intensity of the cosmic UV background is therefore:

1. Compute the emissivity function for your favourite cocktail of sources of UV photons.

2. Solve the radiative transfer problem through the clumpy IGM using the observed properties of the Lyα forest and of Lyman-limit systems (this gives you $\tau_{\rm eff}$).

3. Compute the mean specific intensity of radiation using equation (9.1).

Since step 2) already requires knowledge of J_ν (see, for instance, equation 9.3), the solution is usually found by iteration.

In Figure 9.8 we show the redshift evolution of the background spectrum obtained assuming that quasars are the dominant sources of UV photons. Note that the IGM introduces some spectral features both in absorption and in emission on top of the power-law behaviour due to the quasars. The intensity of radiation peaks at $z \sim 2-3$. Similarly, the photoionization rates for H and He reach a maximum during the quasar era (Figure 9.9). Note that, assuming ionization equilibrium, the peak value $\Gamma(\mathrm{H_I}) \simeq 10^{-12}$ s^{-1} at $z \sim 2.5$ corresponds to a neutral fraction of $f_{\rm neut} = 0.418$ cm$^3 \cdot n_{\rm H} = 3.4 \times 10^{-6}(1+\delta)$ (assuming $T = 10^4$ K). [1] This shows that the UV background due to quasars only can indeed explain the high ionization of the IGM at $z < 5$.

In Figure 9.10, we show a different estimate of the UV background where the contribution of star-forming galaxies has been added on top of the quasar one. Note that the huge effect due to the uncertainty in $f_{\rm esc}$.

[1] This is obtained by considering pure hydrogen and imposing ionization equilibrium: $f_{\rm neut}\, n_{\rm H}\, \Gamma(\mathrm{H_I}) = [(1-f_{\rm neut})^2\, n_{\rm H}^2\, \alpha_{\rm A}(T)]$, which for $f_{\rm neut} \ll 1$ gives $f_{\rm neut} = n_{\rm H}\, \alpha_{\rm A}(T)/\Gamma$.

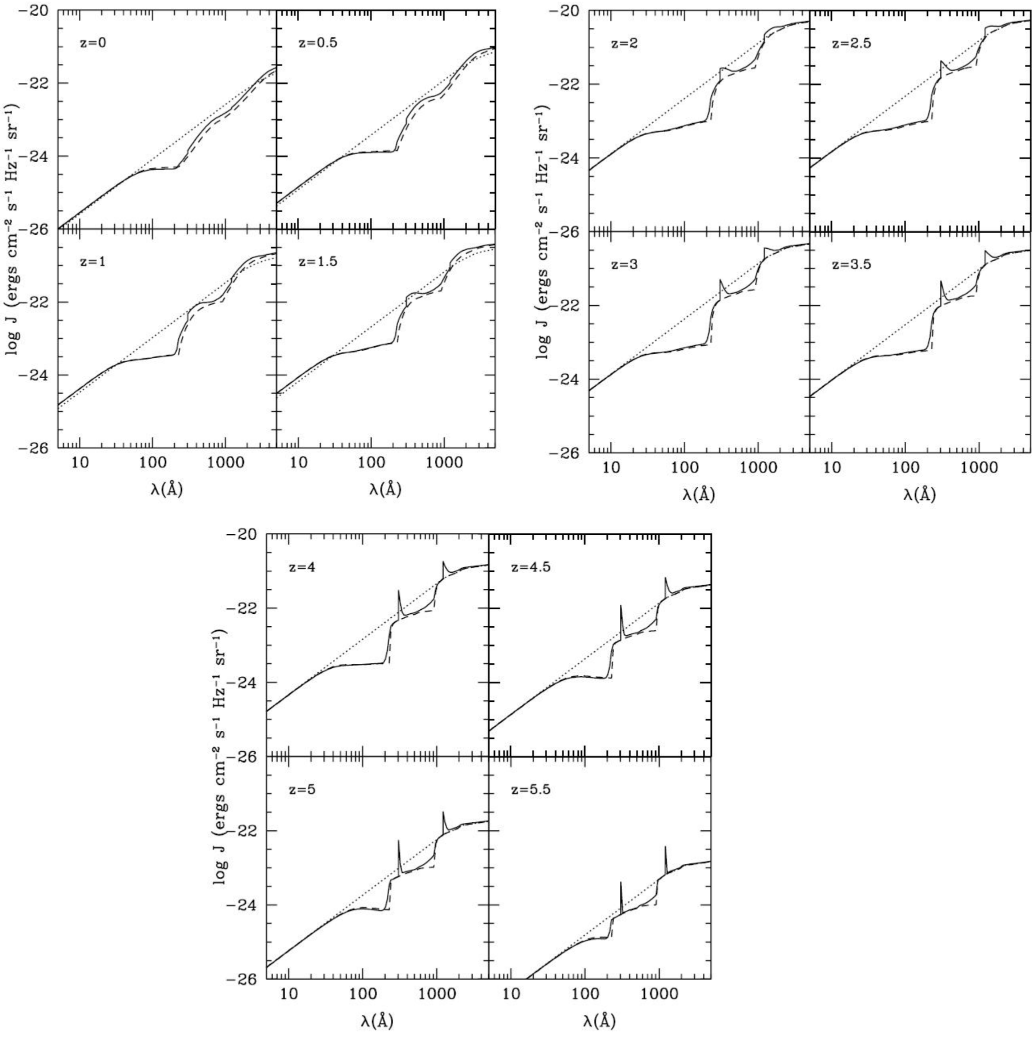


Figure 9.8: The cosmic ionizing background from 5 to 5000 Å estimated by Haardt & Madau (1996) at redshifts $z = 0, 0.5, 1, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5$. For comparison, the corresponding UV backgrounds in a purely transparent universe (no IGM absorption and emission) and in a purely absorbing IGM (no IGM emission) are indicated with dotted and dashed lines, respectively.

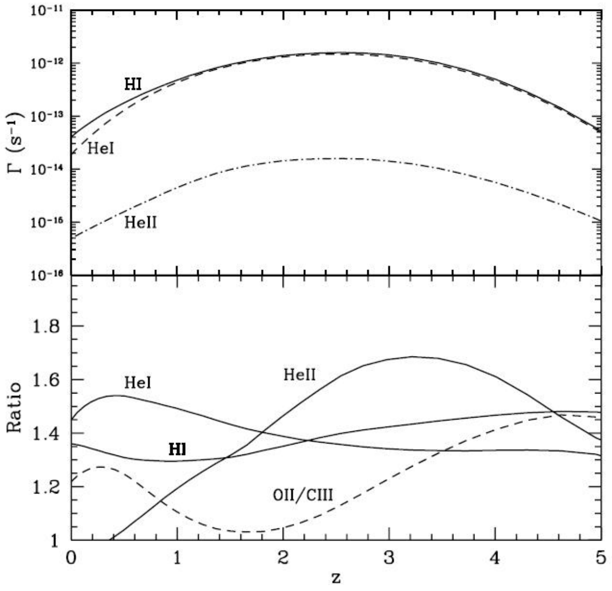


Figure 9.9: Top: Photoionization rates for H$_{\rm I}$ (solid), He$_{\rm I}$ (dashed) and He$_{\rm II}$ (dot dashed) derived by Haardt & Madau (1996). Please, ignore the bottom panel.

9.2 Observations

Two main methods have been used to estimate the amplitude of the UV background from astronomical observations: the flux decrement technique and the line-of-sight proximity effect method.

9.2.1 Flux decrement

The basic idea behind this method is to adopt a model for the gas density distribution and adjust the level of ionizing radiation such that the mean transmitted flux in the Lyα forest matches observations. This technique is very indirect and requires a number of external inputs.

From the analysis of absorption lines in quasar spectra, it is possible to measure the mean transmitted flux as a function of redshift. Recent data, for instance, give $\langle F\rangle_{\rm obs} = \exp\left(-\tau_{\rm eff}\right) = 0.878 \pm 0.019, 0.696 \pm 0.025$ and 0.447 ± 0.031 at $z = 2, 3$ and 4, respectively. These values have been corrected for absorption from metal lines and Lyα systems with damping wings. The starting point of the flux-decrement technique is that, as we will show in the next class, the Lyα optical depth should scale according to the parameter combination

$$\mu = \Omega_{\rm b}^2\, h^3\, T^{-0.7}\, \Omega_{\rm m}^{-0.5}\, \Gamma^{-1}\ . \tag{9.6}$$

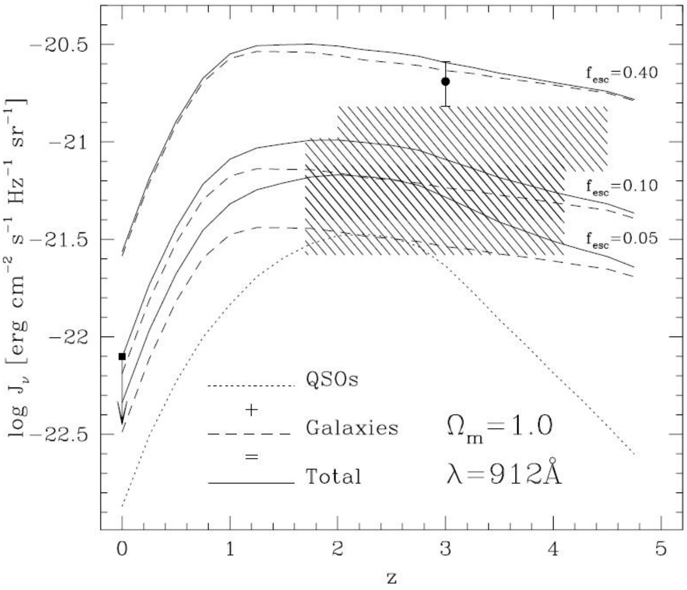


Figure 9.10: The cosmic UV background at $\lambda = 912$ Å as a function of redshift and for different values of $f_{\rm esc}$. The separate contributions of quasars and galaxies are indicated by dotted and dashed lines, respectively. The shaded region indicates the limits coming from the proximity effect. The arrow shows an experimental upper limit for the background at $z = 0$. The datapoint at $z = 3$ is derived from a composite spectrum of Lyman-break galaxies.

Imagine that you match the mean flux of artificial Lyα spectra constructed from hydrodynamical simulations to the observed values by rescaling the simulated optical depths in each pixel by a constant factor. This corresponds to changing the parameter combination μ by the same factor. If independent estimates are available for $\Omega_{\rm b}$, $\Omega_{\rm m}$, h, and T, the magnitude of Γ can be determined. In practice, there are two possibilities:

1. a number of simulations are run by varying $\Omega_{\rm b}$, $\Omega_{\rm m}$, h, and T within a plausibility range fixed by other observations and the different Γ estimates are used to determine its measure and the corresponding uncertainty.

2. a numer of different quantities (power spectra, b-parameter distributions, etc.) are fit simultaneously to determine the largest possible number of parameters.

The main uncertainty is the temperature of the IGM which in the simulations is determined by the assumed spectral shape of the UV background and might not match the real one.

9.2.2 Proximity effect

The quasar proximity effect offers an alternate means of measuring the ionizing background using the Lyα forest. The amount of absorption in the forest generally increases with redshift. However, near a quasar, the absorption tends to decrease. This is generally attributed to locally enhanced photoionization by the quasar itself. Knowing the luminosity of the quasar (which is directly observed), the intensity of the UV background can be inferred by determining the distance out to which the quasar dominates the ionizing flux, $R_{\rm eq}$. For a given quasar luminosity, a larger proximity region indicates greater dominance by the quasar, and hence a lower background.

In a uniform IGM, the optical depth would increase with distance as

$$\tau = \frac{\tau_{\rm bg}}{1+\left(\frac{R}{R_{\rm eq}}\right)^2}\ , \tag{9.7}$$

where $\tau_{\rm bg}$ is the optical depth that would be measured if the quasar was not adding extra ionizing radiation. However, clustering in the IGM produces large variations in trasmitted flux that need to be accounted for.

The classical approach uses the observed change in the column density distribution of Lyα lines near $z \simeq z_{\rm QSO}$ to determine the local impact of the quasar. This technique is best suited for high-resolution data at redshifts where individual lines can be reliably identified ($z < 4$). At higher redshifts, or lower spectral resolution, the crowded nature of the forest makes it difficult to separate out discrete absorbers. In this case, the distribution of pixel optical depths is generally used.

Uncertainties in the measurements are rather large as a number of possible systematic effects could affect the estimates. These include:

1. imprecise determination of the quasar redshift;

2. variability in the UV emission of the quasar on the ionization timescale of the gas;

3. quasars reside in massive galaxies and gravitational clustering of clouds near them may lead to an overestimate of the background intensity by a factor up to three;

4. other sources clustered around the quasar could further enhance the local background radiation;

5. gravitational lensing might boost the apparent quasar luminosity.

9.2.3 Other methods

It is also possible to infer the UV background from ionic ratios in metal-enriched systems. The results, however, depend somewhat on model parameters, such as absorber geometry and metallicity, and may be biased towars local ionizing sources.

9.2.4 Results

In Figure 9.11 and Table 9.12 we summarize the existing measurements of the hydrogen ionization rate performed so far. Note that estimates based on the proximity effect tend to be a factor ~ 1.5 higher than those based on the flux decrement at the same redshift. This might point towards systematic errors in one of the two methods (attention has recently focussed on the effects of gas clustering around quasars).

Comparison with models for the formation of the UV background shows that:

1. The ionization rates estimated from the Lyα forest opacity are more than a factor of two larger than estimates from the integrated flux of optically bright quasars alone. This discrepancy becomes more severe with increasing redshift.

2. The measured ionization rates are in reasonable agreement with the estimates of the integrated ionizing flux from observed quasars plus a significant contribution from galaxies.

We thus conclude that the integrated ultraviolet flux arising from quasars and massive stars is likely responsible for maintaining the intergalactic diffuse gas and the Lyα forest in a highly ionized state. However, the relative

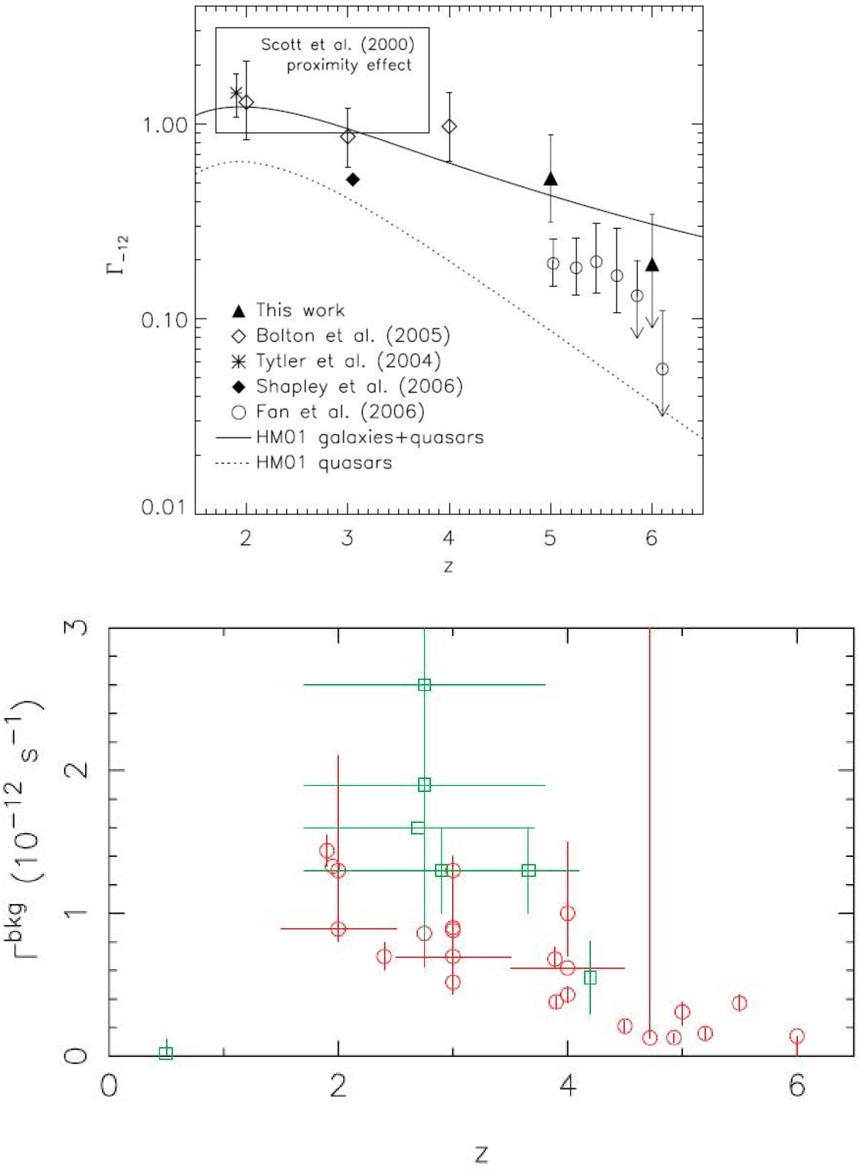
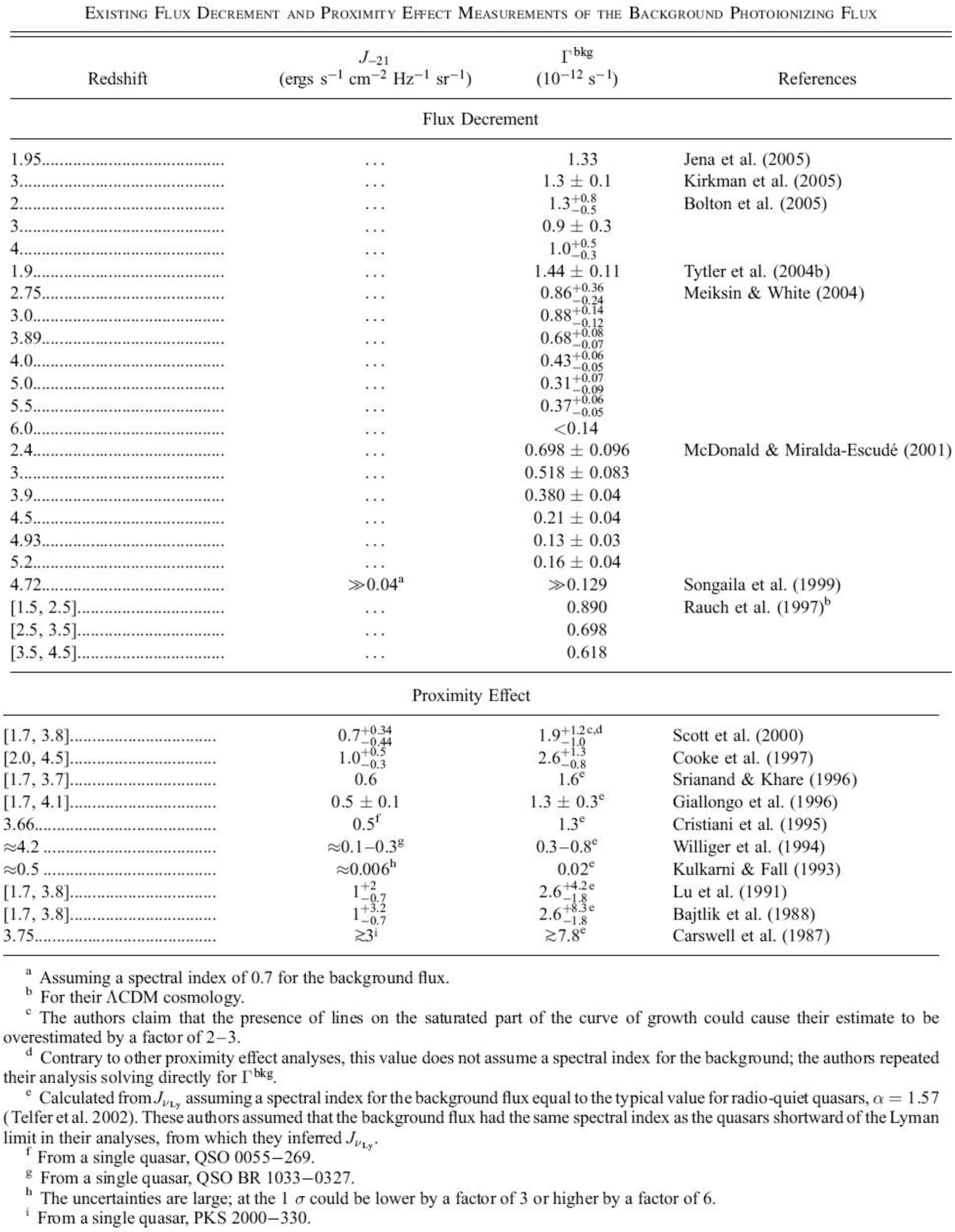


Figure 9.11: Existing measurements of the metagalactic hydrogen ionization rate, $\Gamma(\mathrm{H}_{\mathrm{I}})$. Top: The box indicates the constraints obtained from the proximity effect. All the other points are obtained matching the observed effective optical depth in simulations. The solid and dotted lines correspond to the models of Haardt & Madau (2001) for, respectively, galaxies and quasars and quasars only. Bottom: Summary of measurements obtained using the flux decrement (circles) and the proximity effect methods (squares). Horizontal bars, where present indicate the redshift range over which the measurement applies. The vertical errorbars show the reported uncertainties.



Existing Flux Decrement and Proximity Effect Measurements of the Background Photoionizing Flux

Redshift	J_{-21} (ergs s^{-1} cm^{-2} Hz^{-1} sr^{-1})	$\Gamma^{\rm bkg}$ (10^{-12} s^{-1})	References
		Flux Decrement	
1.95	...	1.33	Jena et al. (2005)
3	...	1.3 ± 0.1	Kirkman et al. (2005)
2	...	$1.3^{+0.8}_{-0.5}$	Bolton et al. (2005)
3	...	0.9 ± 0.3	
4	...	$1.0^{+0.5}_{-0.3}$	
1.9	...	1.44 ± 0.11	Tytler et al. (2004b)
2.75	...	$0.86^{+0.36}_{-0.24}$	Meiksin & White (2004)
3.0	...	$0.88^{+0.14}_{-0.12}$	
3.89	...	$0.68^{+0.08}_{-0.07}$	
4.0	...	$0.43^{+0.06}_{-0.05}$	
5.0	...	$0.31^{+0.07}_{-0.09}$	
5.5	...	$0.37^{+0.06}_{-0.05}$	
6.0	...	<0.14	
2.4	...	0.698 ± 0.096	McDonald & Miralda-Escudé (2001)
3	...	0.518 ± 0.083	
3.9	...	0.380 ± 0.04	
4.5	...	0.21 ± 0.04	
4.93	...	0.13 ± 0.03	
5.2	...	0.16 ± 0.04	
4.72	$\gg 0.04^{a}$	$\gg 0.129$	Songaila et al. (1999)
[1.5, 2.5]	...	0.890	Rauch et al. (1997)[b]
[2.5, 3.5]	...	0.698	
[3.5, 4.5]	...	0.618	
		Proximity Effect	
[1.7, 3.8]	$0.7^{+0.34}_{-0.44}$	$1.9^{+1.2\,{\rm c,d}}_{-1.0}$	Scott et al. (2000)
[2.0, 4.5]	$1.0^{+0.5}_{-0.3}$	$2.6^{+1.3}_{-0.8}$	Cooke et al. (1997)
[1.7, 3.7]	0.6	1.6^{e}	Srianand & Khare (1996)
[1.7, 4.1]	0.5 ± 0.1	1.3 ± 0.3^{e}	Giallongo et al. (1996)
3.66	0.5^{f}	1.3^{e}	Cristiani et al. (1995)
≈ 4.2	$\approx 0.1-0.3^{g}$	$0.3-0.8^{e}$	Williger et al. (1994)
≈ 0.5	$\approx 0.006^{h}$	0.02^{e}	Kulkarni & Fall (1993)
[1.7, 3.8]	$1^{+2}_{-0.7}$	$2.6^{+4.2\,{\rm e}}_{-1.8}$	Lu et al. (1991)
[1.7, 3.8]	$1^{+3.2}_{-0.7}$	$2.6^{+8.3\,{\rm e}}_{-1.8}$	Bajtlik et al. (1988)
3.75	$\gtrsim 3^{i}$	$\gtrsim 7.8^{e}$	Carswell et al. (1987)

[a] Assuming a spectral index of 0.7 for the background flux.

[b] For their ΛCDM cosmology.

[c] The authors claim that the presence of lines on the saturated part of the curve of growth could cause their estimate to be overestimated by a factor of 2–3.

[d] Contrary to other proximity effect analyses, this value does not assume a spectral index for the background; the authors repeated their analysis solving directly for $\Gamma^{\rm bkg}$.

[e] Calculated from $J_{\nu_{\rm Ly}}$ assuming a spectral index for the background flux equal to the typical value for radio-quiet quasars, $\alpha = 1.57$ (Telfer et al. 2002). These authors assumed that the background flux had the same spectral index as the quasars shortward of the Lyman limit in their analyses, from which they inferred $J_{\nu_{\rm Ly}}$.

[f] From a single quasar, QSO 0055−269.

[g] From a single quasar, QSO BR 1033−0327.

[h] The uncertainties are large; at the 1 σ could be lower by a factor of 3 or higher by a factor of 6.

[i] From a single quasar, PKS 2000−330.

Figure 9.12: Summary of the measurements of $\Gamma(\mathrm{H_I})$ performed using the flux decrement and the proximity effect methods.

importance of these two populations is still somewhat uncertain. Moreover, it is unclear whether all the souces responsible for the ionization state of the IGM are presently accounted for by magnitude-limited surveys (the contribution of fainter, still undetected objects might be substantial).

Chapter 10

The Lyα forest: theory

Soon after the discovery of the Lyα forest (and even before it) a number of models have been proposed to explain Lyα absorption in the spectra of background sources. The very first generation of models considered absorption from

1. gas clouds ejected from the host quasar with enormous energy;

2. halos of galaxies and galaxy clusters.

Observational evidence soon pointed away from these ideas. In the 1980s, attention was shifted to overdense intergalactic clouds of gas. The fact that Lyα absorption is seen throughout a large redshift interval (which corresponds to a large fraction of cosmic time) implies that either the clouds live for billions of years or that they are rapidly re-formed after they are dissolved. The key question then was "what keeps the clouds together"? Two competing scenario emerged:

Pressure-confined clouds, where a two-phase IGM was postulated with a hot, tenuous intercloud medium in pressure equilibrium with the cooler and denser Lyα clouds.

Dark-matter-confined clouds, where absorption is due to photoionized gas clouds condensed in the potential wells of small dark-matter halos (minihalos).

The absence of the Gunn-Peterson effect and difficulties with reproducing the observed column density distribution made the first model unpopular. On the other hand, the size of the minihalos in the latter model (~ 10 kpc) was too small with respect to the coherence length of Lyα systems derived from lensed quasars (~ 100 kpc). Moreover, the advent of hydrodynamical simulations indicated that most baryonic matter may not have settled in virialized dark-matter halos at redshift $z \sim 2 - 3$.

Over the past ten years analytical work and in particular hydrodynamical simulations of cosmic structure formation have gradually led to a new picture of the Lyα forest and the IGM in general. In this Chapter we explore the currently favoured model.

10.1 Gravitational instability

Consider an ideal, non-relativistic fluid. The evolution of its density, $\rho(\mathbf{r},t)$, pressure, $p(\mathbf{r},t)$, and velocity, $\mathbf{u}(\mathbf{r},t)$, fields is described by the system of partial differential equations:

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \nabla_{\mathbf{r}} \cdot \rho\, \mathbf{u} &= 0 && (10.1) \\ \frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla_{\mathbf{r}})\mathbf{u} &= -\nabla_{\mathbf{r}}\Phi - \frac{1}{\rho}\nabla_{\mathbf{r}} p && (10.2) \\ \nabla^2_{\mathbf{r}}\Phi &= 4\pi G \rho\,. && (10.3) \end{aligned}$$

Equation (10.1) expresses the conservation of mass and is known as the continuity equation. It states that the rate at which the mass within a closed surface changes has to match the mass flow through the surface. Equation (10.2) states the conservation of linear momentum and is known as the Euler equation. It gives the equation of motion for a given fluid element: the left-hand side is the acceleration and the right-hand side is the force per unit mass. Finally, equation (10.3) tells how the gravitational potential is generated by the existing mass density and is known as the Poisson equation.

We want to follow the evolution of small fluctuations in a fluid that pervades an expanding universe with scale factor a. It is convenient to use comoving coordinates $\mathbf{x} = \mathbf{r}/a$ and to introduce new variables as the density contrast, δ defined as $\rho = \rho_{\mathrm{b}}\,(1+\delta)$ (with ρ_{b} the mean, or background, density), the proper peculiar velocity, $\mathbf{v} = \mathbf{u} - \dot{a}\mathbf{x} = a\dot{\mathbf{x}}$, and the peculiar gravitational potential, $\phi = \Phi + a\ddot{a}x^2/2$. In terms of these new variables the fluid equations become:

$$\begin{aligned} \frac{\partial \delta}{\partial t} + \frac{1}{a}\nabla \cdot (1+\delta)\,\mathbf{v} &= 0 && (10.4) \\ \frac{\partial \mathbf{v}}{\partial t} + \frac{1}{a}(\mathbf{v}\cdot\nabla)\mathbf{v} + \frac{\dot{a}}{a}\mathbf{v} &= -\frac{1}{a}\nabla\phi - \frac{1}{a\,\rho_{\mathrm{b}}}\nabla p && (10.5) \\ \nabla^2\phi &= 4\pi G \rho_{\mathrm{b}} a^2 \delta\,. && (10.6) \end{aligned}$$

Note that the differential operator ∇ is now taken with respect to comoving coordinates while it was taken with respect to the physical ones in the original equations (10.1), (10.2) and (10.3).

Observations of the CMB show the presence of temperature fluctuations with $\Delta T/T \sim 10^{-5}$. This suggests that the structure we observe in the universe formed out of small primordial density fluctuations. Therefore, at early times, we expect that $\delta, \mathbf{v}$ and ϕ are all $\mathcal{O}(\epsilon)$ with $\epsilon \ll 1$. it is thus reasonable to ignore all terms $\mathcal{O}(\epsilon^n)$ with $n > 1$ in the evolutionary equations. The linearized system of equations is then:

$$\begin{aligned} \frac{\partial \delta}{\partial t} + \frac{1}{a} \nabla \cdot \mathbf{v} &= 0 && (10.7) \\ \frac{\partial \mathbf{v}}{\partial t} + \frac{\dot{a}}{a} \mathbf{v} &= -\frac{1}{a} \nabla \phi - \frac{1}{a\, \rho_{\rm b}} \nabla p && (10.8) \\ \nabla^2 \phi &= 4\pi G \rho_{\rm b} a^2 \delta \, . && (10.9) \end{aligned}$$

Now take the divergence of the linearized Euler equation, use the linearized continuity equation to replace $\nabla \cdot \mathbf{v}$ and, finally, the Poisson equation to replace $\nabla^2 \phi$. Eventually one obtains a second-order differential equation for the density contrast:

$$\frac{\partial^2 \delta}{\partial t^2} + 2 \frac{\dot{a}}{a} \frac{\partial \delta}{\partial t} = 4\pi G \rho_{\rm b} \delta + \frac{1}{a^2 \rho_{\rm b}} \nabla^2 p \, . \qquad (10.10)$$

Note that this equation is local, i.e. it only depends on conditions at one point in space.

10.1.1 Collisionless fluids

If the particles of the fluid do not interact with any other force but gravity (collisionless fluid), the pressure term on the right-hand side of equation (10.10) should be erased. This, for instance, applies to the dark matter which is believed to be made of very weakly interacting particles. In this case,

$$\frac{\partial^2 \delta}{\partial t^2} + 2 \frac{\dot{a}}{a} \frac{\partial \delta}{\partial t} - 4\pi G \rho_{\rm b} \delta = 0 \, . \qquad (10.11)$$

In full generality, the solution is given by the linear superposition of two modes, one growing with time and the other decaying (hereafted indicated by the subscripts + and -, respectively):

$$\delta(\mathbf{x}, t) = \delta_+(\mathbf{x}) D_+(t) + \delta_-(\mathbf{x}) D_-(t) \, . \qquad (10.12)$$

The functions D_+ and D_- are the independent solutions of the second order, ordinary differential equation $\ddot{D} + 2(\dot{a}/a)\dot{D} - 4\pi G \rho_{\rm b} D = 0$, and take different functional forms depending of the assumed background cosmology. In a universe dominated by a cosmological constant and by matter:

$$D_+(a) \propto \Theta(a) \int_0^a \frac{dx}{x^3\, \Theta^3(x)} \, , \qquad D_-(a) \propto \Theta(a) \, , \qquad \Theta(a) = \frac{H(a)}{H_0} \, . \quad (10.13)$$

10.1.2 Jeans length

In order to describe the evolution of the baryon density we need to understand what happens when the pressure term in equation (10.10) is not negligible. An equation of state for the gas, $f(\rho, p) = 0$, is required to close the system of equations. This can be specified by introducing the sound speed, $c_s = (\partial p / \partial \rho)^{1/2}$, so that we can write $\nabla^2 p = c_{\rm s}^2 \, \nabla^2 \rho = c_{\rm s}^2 \, \rho_{\rm b} \, \nabla^2 \delta$.

Let us now write the density contrast as a Fourier series

$$\delta(\mathbf{x}, t) = \sum_{\mathbf{k}} \tilde{\delta}(\mathbf{k}, t) \, \exp{(i \, \mathbf{k} \cdot \mathbf{x})} \tag{10.14}$$

thus obtaining for each Fourier mode of proper wavelength $\lambda = 2\pi a(t)/k$

$$\frac{\partial^2 \tilde{\delta}}{\partial t^2} + 2 \, \frac{\dot{a}}{a} \, \frac{\partial \tilde{\delta}}{\partial t} = \left(4\pi G \rho_{\rm b} - \frac{k^2 c_{\rm s}^2}{a^2} \right) \tilde{\delta} \, . \tag{10.15}$$

The source term on the right-hand side vanishes for $k_{\rm J} = 4\pi G a^2 \rho_{\rm b} / c_{\rm s}^2$ corresponding to the wavelength

$$\lambda_{\rm J} = \frac{2\pi a(t)}{k_{\rm J}} = c_{\rm s} \left(\frac{\pi}{G \rho_{\rm b}} \right)^{1/2} \tag{10.16}$$

which is often referred to as the Jeans length. Density perturbations with wavelength $\lambda < \lambda_{\rm J}$ propagate as acoustic waves and are slowly damped by the Hubble expansion (note that, in this case, equation (10.15) is analogous to that of a damped oscillator). On the other hand, perturbations with $\lambda > \lambda_{\rm J}$ are gravitationally unstable (i.e. solved by the superposition of a growing and a decaying mode, with no oscillatory terms). This means that their self-gravity exceeds opposing forces (the internal gas-pressure gradient) and the perturbations collapse. Fourier modes with $\lambda \gg \lambda_{\rm J}$ follow the evolution discussed in the previous section. One thus expects that density fluctuations in the gas and in the dark matter have the same structure for wavelengths $\lambda \gg \lambda_{\rm J}$.

10.2 Numerical simulations

When the density contrast approaches unity, the linear approximation breaks down and one has to resort to numerical simulations to study the subsequent evolution. In general,

1. dark-matter is simulated with N-body methods, where the fluid is discretized into a finite set of particles of a given mass;

2. two different approaches are commonly used to follow the hydrodynamics of the gas

(a) Smoothed Particle Hydrodynamics (SPH): is a Lagrangian technique where the gas is represented by a set of particles and the thermodynamic quantities are computed by averaging over a fixed number of neighbouring particles.

(b) Adaptive Mesh Refinement (AMR): is a Eulerian scheme where the computational volume is covered with a hierarchy of completely nested grid patches and the resolution is increased where required. The fluid equations are solved using finite difference methods on the grids (high-order advection and shock-capturing schemes are required to obtain an accurate solution).

Results, where comparable, agree rather well but some discrepancies are found mainly because of the different shock-capturing abilities of the two methods. Moreover, in general, Eulerian codes have higher spatial resolution in underdense regions (corresponding to the lowest column density Lyα forest) whereas Lagrangian codes better resolve collapsed regions like mini-halos or galactic halos (corresponding to damped Lyα systems and metal systems).

10.2.1 Non-linear structure formation

Numerical simulations (but some insight can also be obtained with analytical methods like the Zel'dovich approximation) show that gravitational instability of small primordial density fluctuations in a universe dominated by cold dark matter and a cosmological constant ends up producing a complex foamy (or sponge like) structure (see Figure 10.1). As a far as the dark-matter is concerned:

- Matter flows out of the underdense regions which become more and more empty with time. At the redshifts of interest for the study of the Lyα forest, underdense regions (sometimes called voids) occupy most of the volume.

- The "voids" ($\delta < -0.8$) are surrounded by thin sheet-like "walls" of matter with typical overdensities ranging between of -0.5 and 0.

- Matter moves along the sheets and tend to concentrate along unidimensional structures, the filaments, located where two walls intersect. Filaments have characteristic density contrasts a few$< \delta < 10$.

- Matter flows along the filaments until it accretes onto nearly one dimensional structures (compact, roughly spherical clumps): the dark-matter halos.

- Very massive dark-matter halos (corresponding to galaxy clusters) lie at the intersection of filaments. Less massive halos (corresponding to galaxies and dwarf galaxies) trace the structure described above and are evenly distributed in voids, sheets and filaments proportionally to the local density.

Consider two neighbouring fluid elements (that with a little abuse of terminology we will call particles). At early times (when the universe is basically homogeneous) their spatial separation increases following the Hubble flow. In a smooth universe this would persist at all times. In the presence of density fluctuations, however, gravity creates a relative acceleration between the particles. Depending on the spatial location, the acceleration might add to the Hubble expansion or act in the opposite direction. The prolonged action of the acceleration may then turn the expansion into a collapse driving the particles closer and closer together. This gravitational collapse would then ultimately lead to the phenomenon of orbit crossing where the particles cross each other with a finite velocity. The acceleration would then reverse and lead to the formation of a gravitationally bound structure. Chaotic interactions in the rapidly varying potential well then cause the dynamical relaxation and virialization of the collapsed objects. Sheets are structures that have collapsed along one spatial dimension, filaments along two, and halos along three.

What happens to the dynamically sub-dominant baryonic gas? Simulations show that baryons trace the dark-matter distribution well on scales above the Jeans length ($\sim 100\,h^{-1}$ kpc). However, due to the action of pressure, they are much more smoothly distributed than the dark matter on smaller scales.

Another difference is that baryonic gas is self-interacting and orbit crossing is not possible (if you run against a wall, you will not pass through it as if you and the wall were made of collisionless dark matter). Accretion shocks are thus produced at the external layers of cosmic structures where the gas is compressed and shock heated to high temperatures. While the largest cosmic structures form, the IGM is heated by gravitational shocks that efficiently propagate from the collapsing regions to the surrounding medium. This process converts gravitational potential energy into heat.

10.2.2 Simulating the IGM

In order to set up a simulation of the intergalactic medium one has to

1. Choose a cosmological model by specifying a set of cosmological parameters like $\Omega_{\rm m}, \Omega_\Lambda, \Omega_{\rm b}, h$, etc.

2. Choose a set of initial conditions (density fluctuations) whose statistical properties are compatible with observations of the CMB temperature fluctuations.

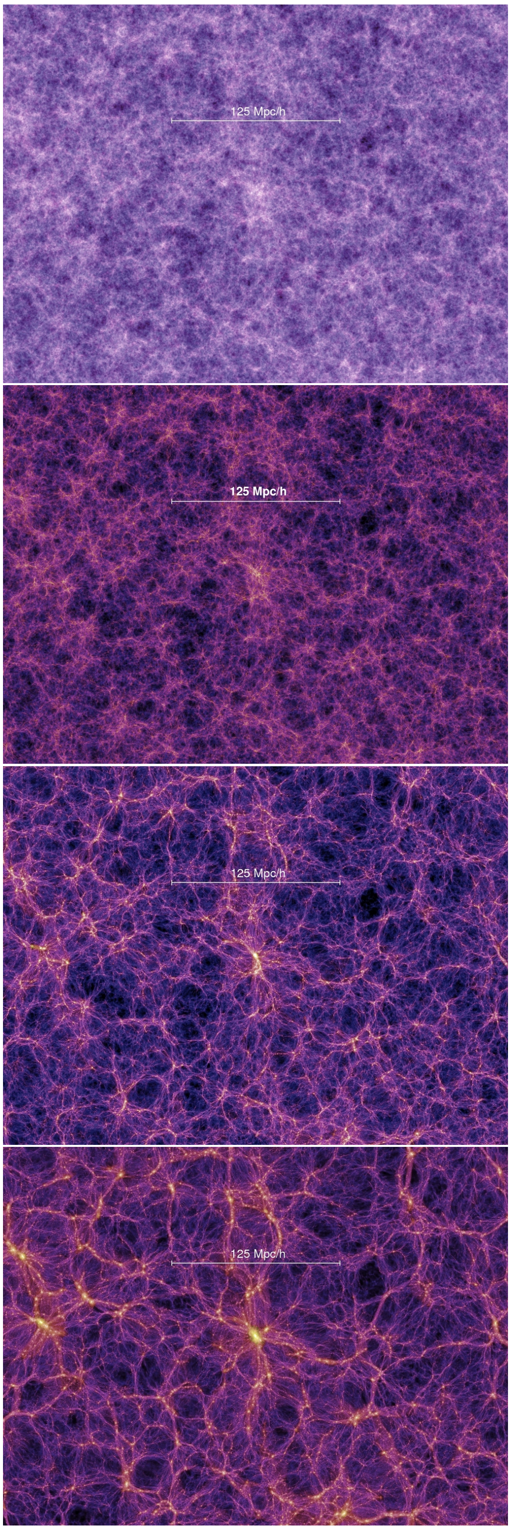


Figure 10.1: Output of a N-body simulation for structure formation in the universe at different redshifts: 18.3, 5.7, 1.4, 0 (from top to bottom). The slices have been obtained by projecting all the dark-matter particles within a slice with a thickness of $15\,h^{-1}$ Mpc onto the plane of the page.

3. Select the size of the volume that one wants to simulate and the spatial and mass resolution of the simulation (this also depends on the computing resources that are available).

4. Specify a rule for star formation out of the cold gas (this happens on length scales that are not resolved by the simulations so it has to be implemented in an approximate way using a simple "sub-grid" model).

5. Specify a rule for the (mechanical, thermal and chemical) feedback effects that star formation has on the surrounding gas (also happening on sub-grid scales). Remember that metal enrichment modifies the cooling properties of the gas.

6. Specify the evolution and the spectrum of the ionizing background $J_\nu(z)$.[1]

Fortunately, the uncertain points 4) and 5) have a marginal impact on the physics of the Lyα forest while they are crucially important to describe galaxy formation and metal absorption systems.

In Figure 10.2 we show the distribution of dark matter (left), gas (centre) and stars (right) extracted from a state-of-the-art simulation. Note that all the three components trace the same large-scale structure and that the IGM is not at all uniformly distributed.

10.3 Results

10.3.1 Phases of the IGM

We start our quantitative analysis of the output from numerical simulations by looking at the relation between the temperature and the density of the gas. In Figures 10.3 and 10.4 we show the results by two different groups. The gas mostly resides in three phases:

Photoionized diffuse gas. The gas that is photoionized by the cosmic UV background lies at relatively low densities ($\delta < 100$) and temperatures ($T < 10^{4.5-5}$ K). The temperature of the photoionized gas is primarily determined by the balance between adiabatic cooling (due to the gravitational evolution of density perturbations), photoionization heating, and recombination cooling. A tight temperature-density relation (sometimes called "effective equation of state") exists as a result

[1] Note that, in most cases, $J_\nu(z)$ is not self-consistently computed from the star-formation history in the simulation box (which is usually rather small in cosmological terms and would not contain a single quasar on average). Rather, a model based on the observed star-formation history and quasar abundance is generally used. Also, the radiative transfer problem is often simplified and the UV background is assumed to be uniform.

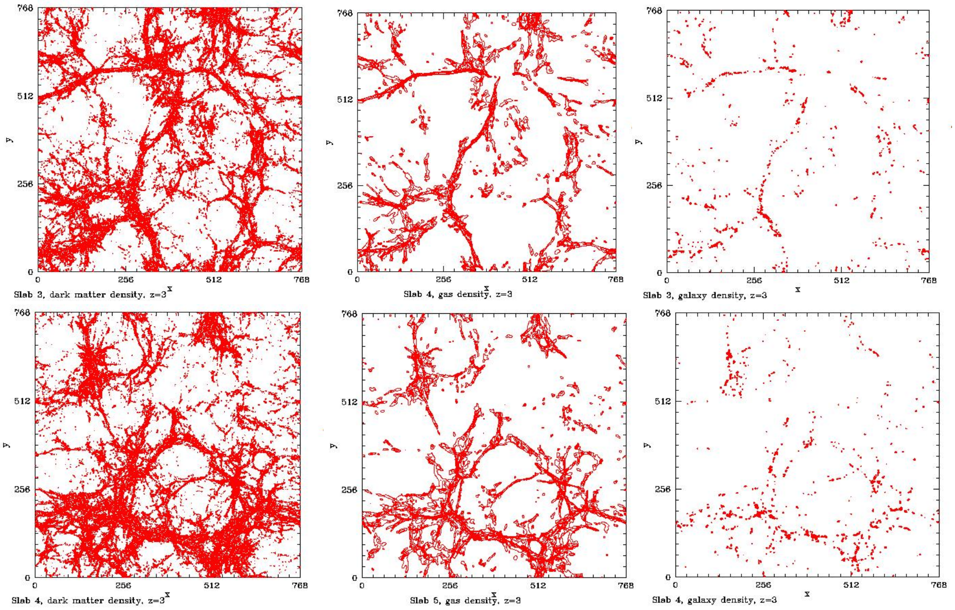


Figure 10.2: Spatial distribution at $z = 3$ of dark matter (left), gas (centre) and stars (right). Data are extracted from a cubic simulation box of $(25\,h^{-1}$ Mpc$)^3$ resolved into 768^3 computational cells. The simulation includes star formation, energy feedback from supernova explosions, ionizing radiation from massive stars and metal recycling due to supernova and galactic winds.

of these three processes:

$$T = T_0 \left(\frac{\rho}{\bar{\rho}}\right)^{\gamma-1} = T_0 \; (1+\delta)^{\gamma-1} \tag{10.17}$$

where T_0 (the temperature for gas at mean density) and γ depend on cosmology and the past ionization history of the gas. The index γ is expected to vary from near unity at high redshift to ~ 1.6 at low z. In general $\gamma = 1.3 \pm 0.3$ is a good approximation at all redshifts of interest. The parameter T_0 ranges in the interval $11,000 < T_0 < 18,000$ K. Fiducial values are $T_0 = 11,200 \pm 5,000$ K at $z = 2$, $T_0 = 17,800 \pm 5,000$ K at $z = 3$ (when quasars make the spectrum of the UV background harder) and $T_0 = 12,500 \pm 5,000$ at $z = 4$.

Shock heated gas. Simulations predict that gas compressed and heated by shocks can reach temperatures of 10^8 K in rich clusters of galaxies, while filaments and mildly overdense regions are heated to temperatures in the range $10^5 - 10^7$ K. Note that at these temperatures the IGM is collisionally ionized and becomes transparent to Lyα radiation. The fraction of gas in the shock-heated phase increases at low redshifts.

Condensed gas. Gas in dark-matter halos that had the time to cool and condense is found at high densities and low temperatures ($T < 10^{4.5}$ K). This is the gas that, with further cooling, can form stars.

10.3.2 Origin of the Lyα forest

It is possible to shed some light on the origin of the Lyα forest by drawing an imaginary line of sight across a simulation box and computing the Lyα optical depth as a function of redshift (including peculiar motions, see Figure 10.5). Hydrodynamic simulations show that the smoothly varying density of the IGM gives rise to a fluctuating optical depth in redshift space. Many of the optical-depth maxima can be fitted quite accurately with Voigt profiles. In particular, the low-column-density ($N_{\rm H_I} < 10^{14.5}$ cm^{-2}) Lyα forest at redshift $z > 2$ is produced by photoionized gas in filamentary and sheet-like structures. Higher column density lines (and metal lines) occur where the line of sight intersects a dark-matter halo with condensed gas.

Photoionization equilibrium

At the relevant redshifts, the gas making up the Lyα forest is in photoionization equilibrium. In this case, as we showed in the previous class, $n_{\rm H_I} = n_{\rm e} n_{\rm p} \alpha(T)/\Gamma$ which implies $n_{\rm H_I} \propto (1+z)^6 \, (\Omega_{\rm b} h^2)^2 \, (1+\delta)^2 \, \alpha(T)/\Gamma$.

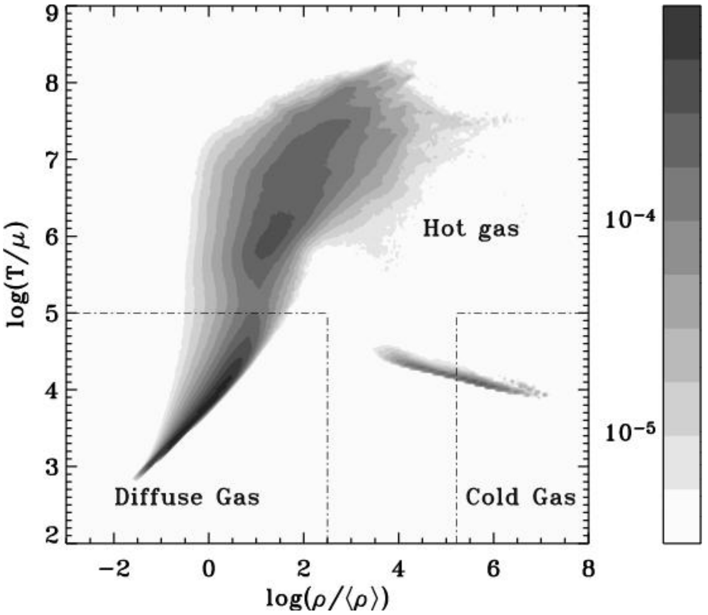


Figure 10.3: Temperature-density relation at $z = 0$ for cosmic gas from a numerical simulation. Colors indicate the baryon mass fraction at given ρ and T. Note the presence of a tight power-law relation at low-densities and temperatures.

Given that, at the typical temperatures of the photoionized IGM, the hydrogen recombination rate is proportional to $T^{-0.7}$, one finally obtains

$$n_{\mathrm{H_I}} \propto \frac{(1+z)^6\,(\Omega_\mathrm{b}h^2)^2\,(1+\delta)^2}{T^{0.7}\,\Gamma(z)} \ . \tag{10.18}$$

Hubble-flow broadening of the Lyα forest

Lyα forest lines in quasar spectra have typical widths of 20-50 km s^{-1}. Low-column-density absorbers in cosmological simulations are large, diffuse structures that are still expanding with residual Hubble flow. In fact, typical marginally saturated lines ($N_{\mathrm{H_I}} \sim 10^{14}$ cm^2) arise in gas whose density is a few times the cosmic mean or less. Weak lines ($N_{\mathrm{H_I}} < 10^{13}$ cm^2) often occur at local maxima that lie below the global mean density. As a consequence of this, the Hubble flow across the spatially extended absorber is usually the dominant contribution to the width of its associated absorption line. Thermal broadening is unimportant over most of the spectrum, and peculiar velocities tend to make absorption features narrower rather than broader (see Figure 10.6). Note, however, that some low-column density lines at high z arise in shock-heated gas and do are thermally broadened.

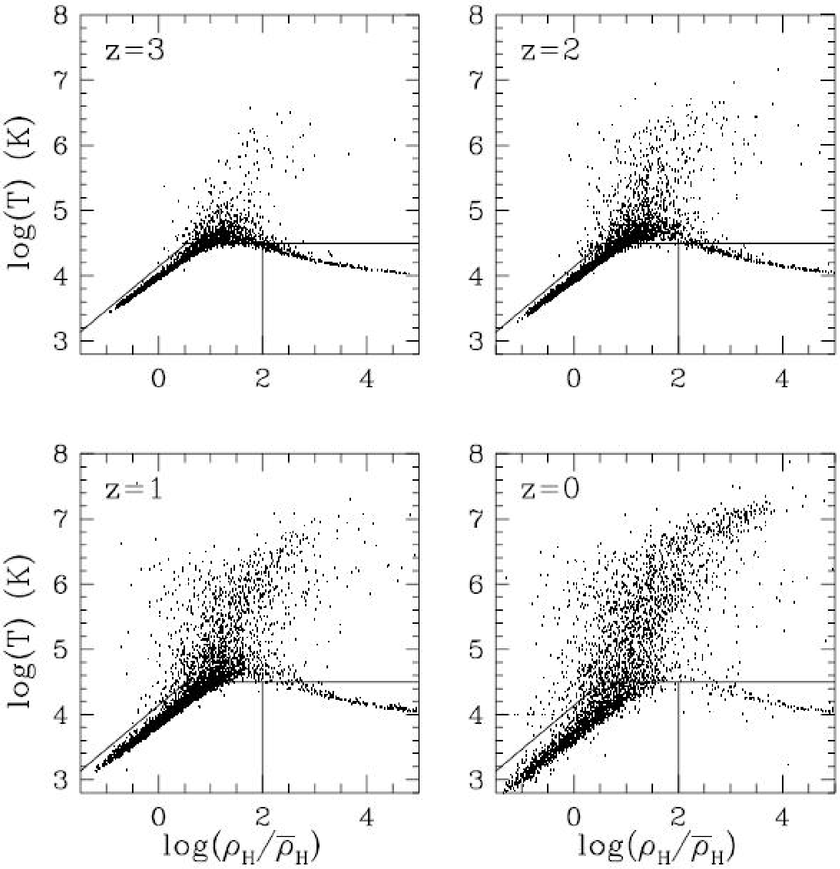


Figure 10.4: Temperature-density relation for cosmic gas from a numerical simulation (points). The solid lines indicate an approximate separation between diffuse photoionized gas (bottom left), condensed gas (bottom right) and shock-heated gas (all the rest). Note that the fraction of shock-heated gas increases at low redshifts.

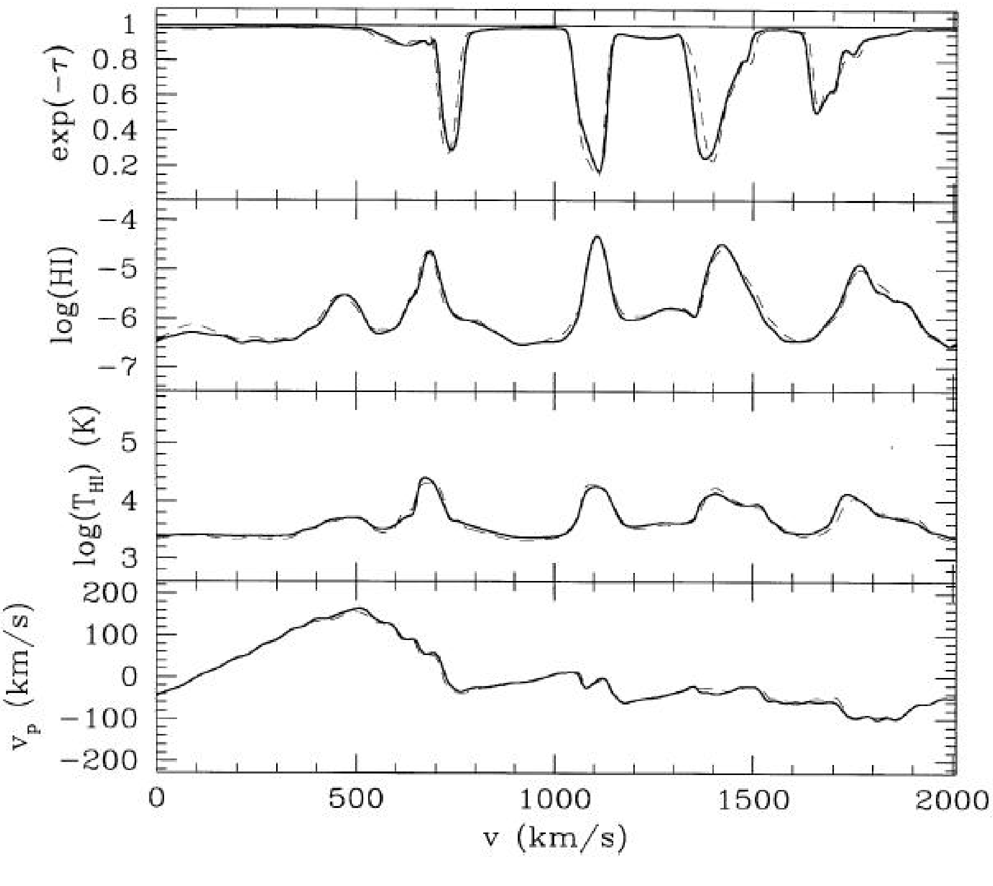


Figure 10.5: Absorption spectrum along a random line of sight through a simulation box and the corresponding H_I fraction, temperature and peculiar velocity distributions (from top to bottom).

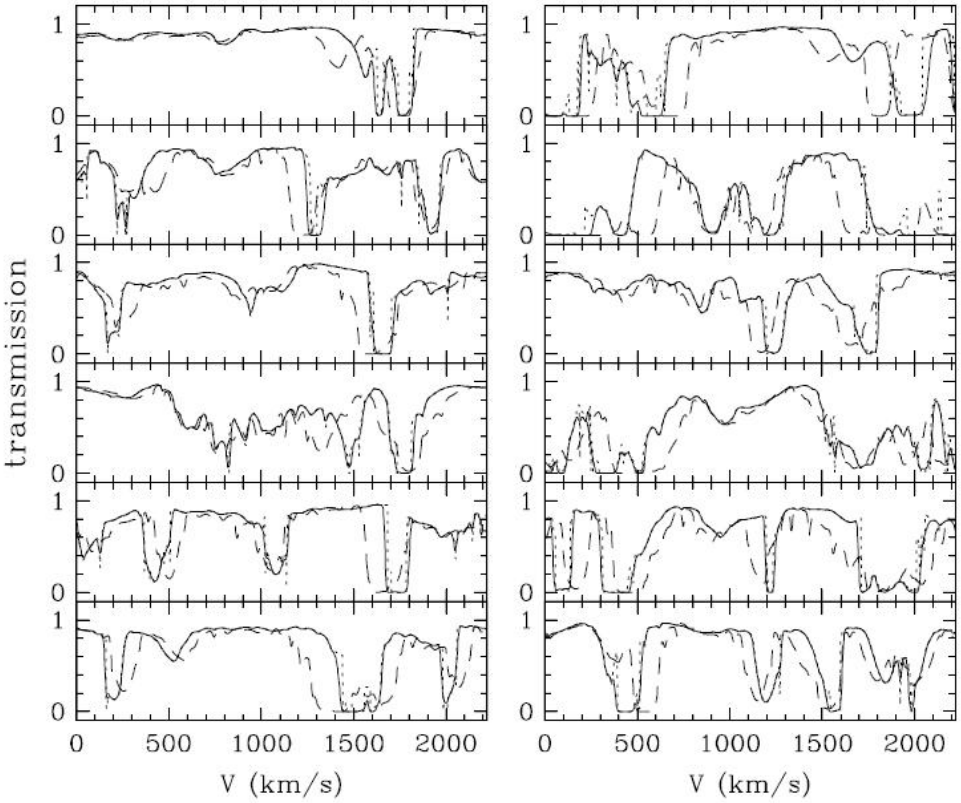


Figure 10.6: The solid lines show spectra along 12 randomly chosen lines of sight through a simulation at $z = 3$. Dotted and dashed lines show spectra along the same lines of sight with no thermal broadening and no peculiar velocities, respectively.

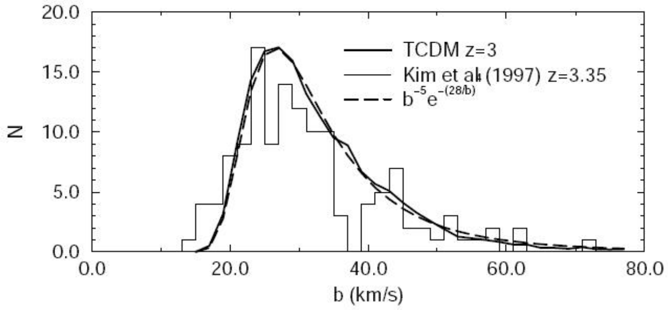


Figure 10.7: The distribution of b-parameters from a simulation of the IGM at $z = 3$ (solid line) is compared with observational data (histogram).

The resulting distribution of b-parameters obtained by fitting the simulated absorption lines with Voigt profiles is compared with observational data in Figure 10.7. The agreement is very good.

10.3.3 Redshift and equivalent-width distribution of the absorbers

Hydrodynamical simulations give a fairly good match to the observed redshift distribution of the absorbers, dN/dz (Figure 10.8). In particular, independently of the exact cosmological model, they show a transition from rapid evolution to low evolution at $z \sim 1.7$. The crucial ingredient to this success is the redshift evolution of the ionizing background. The characteristic break in the rate of evolution of the number of lines below a redshift ~ 1.7 as observed by the HST can be explained by the decrease in the intensity of the ionizing background, itself a consequence of the rapid decline in the abundance of quasars and young stars towards lower redshifts. The decline in the photoionization rate counters the decline in the recombination rate caused by the expansion of the universe, and the combination of the two effects leads to a slow evolution. Gravitational growth of structure has a subsidiary effect, reducing dN/dz as gas moves from lower density regions into collapsed structures that have smaller cross sections for absorption. This transformation of the underlying structure has an important effect on the evolution of the equivalent width distribution, dN/dW, which steepens towards low redshift (Figure 10.9).

10.3.4 The fluctuating Gunn-Peterson approximation

The physics that governs the unshocked IGM leads to a tight correlation between the neutral hydrogen density and the underlying gas and dark matter overdensity. Some time ago we have derived the the Gunn-Peterson formula

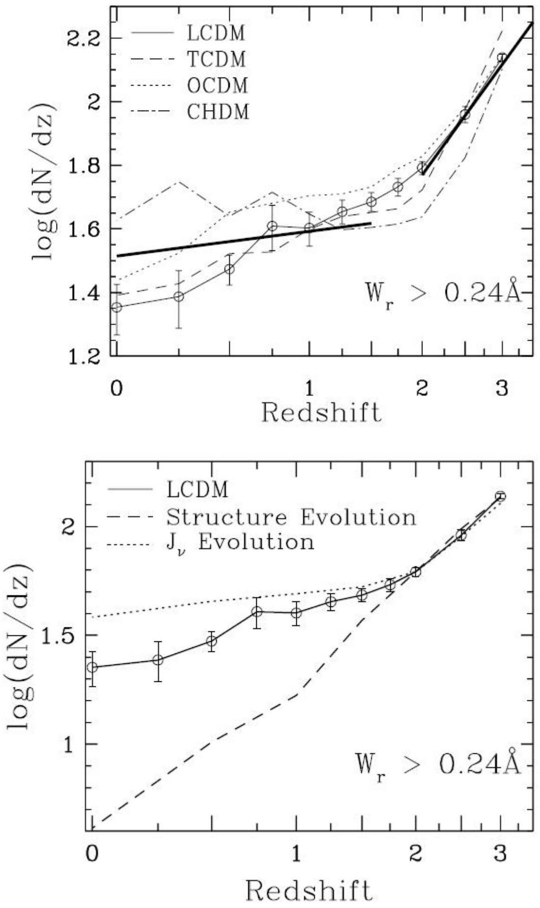


Figure 10.8: Top: Evolution of dN/dz in various simulated cosmological models (thin lines). The thick solid lines show fits to observational data. All cosmologies show a change in the evolution of dN/dz at $z \sim 2$. Bottom: The effect on dN/dz due to structure evolution only (dashed) and due to evolution of the UV background only (dotted). This shows that the change at $z \sim 2$ in the complete simulation (solid) is mainly driven by changes in J_ν.

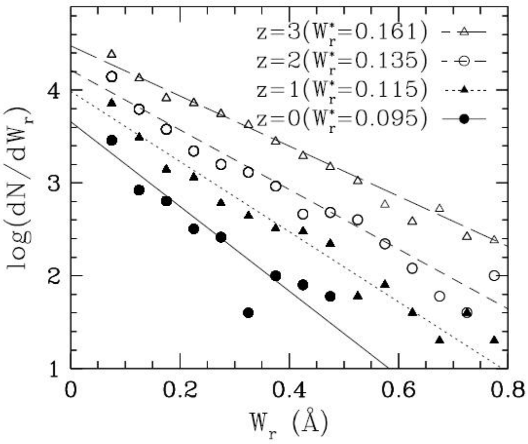


Figure 10.9: Equivalent width distribution of the simulated Lyα forest at $z = 3, 2, 1, 0$.

- equation (7.6):

$$\tau = \frac{\pi e^2}{m_e c} f_\alpha \, \lambda_\alpha \, \frac{n_{\mathrm{H_I}}}{H(z)} \ . \tag{10.19}$$

Assuming that

1. all gas lies on the temperature-density relation eq. (10.17);

2. thermal broadening and collisional ionization can be ignored;

3. photoionization equilibrium holds;

we can directly combine equations (10.17), (10.18) and (10.19) to obtain

$$\tau = \tau_0 \frac{(1+z)^6 \, (\Omega_b h^2)^2}{T_0^{0.7} \, H(z) \, \Gamma(z)} \, (1+\delta)^\beta \ , \tag{10.20}$$

where $\beta = 2 - 0.7(\gamma - 1)$ is determined by the slope of the effective equation of state, and asymptotically approaches the value 1.6. For a flat universe at redshifts $z > 2$, the Hubble parameter can be approximated by $H(z) \simeq H_0 \, \Omega_m^{1/2} \, (1+z)^{3/2}$. Thus, given the correct density distribution and effective equation of state, the Lyα optical depth should scale according to the parameter combination $\mu = \Omega_b^2 h^3 T_0^{-0.7} \Omega_m^{-0.5} \Gamma^{-1}$. This is what we used in the previous class to derive Γ from the observational data on the flux decrement.

In summary, the Lyα forest of absorption lines can be seen as a continuous, non-linear map of the underlying density field. Plugging in all the physical constants, one gets $\tau = A(z)\,(1+\delta)^{\beta}$ with

$$\begin{aligned} A(z) &= 0.946 \left(\frac{1+z}{4}\right)^{6} \left(\frac{\Omega_{\rm b} h^2}{0.0125}\right)^{2} \left(\frac{T_0}{10^4 K}\right)^{-0.7} && (10.21)\\ &\times \left(\frac{\Gamma}{10^{-12}\,{\rm s}^{-1}}\right)^{-1} \left(\frac{H(z)}{100\,{\rm km\ s}^{-1}\ {\rm Mpc}^{-1}}\right)^{-1} && (10.22) \end{aligned}$$

which is known as the *fluctuating Gunn-Peterson approximation*. Simulations show that the approximation is very reasonable at low densities but it breaks down when $\delta > 10$.