LECTURE NOTES OF OBSERVATIONAL ASTROPHYSICS - PROF. SERGIO ORTOLANI

Patrizia Bussatori, Fabrizio Muratore

January 2021

Premise

This book contains the notes taken from the lessons of the observational astrophysics course held by prof. Sergio Ortolani for the master degrees in Cosmology and Astrophysics, in the academic year 2020/2021. As these are not official professor approved notes, we do not guarantee the correctness of the contents. Indeed, any error reporting is welcome! Enjoy the reading!

Patrizia Bussatori & Fabrizio Muratore

Contents

Chapter 1

Basic astronomical notions

1.1 Introduction

Here there are some important definitions.

- The sidereal time is defined as the hour angle of the gamma point and it grows with time since there is a difference of almost 4 minutes a day as a consequence of the motion of the Earth around the Sun. In fact the Earth completes a whole turn around the Sun in about 365 days and if we divide the minutes into a day ($24 \times 60 = 1440$) for 365 ($1440 \div 365$), we obtain 3.95 minutes per day.
- Right ascension is the celestial equivalent of terrestrial longitude. Both right ascension and longitude measure an angle from a primary direction (a zero point) on an equator. Right ascension is measured from the Sun at the March equinox i.e. the First Point of Aries, which is the place on the celestial sphere where the Sun crosses the celestial equator from south to north at the March equinox and is currently located in the constellation Pisces. It is customarily measured in hours, minutes, and seconds, with 24h being equivalent to a full circle.
- The hour angle is the angle between an observer's meridian (a great circle passing over his head and through the celestial poles) and the hour circle (any other great circle passing through the poles) on which some celestial body lies. This angle, when expressed in hours and minutes, is the time elapsed since the celestial body's last transit of the observer's meridian.

The relation between sidereal time, right ascension and hour angle is:

$$ts = H + RA \tag{1.1}$$

where we see that the hour angle progressively grows as the sidereal time grows.

Therefore, to know how far the direction of the celestial body is form the meridian (origin oh the hour angle), just check the time sidereal and the object's coordinates.

1.2 Astronomical coordinates

The equatorial coordinate system is the preferred coordinate system to point objects on the celestial sphere indeed equatorial coordinates are independent of the observer's location and the time of the observation. This means that only one set of coordinates is required for each object, and that these same coordinates can be used by observers in different locations and at different times.

The equatorial coordinate system is basically the projection of the latitude and longitude coordinate system we use here on Earth, onto the celestial sphere. By direct analogy, lines of latitude become lines of declination (Dec; measured in degrees, arcminutes and arcseconds) and indicate how far north

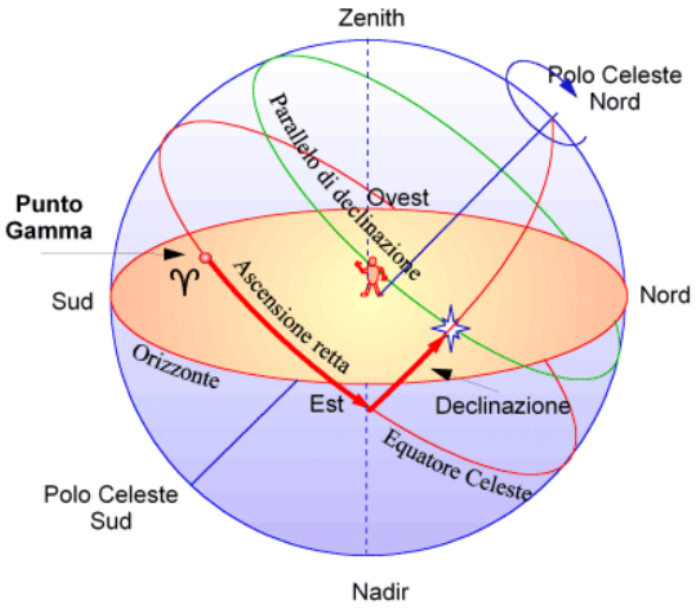


Figure 1.1: Illustration of the equatorial coordinate system.

or south of the celestial equator (defined by projecting the Earth's equator onto the celestial sphere) the object lies. Lines of longitude have their equivalent in lines of right ascension (RA), but whereas longitude is measured in degrees, minutes and seconds east the Greenwich meridian, RA is measured in hours, minutes and seconds east from where the celestial equator intersects the ecliptic (the vernal equinox), as we seen before.

The most frequent use of spherical trigonometry formulas in basic astronomy regards the position of the Sun and celestial bodies, in particular the height of the celestial bodies on the horizon in order to determine the observability of the stars.

The most used equation is the following:

$$sinh = sin\delta sin\phi + cos\delta cos\phi cosH \tag{1.2}$$

where:

- h is the height of the star on the horizon;
- δ is the declination of the star;
- ϕ is the geographical latitude;
- H is the hour angle.

Now there are three limit cases.

- If $H = 0$, $ts = RA$, the body stay at the maximum height, regardless of δ.
- If $\phi = 90°$ (at the **pole**), we obtain:
$$sinh = sin\delta \tag{1.3}$$
so $h = \delta$ and the relation is independent of the time so there are no change in altitude over the horizon.
- If $\phi = 0°$ (at the **equator**), we obtain:
$$sinh = cos\delta cos\phi cosH \tag{1.4}$$

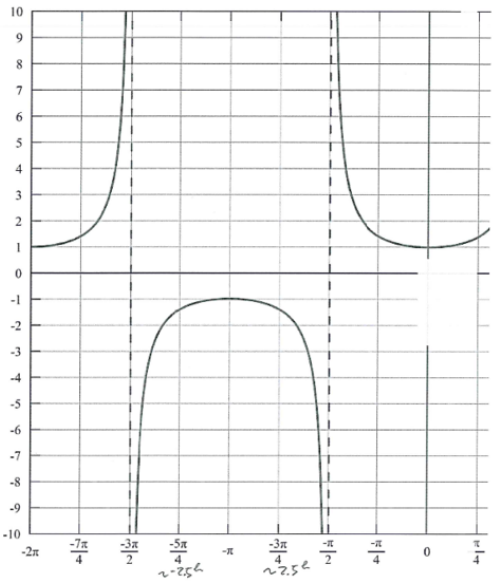


Figure 1.2: Illustration of the nomogram that shows the airmass changing during the stellar path in the sky. It is minimum at the highest point over the horizon.

so there is the dependence on the time and stars go up and down very rapidly.

Using the equation 1.2, it is possible to built a **nomogram**, a diagram representing the relations between three or more variable quantities by means of a number of scales, so arranged that the value of one variable can be found by a simple geometrical construction.

For example, a nomogram can report δ, H, ϕ and the airmass that affect the observation of the object in a certain place where is located the telescope. This is particular useful to program the best observation.

In figure 1.2 we can observe the nomogram for the cosecant law:

$$airmass = secz = 1/cosz \tag{1.5}$$

where $360° = 2\pi = 24h$. This law shows how airmass is different based on z, the zenith distance of the object. In is possible to observe that the less airmass (the best observation) corresponds to the greatest height (on the meridian) the corresponds to curvature in the graphic. Then the observational conditions make worse quickly.

Another application of the equation 1.2 in astronomy is the determination of the position of the Sun with respect to the horizon to determine the duration and the end of the night. If we define like limit of the night a negative height (under the horizon) of 18° like convention and if we know the declination of the Sun at the moment of observation, it is possible to obtain the moment of beginning e the duration of the astronomical night. The calculation of the night duration during a year can show the total number of hours available in a certain place.

The integration during a year show that the total number of night hours available vary from a maximum of about 3400 hours at the equator to a minimum of about 1700 at 84° of latitude. The difference is connected the number of hours lost during sunset and sundown (il crepuscolo), because the total time that the Sun stays up and down the horizon must be equal in every point of the Earth even if (ancorchè) distributed in different ways depending on the season and on the latitude.

In civil application the equation 1.2 is used often to determine the total number of diurnal hours.

In figure 1.3 we can observe the plot the total hours available for observation during a year as function of latitude.

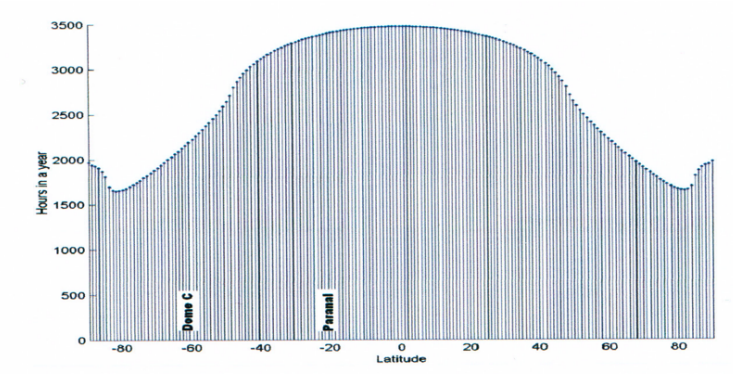


Figure 1.3: Illustration of the available hours for observation during the year.

1.3 Magnitude scale

Magnitude scale is defined by Pogson law:

$$m_1 - m_2 = -2.5 log I_1/I_2 \tag{1.6}$$

where m_1 and m_2 are the apparent magnitude and I_1 and I_2 are the intensity of the light $[erg/s]$ of two different sources[1]. The sign minus is due to the inverse scale. So if the bodies have $I_1/I_2 = 10$, they have a difference of 2.5 in terms of magnitude. If $I_1/I_2 = 100$ the difference is about 5. It dramatically increase because of the logarithmic scale.

As we can see, the magnitude scale is a **inverse logarithmic scale**, **not** linear because the nature has donated to human a logarithmic scale in terms of sensibility to light in order to not saturate observing a bright source. For human eyes, the magnitude limit to observe a celestial object is about 5 or 6 mag in a very dark sky. The fact that is a inverse scale means that brighter stars have minor magnitude and fainter stars have bigger magnitude.

The calibration of the *zero point* of the magnitudes is established by series of standard stars calibrated initially on the polar star (today suspected variable). For the polar star $V = 2.12$ and then the calibration has been extended to many stars near the north pole.

Today there are many catalogue that report every star and its apparent visual magnitude. In particular for the Sun $V = -26.74$ and Vega, used as a calibration star, $V = 0.03$. The stars that are used are reference have a well established magnitude but they have V about 0 because they are used as references.

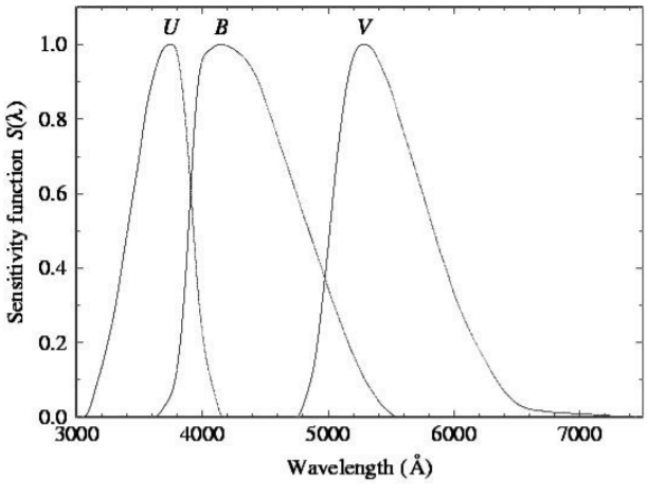
Because of the logarithmic scale, the magnitudes **can't be added or subtracted linearly**.

- You must convert magnitude in linear scale.
- Then you can sum up or subtract (in linear scale).
- Finally, you have to re-convert to magnitude scale.

However, we must remember that for magnitudes with vary small differences (a ratio of 10% that corresponds to a difference of 0.103 of magnitude) the magnitudes are linear but then they turn rapidly to higher values.

Example Suppose to have two stars, $m_1 = 0$ (a calibration star) and $m_2 = 2$. What is the magnitude of the binary system? Someone can answer that $m_{tot} = 0 + 2 = 2$ but it is **FALSE**. I'm adding the magnitude (that has a logarithmic scale) in a linear way. The result is an absurd because I consider

[1]There are different types of notations. It is possible to indicate the apparent magnitude with m but if we are considering the visible band it is common m_v, v or V.



The Johnson-Morgan UBV filter system.

Approximate central wavelengths and bandwidths are:

Band	$<\lambda>$ (Å)	$\Delta\lambda$ (Å)
U	3600	560
B	4400	990
V	5500	880

Figure 1.4: UBV Johnson system.

two object but the sum is equal only to one object, that is the binary system is fainter than the two stars together.

Generally we consider the visible band $(3900 - 7000A^{\circ})$ but there are many other band like the UBV photometric system (from Ultraviolet, Blue, Visual) also called the Johnson system (figure 1.4). It is characterized by passband very wide: inside there are many photons with different wavelength.

1.4 Magnitude-distance relation

The relation between apparent magnitude m, absolute magnitude M and distance of the object d in parsec[2] is:

$$m = M + 5logd - 5 + (A_V) \tag{1.7}$$

so the apparent magnitude depends on the distance. The term A_V can eventually appear and represents the absorption of the dust in the interstellar medium in the visible range (interstellar reddening). We will discuss this term later.

M is the absolute magnitude, the intrinsic magnitude of the object. This is defined as the apparent magnitude of the object at a convectional distance of 10 parsec from the observer in absence of interstellar reddening (in general we have to take it into account). So we can define it as:

$$M = m - 5logd + 5 \tag{1.8}$$

So when $d = 10pc$ the apparent magnitude is equal to the absolute magnitude.

[2]The parsec is defined as the distance at which one astronomical unit subtends an angle of one arcsecond

The absolute magnitude of the Sun is $M_\odot = 4.83$. This is an important value of reference that we must remember. It is used to make comparisons with other object to understand their level of brilliance. Moreover it is close to the visible limit for human eyes. This gave us the idea that we can observe only until a distance of 10 pc that is very low. The thickness of the galaxy is about 500 pc and its diameter is about 30000 pc. If we consider also the reddening effect, a Earth observer can see a very small part around himself.

Now, using the equation 1.7, it is possible to define the **distance module**:

$$m - M = 5logd - 5 + (A_V) \tag{1.9}$$

so m increase with distance d because more distant is an object, bigger is the apparent magnitude: the object is fainter but the scale is a logarithmic **inverse** scale. Of course if absorption A_V increases, also $m - M$ increases.

The distance module of the galactic center (GC) is about 14.35, without the absorption effect, that corresponds to a distance of about 7.4 Kpc[3]. So stars like the Sun are visible with the naked eyes until a distance of about 11 parsec, that is a very small distance compared to the galaxy size. To reach a distance of 100 pc, it is necessary $V = 9.8$. Moreover, if we consider the absorption effect ($A_V = 30$) the distance module became $(m - M)_{GC} = 14.35 + 30 = 44.35$. This fact tells us that we are very restricted in observations.

It is clear that to analyze the galaxy, it is necessary to observe at lower magnitude than the first surveys based on photography plate taken with small telescopes and to observe in the direction of GC to understand the size, the shape and the star distribution of the galaxy.

It is also necessary to remember that in the Solar System the absolute luminosity is defined as the luminosity of those solar body at a distance from the Sun and the Earth of 1 AU.

The Kapteyn universe Around 1900, Hugo von Seegler studied the structure of the galaxy by making counts of stars between successive magnitudes: he determined the rates at which the galaxy was diminishing in multiple different areas of the sky. In 1901, Jacobus Kapteyn (1851 – 1922) employed the proper motions technique and derived a statistical approach that allowed him to estimate the average distance to stars between successive magnitudes, effectively providing a scale for von Seegler's discoveries. Conclusive results: von Seegler and Kapteyn estimated the galaxy to be an oblate star system approximately 10 kpc in width and 2 kpc in thickness with the Sun being relatively close (0.6 kpc) to the center. In conclusion, Kapteyn estimated a uniform distribution of the stars in the galaxy.

However, there were some problems.

- Kapteyn's work was based on an unproven presumption, that is, there is no light absorption in space and we know this is false. Light absorption is an important effect, a considerable factor. With light absorption in space, stars look more faint, thus, seeming further away than they actually are.
- Kapteyn and Seegler missed the galactic center that is no visible in the norther region where they made the observations.
- They used a small refracting telescope of about 10 cm of diameter so they could observe only the brightest stars.

Nowadays this description is not considered true but it is an important historical fact in the science history.

[3] $d = 7.4Kpc$, distance from the GC, has been established using the halo distribution of the globular cluster.

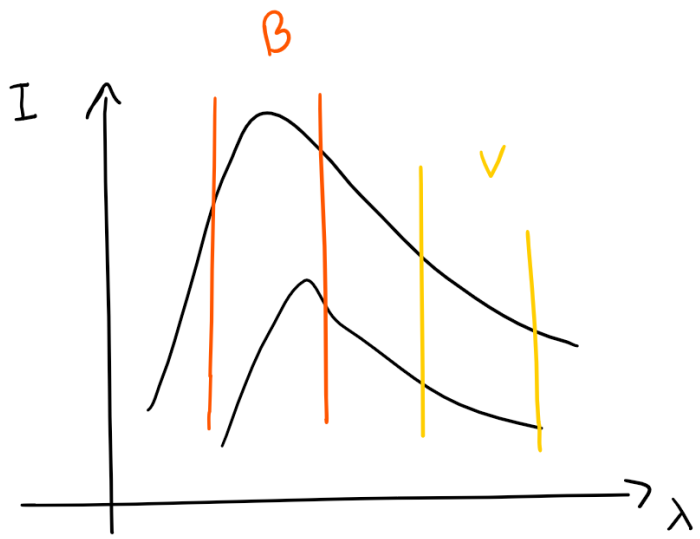


Figure 1.5: Two spectrum similar to black bodies.

1.5 Color index

The color index is defined as:

$$c_{1,2} = m_1 - m_2 \tag{1.10}$$

that is the difference in magnitude between two different passband, that is two different filters. For example, the color index for the Sun in bands $B - V$ is 0.65.

Suppose to have two spectrum of stars similar to two black bodies, figure 1.5. In both cases the intensity is different in function of the wavelength. If we consider the second one, shifted with respect to the first one, bands B and V change so color index change with different temperature.

The calibration stars for color index are A0V no reddened. They are main sequence stars with H burning at the center and a temperature about 10000 K. Astronomers have chosen this type of stars because their spectrum is very similar to a black body (a plankiana). For 10000 K, the Bolzmann equation tells us that the majority of chemical elements and molecules do not emit lines and the spectrum is dominated by H light and the Balmer series. In this way the spectrum is almost continuous with few H lines.

One color index hasn't been calibrated on A0V; it is the **bolometric correction**, the difference between visible band and the bolometric band (that includes all the spectrum). Bolometric corrections are very important to pass from theoretical models to measurements. We will see it later.

1.6 Whitford law - absorption law

Analysing the spectrum of binary system, in the past we understood that the fixed lines (not subject to Doppler effect) were caused by interstellar absorption by gas, that creates very narrow lines, and above all by dust. The dust is made up of particles of carbon, iron, silicates of different shape, combinations and orientation and they are strongly absorbent. For example, if we observe in the direction of the GC, the absorption is much more bigger than if we observe perpendicularly to the galaxy plane (figure 1.6). In particular around the Sun the absorption $A_V = 1$ for 1 kpc, that is 1 magnitude for 1 kpc in the direction of the GC. So the intensity is reduced by a factor of 2.5on 1 kpc in average.

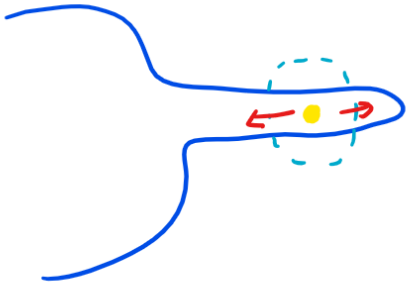


Figure 1.6: Illustration of the galaxy.

The particles that cause the absorption in the interstellar medium have the same size or are bigger than the wavelength in visible band, in order of microns.

In general the interstellar absorption law is:

$$A_\lambda = 1/\lambda \tag{1.11}$$

known as Whitford law, valid, in first approximation, from visible to infrared band. It is an empirical law and shows that the absorption is more intensive at short wavelengths respect to long wavelengths. A_λ increases when λ decreases so the absorption effect is much more important in the blue band compared to the V band or infrared.

Whitford law has been obtained by using mainly the early type stars because they are brighter, with less absorption lines in the spectrum and very well studied. But if we consider other spectral type, so different temperatures, the absorption law can be different. We will discuss it later.

The interstellar absorption is often estimated using the **color excess** defined as the difference between the observed color index (from photometry) and the intrinsic one (by spectral type) in absence of absorption. For the B, V band (the most used) the color excess is:

$$E(B-V) = (B-V) - (B-V)_0 \tag{1.12}$$

It is the consequence of the inverse scale of the magnitude. E stands for *excess* and $E(B-V) > 0$ because of the absorption that is bigger in the B band than in the V band and the difference between B and V increases because of the reddening law in *not* uniform way.

If $(B-V) = (B-V)_0$, that is $E(B-V) = 0$, the color index observed is equal to the intrinsic color index so there is no reddening. This occurs with the Sun and some few stars around it.

Connected to this equation, there is an important relation between the absorption and the color excess:

$$A_V = 3.1E(B-V) \tag{1.13}$$

The constant 3.1 is very important but also very delicate to establish. If this number is incorrect, the distance is wrong, also for different orders.

The clear result of the interstellar absorption on the spectrum of the star is the reduction (drop, diminuzione) of the flux (or the intensity) of the source, respect to the continuous: an effect that is bigger at shorter wavelengths. On the other side, the absorption decreases at longer wavelengths, becoming negligible (trascurabile) from far infrared to the radio regime of microwave, figure 1.7. A spectrum affected by absorption appears with a different slope (pendenza) from the original one. The

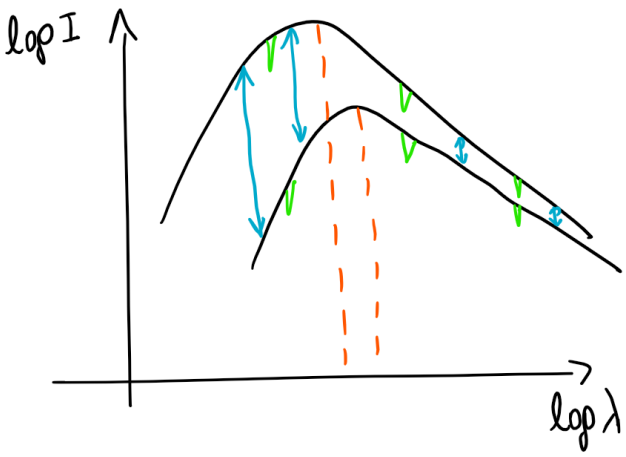


Figure 1.7: Illustration of the galaxy.

color index grows and the peak of the continuous is shifted to bigger wavelengths: the spectrum is similar to the spectrum of a star of advanced spectral type (assomiglia ad uno spettro di tipo avanzato). However the lines remain **in the same position**, the equivalent width (larghezza equivalente) remains the same and the spectral analysis allows to recognize the original spectral type of the source (so the intrinsic color index). Indeed the lines are quantum transitions so there is no way the lines could change their position.

The reddening effect (not only the interstellar one, but also the atmospheric one) doesn't have to be confused with the Doppler effect. This effect causes a movement (a shift in X-axis) of the entire spectrum, also of the lines, to bigger wavelength for objects that are moving away.

The absorption law does not include the correction for the Earth atmosphere. This correction depends on the height of the object from the horizon, on the direction of observation, on the airmass and aerosols. In average, the correction in visible band is 0.15 magnitude per airmass (given by combination of reddening due to Rayleigh scattering of the molecules and aerosols).

In conclusion, we can say that the reddening is a consequence of the Whitford law so stars appear reddened throughout the interstellar matter than their intrinsic color. Moreover Whitford law has a very simple form only in optical widow. In other band of the spectrum isn't linear and it isn't constant in all the galaxy. For example in regions of stellar formation the constant is about 4 and in other regions the constant is about $2.5 - 2.7$.

1.7 Metallicity indicators

Stars contains mainly H, an important fraction of He and other metals. There are two ways of measurement:

- number of observing atom, used by spectroscopy that measures the abundance of chemical elements;
- fraction of mass, favorite by theorists in models.

X, Y, Z In particular X, Y e Z indicate the fractional abundance of, respectively, hydrogen, helium and metals. By definition:

$$X + Y + Z = 1 \tag{1.14}$$

The original cosmic abundances are $X = 0.78$, $Y = 0.22$ and $Z = trace$. For the Sun the typical abundance is $X = 0.707 \pm 2.5\%$ and $Y = 0.27 \pm 6\%$. The abundance of metals it has been assumed

often about 0.020 but some resent researches based on astroseismology give a lower abundance, about 0.016. So nowadays we assume $Z = 0.0189 \pm 8\%$.

The fractional abundances are very important but not easy to find. Indeed we can observe only the transparent atmosphere, above the photosphere, that can't be representative of all the star. In general the chemical elements are in equilibrium but the different nuclear reactions, that take place during the evolution of the star, can change the elements distribution. Moreover a particular phenomena can occur: **the elements segregation (or distribution)**. It consists on the separation of the heavier elements in the inner part of the star (o sort of sedimentation) while the lighter ones remain above.

For all those reasons, it is not easy to determine X, Y and Z with accuracy.

$[Me/H]$ Another important metallicity indicator is the ratio:

$$[Me/H] = \frac{log(Me/H)}{(Me/H)_0} \tag{1.15}$$

It indicates the logarithm of the metal abundance with respect to the H: it is normalized to the Sun. Usually it is used the iron (Fe) like reference of the metals, because of its abundance and because it is easy to measure.

For example, if $[Fe/H] = -1$, the star has a contents of Fe (in general metals) 10 times lower (because of the logarithmic scale) than the Sun in atoms number. If $[Fe/H] = +1$ the contents is 10 times more than the Sun. If $[Fe/H] = -2$, the contents is 100 times lower than the Sun.

In general, the process is:

- to measure the absorption lines from the spectrum of the object;
- to get how much is the metal abundance;
- to extrapolate the total abundances;
- finally to check the results with theoretical models.

However it is an hard work for many reasons:

- it is time consuming;
- you must have a very good spectroscopy;
- you have to measure the stars one by one;
- it is a long process which implicates the knowledge of the temperature of the star. Indeed the intensity of spectrum lines depends on Boltzmann equation (that describes the distribution of atoms population between different levels) and on Saha equation (that describes the ionization of equilibrium). Both the equations depend on temperature. So, to know the metal abundance, you need to know a precise temperature.

Some important deviations from the global contents of metals and respect to the iron, it has been registered for old stars poor in metals, where we observe an excess of alpha elements due to supernovae of type II.

But, what we can do for very faint stars, for which it isn't possible to make a precise spectroscopy? Or, for very extinct regions? We can introduce another indicator.

UV excess Historically, the UV excess has been defined as the difference between the color index $U - B$, to parity of $B - V$ with respect to the Hyades:

$$\Delta U = (U - B)_{Hyades} - (U - B)_{obs} \tag{1.16}$$

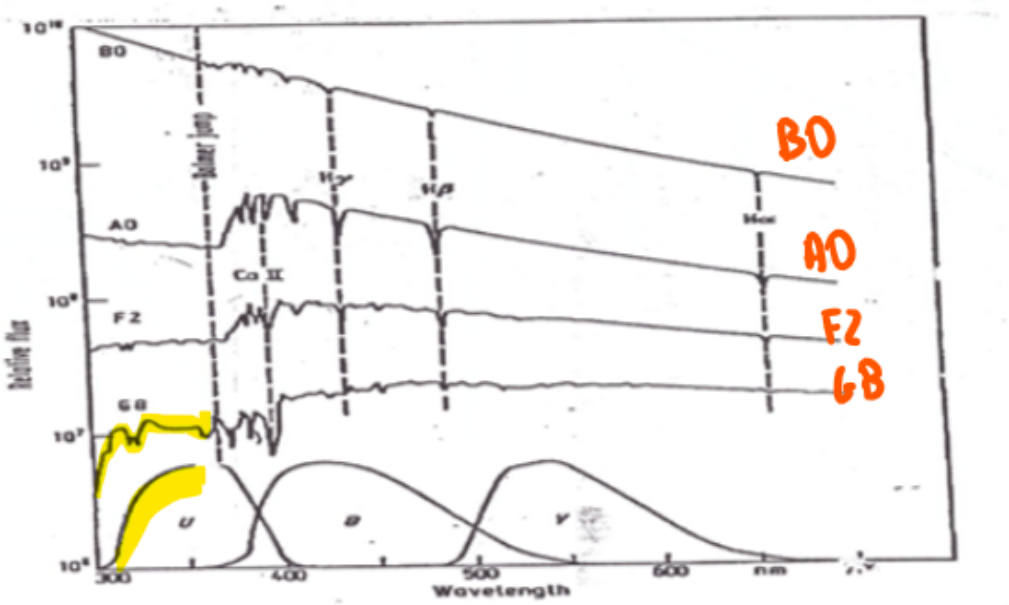


Figure 1.8: Blanketing effect in the U band.

The Hyades are an open cluster near the Sun. They are Sun stars (population I, young in the disk), with very low reddening and same spectral type (so the same temperature T, probably). They are reference stars for which $\Delta U = 0$.

However, during the 50' astronomers realized that there were some peculiar stars around the Sun; they were stars of population II, so stars of halo, with an UV excess about $0.15 - 0.6$. So the indicator say that this type of stars are poor in metals.

Indeed the UV excess can be interpreted as the **blanketing effect** due to the present of metals. If we observe the figure 1.8, we can notice different spectrum of some different type stars. Stars like the Sun, G stars, present some metal lines in the U band, like iron (Fe) and calcium (Ca). The remaining part of the spectrum is substantially flat, there are only some H lines, common in all stars. Now, if we consider a star of different type, iron and calcium lines have different depth. In particular those lines are deeper to decrease the U band. This is the blanketing effect: the abundance of this metal produce this reduction in the U band so U magnitude and color excess increase values. In other words, for metal poor stars (like population II) ΔU is higher. In general, for metal poor stars $\Delta U > 0$.

Preston index There is another index, introduced by scientist Preston, calibrated for variables RR Lyrae, star of F type with well defined metal lines, at the minimum of their luminosity, corresponding to the lowest temperature. It is defined as 10 times the difference of spectral class obtained by the H lines in Balmer series, with respect to the spectral class obtained by K and $CaII$ lines.

$$\Delta S = 10[SP(H) - SP(CaII)] \tag{1.17}$$

This is valid only for spectral types A5-F5 ($(B-V) = 0.15 - 0.45$). Of course, $CaII$ classification can be different from H classification because metallicity changes in stars so it is a different indicator of temperature.

RR Lyrae have been chosen because they are very bright, easily identified, variable with small periodicity (some hours for a variability of 10% or 30% of their luminosity). They are 1000 times brighter than the Sun. Due to their nature, they are visible also at big distances, also extra galactic distances so they are perfect standard candles to measure distance.

For fainter stars, for which we can have low resolution spectra, it is valid:

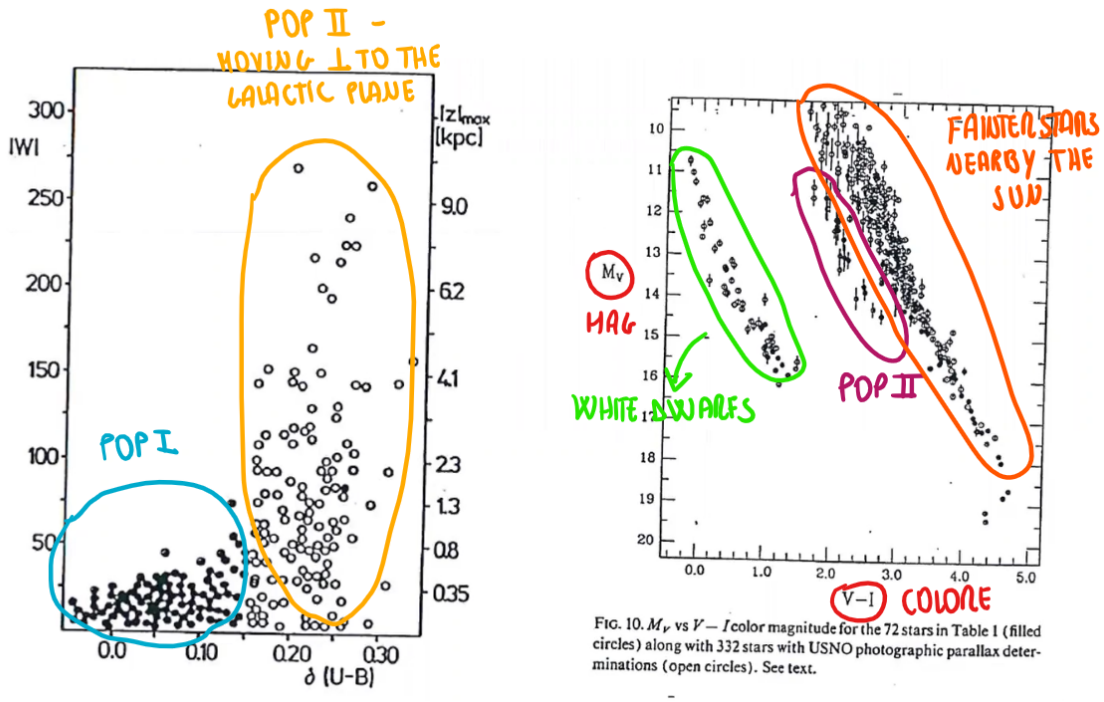


Fig. 10. M_V vs $V-I$ color magnitude for the 72 stars in Table 1 (filled circles) along with 332 stars with USNO photographic parallax determinations (open circles). See text.

Figure 1.9: Evidence of population II.

$$[Fe/H] = -0.23 - 0.16\Delta S \tag{1.18}$$

In general, it is important to have a spectral classification. For spectral classification you just need low resolution spectra, easily obtained also for fainter stars while for abundance analysis, it is necessary a high resolution spectra.

Application of UV excess The UV excess has been used to discovery the population II of stars.

As visible in figure 1.9 in the first graphic, stars of population II are metal poor and their cinematic index indicates that they are moving perpendicular to the galactic plane. In the second graphic, a color-magnitude diagram, we can observe different stars.

- The white dwarf group: last phase of evolution of stars, often rich in carbon and which are cooling down.
- In the second part of the graphic there are mainly faint stars, very common in the universe, nearby Sun. They are called also "normal stars" because they are burning H in inner part. Their distance d has been measured by parallax method from ground.
- In parallel to the branch of faint stars, there is the branch of stars of population II. They are "intruders" in the galactic disk and are a minority. They are main sequence stars moving perpendicular to the galactic plane. Because of their lower metallicity, they follow the main sequence phase in parallel to the fainter stars, so they appear more blue. This type of stars is 100 times less rich in metals with respect to the Sun and they are shifted for 2/2.3 in color index (x-axis) so they have higher temperature.

1.8 CCD sensor

How deep can we observe? To answer the question is necessary to understand how works a sensor.

Nowadays scientists have available two types of sensor: CCD and CMOS.

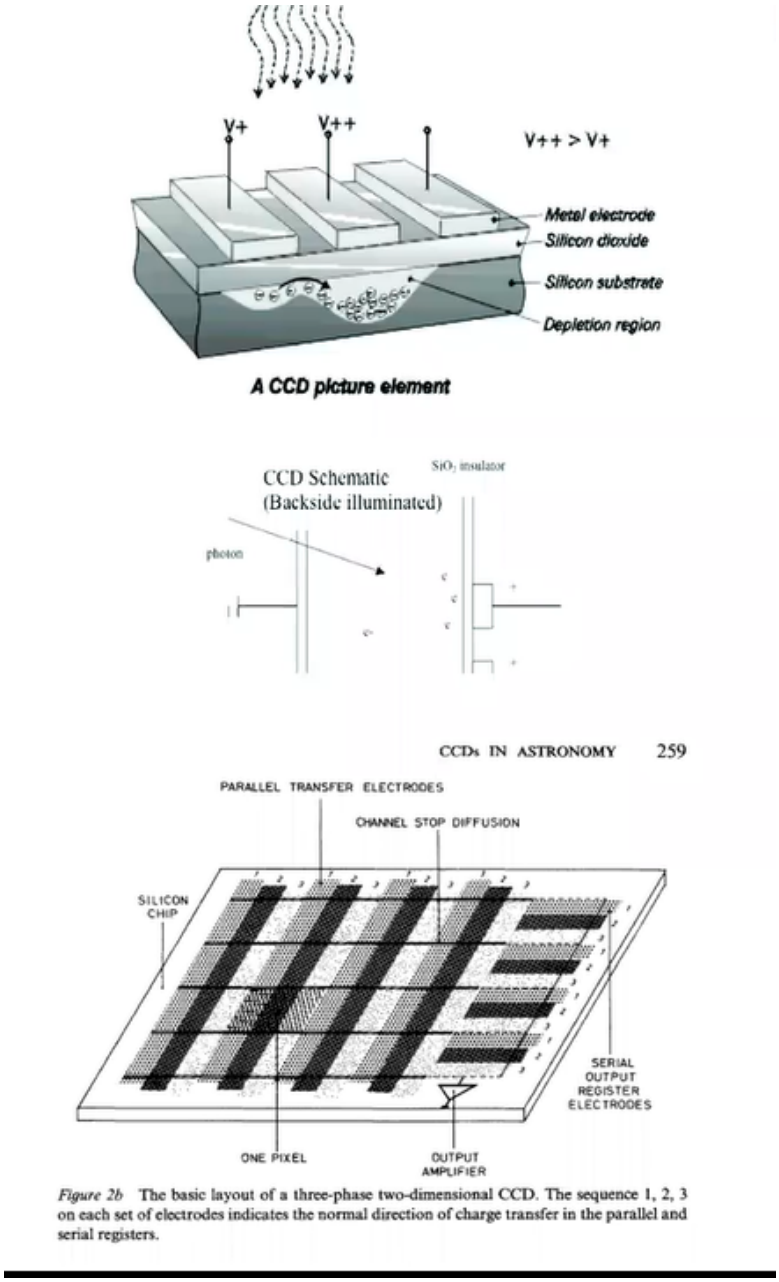


Figure 1.10: Illustration off CCD functioning.

Historically we used only CCD and only recently it has been developed the CMOS sensor. In particular, in the past semi-conductor detectors have been built on CCD principals based on the transfer of photoelectrons (charges produced by photons) using electrical fields.

Concentrate on CCD sensor. It works as illustrated in figure 1.10.

The CCD is composed by a unique peace of silicon equipped with electrodes that create electric fields. During observation, photons from the source hit the silicon and produce electrons below the silicon substrate (with the energy given by photons, electrons pass from a ground state to conduction state). In this phase electric fields are fixed and electrons are collected below electrodes, confined in a limited area. At the end of exposure, positive electric fields (positive polarization) conduct the electrons in specific directions and move them out of the visible field.

As visible in figure1.11, if a photon hit the sensor on the border between two pixels, it isn't lost: it goes in one of the two pixels depending on its position.

Sometimes, at very bright illumination level, there are so many electrons collected under the electrodes

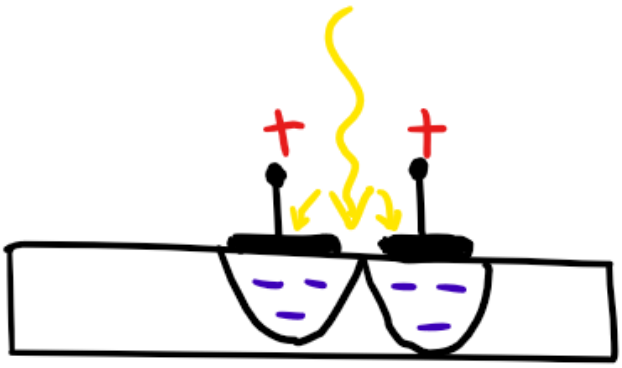


Figure 1.11: Two CCD pixels.

that they creates a big negative field so they invade the pixels nearby (in any case they aren't lost). This is the **blooming effect**.

Then there is the read out phase. The electrons are counted and the signal is amplified by an amplifier. The sensor gets several packages from different pixels in a sequence of time (that is spacial distribution is converted into time).

In CCD sensor the charge transfer in phase of read out can take time and also some electrons can be lost. In any case, the amplifier is only one so, with a good construction of silicon substrate and electrodes, the quantum efficiency is almost **uniform**.

With modern CCD the efficiency in visible band, number of photons converted into electrons, is about 90%. In particular the peak of efficiency is shifted a bit on the red side so real peak is between 600 and 700 Angstrom while the peak of the visible is at 550 Angstrom. the difference is not so big but there is some slope to take in account. The efficiency drops down in the UV band where there are some reflection effects due to the silicon substrate. In general it is necessary to take in account that old photography plates, on which many international system are calibrated (for example Johnson system), have the peak of sensitivity in blue so they are almost flat in the optical range, eventually going down.

The uniform efficiency of the sensor is the main reason why astronomers use CCD still today.

Indeed, CMOS sensor works differently. As visible in figure 1.12, CMOS amplifies the signal from pixels one by one and then the amplified signal is sent to the electrical connections. So it requires more electric power, more electric devices which make the CMOS sensor less uniform than CCD even though it is faster in read out phase. For these reasons CMOS are used for fast photometry, adaptive optics and other special uses.

The different uniformity between CCD and CMOS is due to the fact that in CCD the silicon substrate is a unique piece in which the electrons remain enclosed in a restricted area thanks to the electric fields (there aren't physical divisions between pixels). In the second sensor there are physical gaps between pixels. This gives a less uniformity in the registered signal.

Also CCD are very **stable** because the efficiency depends on the capacity to transform photons into electrons, that gives the necessary energy to electrons to make a bump from ground state to conductive state. However, this amount of energy is fixed by quantum transformations (quantum levels are extremely stable) so the CCD is intrinsic an extremely stable device.

Moreover, since the number of electrons depends on the number of photons detected in a proportional way, the CCD is also very **linear**.

In figure 1.13, it is possible to observe the several capacity of reaction of CCD pixels to photons coming form the source. The graphics show curves for different wavelength: since there is no physical division

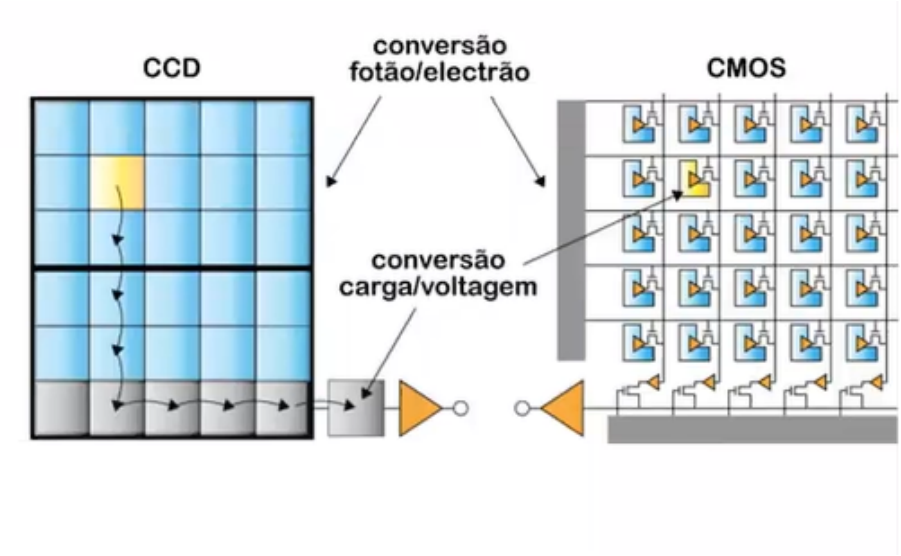


Figure 1.12: CCD and CMOS sensors compared.

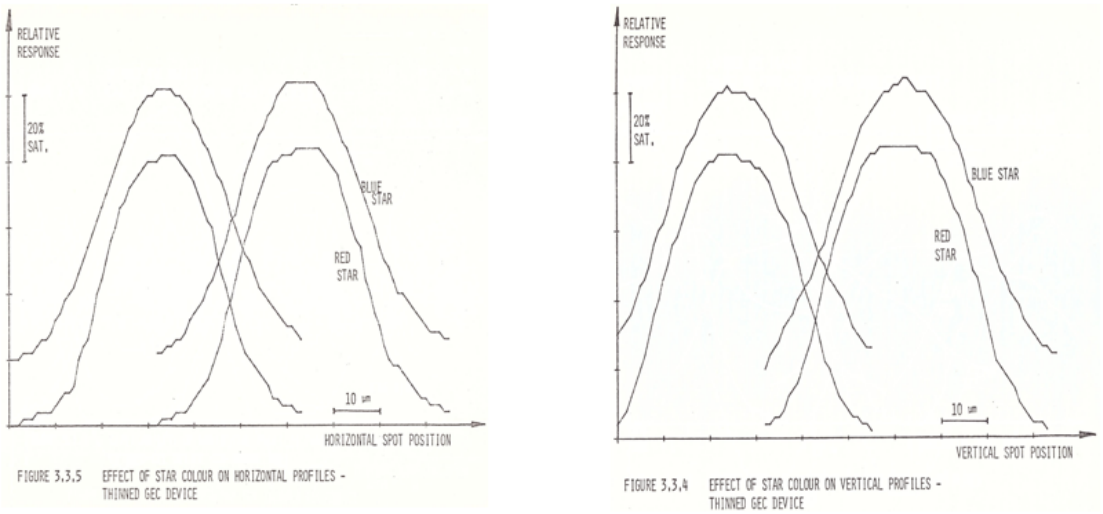


Figure 1.13: CCD sub-pixel sensitivity.

between pixels, there are no electrons lost during the exposure with a uniform trend.

In both cases, when you read out the sensor (conversion of photoelectrons into signal) the image is destroyed and this isn't logical apparently indeed during the exposure and the reading out, you can't change the exposure time, correct the position and so on. So in the future we will have to think in other new devices.

1.9 The S/N ratio and the limit magnitude

When light comes from an astronomical source, it crosses the atmosphere where is subjected to absorption, deviations and scattering effects. Here the 15% is lost. Then it passes throughout the optical train (the telescope). Here there is the most important loss of photons, about 55% caused by reflections of the mirrors, filters, the spiders that support the secondary mirror and by the secondary mirror itself. Finally only the 30% of incoming photons are detected, even though the sensor efficiency is about 90% in visible band.

The light path is described in figure 1.14.

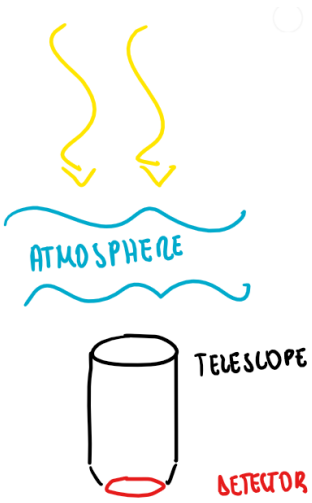


Figure 1.14: Light path.

The signal registered by the detector is given by:

$$S_{tot} = s \cdot t \tag{1.19}$$

where S_{tot} is the total signal collected, s is the signal per unit time (depending on the size of the telescope, the efficiency and the brightness of sources) and t is the exposure time. From this equation we can observe that the signal goes linearly with time, due to the linear trend of the detector.

However, during the exposure and the reading out there are different sources of noise, defined as:

$$N = \sqrt{S} = \sqrt{s \cdot t} \tag{1.20}$$

so

$$N^2 \propto t \tag{1.21}$$

The most important quantity to establish is the ratio S/N.

Therefore, depending on the equations above, for a fixed source, going to infinite exposure time, S/N goes to infinity. So, from mathematically point of view, you can compensate the size of the telescope increasing t.

In general, the study of the ratio S/N is a fundamental element planning the observations. Many modern resources need to know with accuracy this ratio in order to study, for example, exoplanets and faint stars.

The dissertation about S/N depends on noise nature and, so, on detector characteristics. Then the limit magnitude depends on the ratio S/N.

In this case, we will treat only stellar images, relatively small with respect to the sensor size and for which stars occupy only few pixels. However, the observations for extended objects are similar.

We have discussed in the last section the characteristics of CCD and how it works. Now, assuming a strong linearity of the signal as function of incident photons, there are different sources of noise:

- Read Out Noise, RON during the read out phase of the signal;
- Dark Noise or thermal noise;

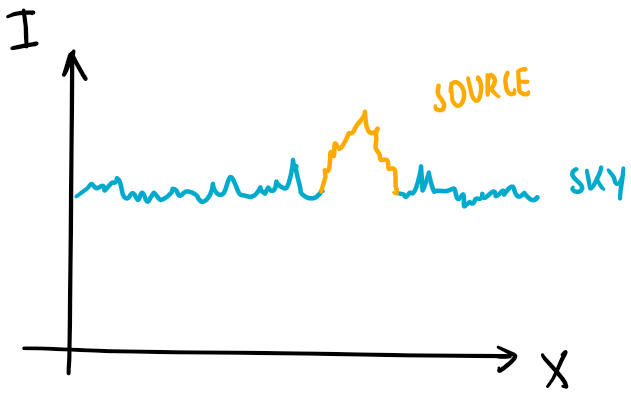


Figure 1.15: The source intensity is summed up to the sky level.

- statistic Poisson noise (shot noise) of the signal and of the sky;
- spatial noise or pattern noise, SPN.

RON In modern CCD, the Read Out Noise is contained in a set of ten pixels so it is negligible for deep images or for bright objects with high S/N. Instead it is important in case of images with low sky level and low signal, for example in spectroscopy or for small exposure time.

In general it is almost constant so it is a fixed number for every device.

Dark Noise Also dark noise is negligible in many modern CCDs. It is caused by the functioning electronics that produces some casually jumps of electrons from ground state to conductive state. More or less, they have Maxwellian distribution of velocity that depends on temperature. So many CCDs are cooled down to lower temperature in order to keep all thermal electrons at very low level. Therefore in modern CCDs the pollution of dark electrons is very limited.

Shot noise For common applications, the shot noise is the most important contribution of noise. For its nature, it varies with the square root of the signal.

In general astronomers observe very small and faint objects and they have always take into account the sky background, disturbing from ground and space. If the source dominates over the background, there can be a reasonable approach but if the source is comparable to the sky, there can be problems (figure 1.15).

In general the sky can be removed if it is known, however every mathematical operation increases the noise N because you don't know exactly the sky level so you use averages values and then you extrapolate. So if you try to remove the sky from the source imagine, the noise grows up.

In the case of a very strong signal, the sky level is unimportant because the intensity registered is dominated by source intensity (figure 1.16). Instead, for fainter stars, the signal is dominated by sky background.

So, it is valid:

$$N = \sqrt{S + SKY} \tag{1.22}$$

and for very faint stars, the signal S is negligible compared to the sky level, so:

$$N = \sqrt{SKY} = \sqrt{sky \cdot t} \tag{1.23}$$

where sky is the intensity of sky level and t is the exposure time. Usually SKY is a fixed quantity.

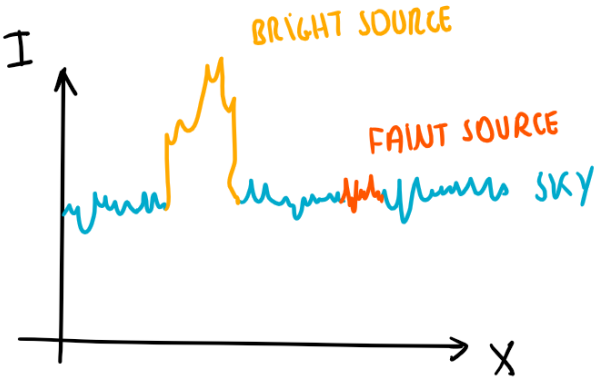


Figure 1.16: Bright and faint stars compared to the sky level.

So in this case, the S/N ratio is:

$$S/N = \frac{s \cdot t}{\sqrt{sky \cdot t}} = \frac{s}{\sqrt{sky}} \cdot \sqrt{t} \tag{1.24}$$

Another time, we see that increasing time t, S/N increases: if $t \to \infty$ then $S/N \to \infty$, independently on the size of the telescope.

Therefore, in general the sky and the telescope aren't the barrier to the sensor capacity to reach the deep sky. Indeed you can observe sources much fainter then the sky background, if the observing time is enough long in order to get the signal S enough high compared to the noise N.

So the real barrier to reach a very high S/N is the SPN.

SPN noise SPN, spatial noise or pattern noise, is the real limit in observations to get a very high S/N and is a specific number for each sensor.

Even though the CCD is almost uniform (the uniformity is almost constant with time), there are some small sensitivity variations from pixel to pixel of about a fraction of 1%. Therefore the quantum efficiency change from pixel to pixel as well the scale background and the noise profile of a star is affected by the different sensitivity of pixels.

Mathematically, considering also this type of noise e knowing that all contributions are independent, the total noise is the square root of the sum of all contributions:

$$N^2 = A \cdot RON^2 + A \cdot sky \cdot t + A \cdot Dark \cdot t + A \cdot s \cdot t + (SPN \cdot (sky + s) \cdot t)^2 \tag{1.25}$$

where A is the area occupied by source on the sensor (for a star, generally 3 or 4 pixels of diameter). In particular every term of this equation has a geometrical representation (for example RON is a constant). Then S/N became:

$$S/N = \frac{s \cdot t}{\sqrt{sky \cdot t + cost}} \tag{1.26}$$

where $cost \propto t$. So S/N can NOT increase. You may think to compensate this noise taking the **flat fields**, calibration files taken before or after the observation, because it is constant in time. Flat fields are realized illuminating the dome or a background surface uniformly in order to obtain sensitivity distribution, pixel by pixel. In this way the best correction reachable is about 0.5%, so from $1\% \to 0.5\%$. So flat fields are useful to reduce SPN, but not to remove it. Why? Because typically you are exposing at different wavelength so a white light on the dome is not the real light of a star or

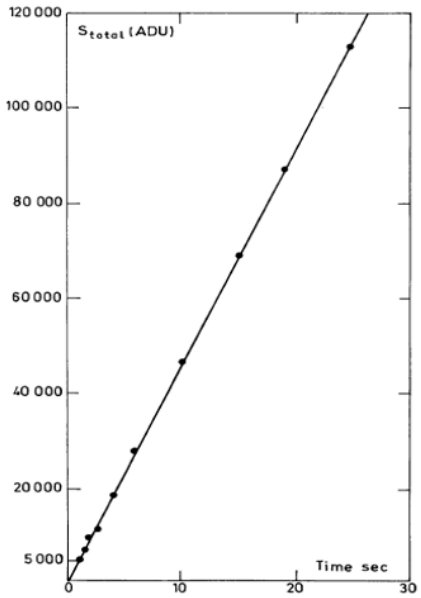


Fig. 3a. Linearity of the CCD. The straight line is the total output signal on the central pixel and those on the same column.

Figure 1.17: CCD linearity in the full field and in single pixels (Mellier et al. 1986).

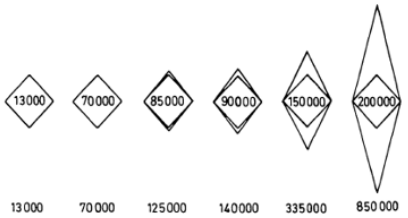


Fig. 3b. Evolution of the FWHM with input signal increase: the output signal on the illuminated pixel is represented on the central diamond. We mention below each of them the total charge read by the CCD. Note the diamond shape along the columns whereas no degradation is observed along the rows

Figure 1.18: Local distribution of electrons when pixels saturate.

of the sky background (for example sky background is composed by aurora light, zodiacal light etc.). So why to use the wrong spectrum? You should expose twice: one on the source and one on the sky background in order to make a better operation of flat field. However, in this way S/N becomes lower because you expose less time. Moreover sky background changes so it is not easy to find the best correction.

In figure 1.17 it is possible to observe a bulk CCD exposed in the lab with uniform light: the graphic shows the measurements of light all over the CCD. What we extract from this plot is that sensitivity is perfectly linear as function of time.

However, as stated before, when the pixel is saturate because it collected too many electrons, they spread out in the nearby pixels and linear scale is lost. This is a local effect (figure 1.18).

Therefore SPN is the real barrier; a factor of 1% of this type of noise means that the faintest star observable is about 100 times fainter than the sky background. In term of magnitude, for ground based observations, sky background is about $V = 21.5$. So with a factor of 1%, the limit magnitude is 27. For a factor of $2 - 3\%$ the limit magnitude is $26 - 25$. From space $V = 23$ so, with a factor of 1%, the limit magnitude is 28, not too different from ground based observations.

Can we get a better correction of SPN? There are two innovative techniques, illustrated in figure 1.19.

- **Drift scanning** - The first solution was proposed by Mackay. Usually, after an exposure, during

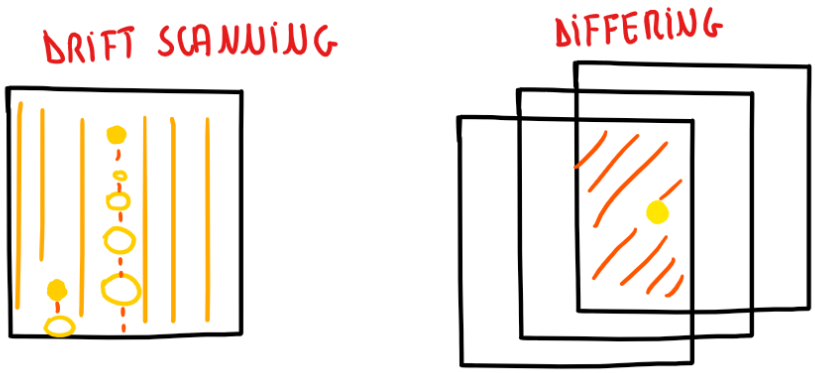


Figure 1.19: At left, the drift scanning method. At right, the differing method.

the reading out of charges in specific directions, there is a shutter (otturatore) in order to stop the accumulation of photons from the source (or other objects present in the view field). However Mackay had an innovative idea: in order to smooth down the variation from pixel to pixel he created a device that don't close the shutter during the reading out and it makes the read out very slowly (in 10/15 minutes). At the same time, the CCD moves with the same speed of the reading out but in contrary way respect to the telescope. In this way, the image increases in brightness and different pixels with different sensitivity are interested so, statistically, variations from pixel to pixel are smoothed down. With this innovative method, Mackay was able to reduce SPN from factor 1% to a factor of 0.1%. It is a reduction of factor 10, which means a limit magnitude of 29.

However there are some problems: there is no uniformity in S/N: in figure 1.19, S/N is bigger above the source and smaller below where many photons have been accumulated. Moreover, if there is another star, SPN of this star is smoothed down lesser than the first one. Finally, you smooth down SPN only in one direction (the direction of reading out) so you see a very flat image only in one direction, where SPN is about 0.1%. In other directions SPN is 10 times more. For this reason the image presents some vertical strips.

- **Differing** - This technique consists of taking an image of the source, then a second one a bit shifted, then another one again a bit shifted and so on. To obtain the final image it is necessary to re-shift the images taken before. In this way different pixels are exposed and statistically variations between pixels are smoothed down. In this way, you can gain a factor 2 on SPN, depending on the number of shifted images.

  Of course there are some disadvantages. First, the useful common area is smaller than the full field of the sensor. Second one, you miss a lot of observation time reading out the images and storing them in the computer. Nowadays, this method is the most used because it is easy and you can reduce SPN more or less depending on targets, scientific goals, etc.

1.10 Exposure time

One important thing when you are planning observations is to find the right exposure time. Nowadays there are some online tools to calculate the best exposure time but they are good for standard cases, not for limit astrophysics. Indeed they suppose the most simply case, where the telescope is observing a very faint star, which intensity is comparable to the sky level, so for which is valid:

$$S = s \cdot t \Rightarrow S \propto t \tag{1.27}$$

$$N = \sqrt{sky \cdot t} \Rightarrow N \propto \sqrt{t} \tag{1.28}$$

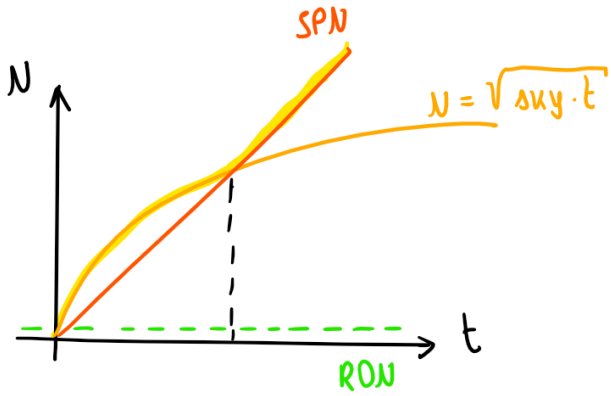


Figure 1.20: Noise as function of time.

In this simple case $S/N \propto \sqrt{t}$ and N has a parabolic trend as function of time t while RON is a constant, as illustrated in figure 1.20.

In this graphics appears also the other important source of noise, the SPN. We know it is an additional noise, proportional to the intensity of the star or the background and linear to the time t, indeed it is represented as a straight line. From last section, we know it is not a random noise but it is a spacial pattern that represents variations in sensibility from pixel to pixel.

In figure 1.20, it is possible to observe that for relative short exposure time, SPN is considerable lower than Poisson noise of the sky. Unless, with very long exposure time, SPN dominates. So the noise change from a random trend to a sort of fixed pattern. This is the ultimate limit imposed to observations.

As we seen in the last section, you can correct SPN partially but you can't remove it. Sometimes corrections can worse the situation, for many reasons: there was too noise, the exposure time was too short and add Poisson noise and so on.

Now, suppose to have 1 hour of exposure which means a factor 1000 for the sky background. Then suppose to correct this factor with a flat field of the same factor. The operation is:

$$\frac{1h \rightarrow \sqrt{1000}}{FF \rightarrow \sqrt{1000}} \tag{1.29}$$

This operation, a ratio, implies the propagation of errors so, statistically, it don't reduce the noise but it increases the noise. So you have to go to very high counts if you don't want to degrade significantly the S/N. Of course, to use a flat field with very strong level of intensity, you have to be sure that the detector is extremely linear. **Suggestion:** go to values not far from sky background and eventually take many images and then co-add them in order to increase S/N and to minimize the Poisson noise.

In general, if you consider the bulk CCD and expose with increasing time under a fixed lamp for flat field, the result is a linear trend, as illustrated in 1.21. The points are distributed with very small dispersion and the plot is a perfect straight line. In particular, there is no deviation from linearity up to the saturation that occurs when a pixel collects 10^5 electrons (in average, it depends on pixel size). Instead, if you repeat the experiment with a laser beam (fascio laser), the result is different: the trend is linear in the first part of the plot and then it deviates. It occurs because the laser beam could be smaller than the pixel so when number of electrons starts to increase in a single pixel, electrons spread out because of the repulsion forces between them in pixels nearby. So the CCD is intrinsic linear but sometimes there are some geometrical effects of spreading of pixels that deviate from linear trend.

Finally in figure 1.22, it is possible to observe N in ADU (electronic unit) or in electrons, as function of time t. The same result can be obtained with S on y-axis. The slope of lines depends on the intensity of the source. The brightest are more sloped, the faintest are less sloped. In general, the faintest you

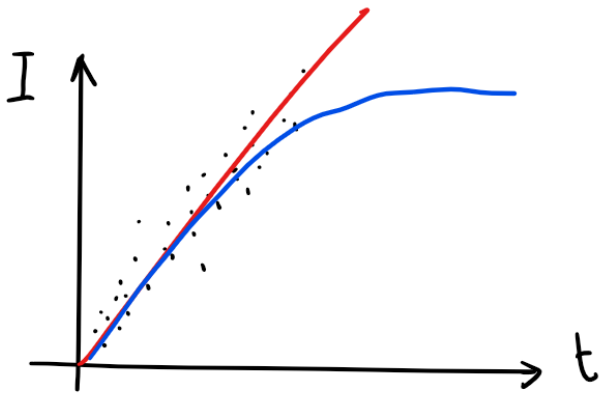


Figure 1.21: CCD response to a fixed lamp.

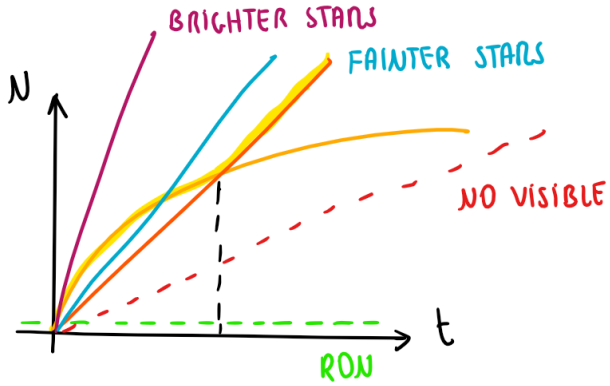


Figure 1.22: Different slopes compared.

can detect is the source coinciding with the slope of SPN. Under this line it is impossible to observe it.

1.11 Accuracy

Suppose to observe a faint object and you want the best S/N: what is the highest accuracy reachable? The best accuracy in general is 70 *ppm* (par per million).

Consider now the figure 1.23. It illustrates the signal of a star (yellow) in function of x-axis and the pattern sensibility of pixels (green).

The variable sensitivity of pixels is the main problem to reach the best accuracy. How can we resolve the problem? One technique is **defocus** the image. With a defocused image, the shape in brightness

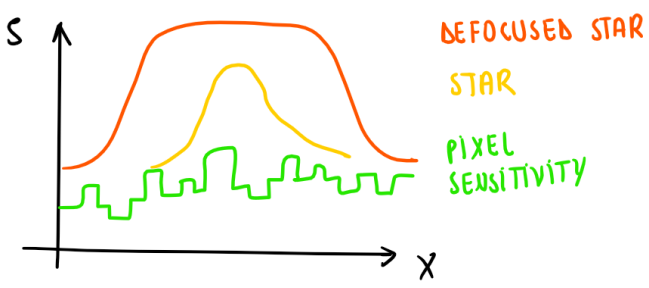


Figure 1.23: Signal of the source in case of focus (yellow) and defocus (orange).

profile of the star gets close to a cylinder, smoothing all pixel to pixel variations. This is a great way to reduce this pattern noise, especially for bright stars.

Then, the Poisson noise remains a problem. You may think that going to very long exposure time resolves it but there is a limit of the electrons that CCD can collect. As we seen before, the limit is about 10^5 electrons in average for pixel. So, because it is a Poisson noise, $N = \sqrt{10^5} \simeq 300$. So $S/N \simeq 1/300$, very far from the best accuracy of 70 *ppm* ($70 \cdot 10^{-6}$). To improve it, it is necessary more counts so we need a wider stellar image (more defocused in order to have more pixels). For example, if the total counts are 10^6 (considering 10 pixels), $N = \sqrt{10^6} \simeq 1000$ so $S/N \simeq 1/1000$ that is better than the first one and so on. In general, to have enough electrons, which means enough photons from the source, you have to take in account how long should be the exposure time. If the telescope is not very large, it is necessary a much longer exposure time. Therefore the limits on S/N are on size of the telescope and the exposure time.

Noise connected to the area Consider now the area occupied by the intensity distribution of the source (A in equation 1.25). It is like a gaussian. At first, we can think that the total area $A = \pi r^2$ but it is not right because the winds of the distribution are negligible in photometry. So we could use the Full Weight at Half Maximum (FWHM) ad diameter but in this way we underestimate the area, loosing important fraction of light. Therefore the best way is to use FWHM as radius:

$$A = \pi \cdot FWHM^2 \tag{1.30}$$

1.12 Papers suggested

There are some interesting and important papers, suggested to the reader on CCD and calibration and correction techniques.

CCD photometric properties

- MACKAY, ARAA, 24, 259, 1986
- **LEACH et al., PASP, 92, 233, 1980**
- GUDEHUS et al., AJ, 90, 130, 1985
- MELLIER et al., AA, 157, 96, 1986
- AMELIO, Scientific American, 1974

Data calibration

- BESSELL, PASP, 89, 591,1979
- COUSINS, MNASSA, 39, 93, 1980
- BECKERT et al., PASP, 101, 849, 1989
- ORTOLANI, CCD ESO Manual:calibration, 1992

Reduction problems, errors observing techniques, sampling

- PEDERSEN, ESO Danish Tel. Manual, 1984 (obs. techn.)
- STETSON papers, for example:
- STETSON, Dom. Astr. Obs. Prepr., 1988 (flats, accuracy)
- STETSON, PASP, 117, 563, 2005 (cal. accuracy)

- ORTOLANI, "The optimization of the use... ESO/OHP workshop, 1986, p. 183 (reductions, general)
- KING, PASP, 95, 163, 1983 (sampling, star size)
- BUONANNO et al., PASP, 101, 294, 1989 (sampling)
- **DIEGO, PASP, 97, 1209, 1985 (star shapes and size)**
- MATEO, PHD thesis (completeness corrections)

1.13 Bolometric correction

In observations it is necessary to pass from measures in some band to bolometric magnitude m_{bol}, the magnitude corresponding to all the flux of the star, from far radio to X. Combing spectrum from ground and from satellites, applying some corrections and extrapolating, it is possible to obtain the black body profile so, integrating, the bolometric magnitude. It is very important to know m_{bol} because it represents the absolute luminosity of the star which is the energetic output, predicted by theories (so which implies rate of contraction, rate of nuclear reactions and so on). Therefore it is important to know the relative correction, exactly the **bolometric correction BC**.

The transformation between visual magnitude into luminosity and vice versa is done thought the bolometric correction that allows to pass from optical magnitude (or another band) to bolometric one:

$$m_{bol} = V + BC \tag{1.31}$$

So the BC is:

$$BC = m_{bol} - V \tag{1.32}$$

which means it is a color index. However, on the contrary of normal color index, BC is calibrated on F5V stars with 6600 K of temperature that correspond to an emission peak at the center of visible band (the vast majority of energy is detected inside visible filters). Outside the normalization point, the BC assume always negative values, bigger going to higher and lower temperatures. It reach a value of -4 for O stars and -3 for M stars, as visible in figure 1.24.

Be careful - Because BC assumes negative values, $m_{bol} = V + BC$ is smaller, which means brighter (it includes much more flux).

In general, the BC depends on parameters of stellar atmosphere so it is sensible to luminosity class an metallicity of the star. Of course, bigger are the corrections, bigger are the uncertainties connected to the transformation. So, the theoretical reconstruction of HR diagram will be much more uncertain going far away from F5V type of stars.

The HR diagrams (figure 1.25, or better the isochrones, is the set of points that a stellar population forms on HR diagram when stars, that compose it, have been formed in the same time. Theoretical HR diagram must be translated into observable terms such as visual magnitude in order to measure the luminosity and the color index to obtain the temperature. They can have on x-axis the color index $B - V$ (increasing to right) or temperature T (increasing to left).

The BC is responsible of many effects visible in color-magnitude diagrams of evolved populations, such as:

- the rapidly decrease in terms of luminosity of horizontal branch to the blue limit (that is for stars at high temperature). In this branch there are stars which burn He into C in the core and H into He in shells around the center. Practically here stars are so cold that relevant fraction of energy is emitted outside the visible filter. In bolometric there is an increase but no more in

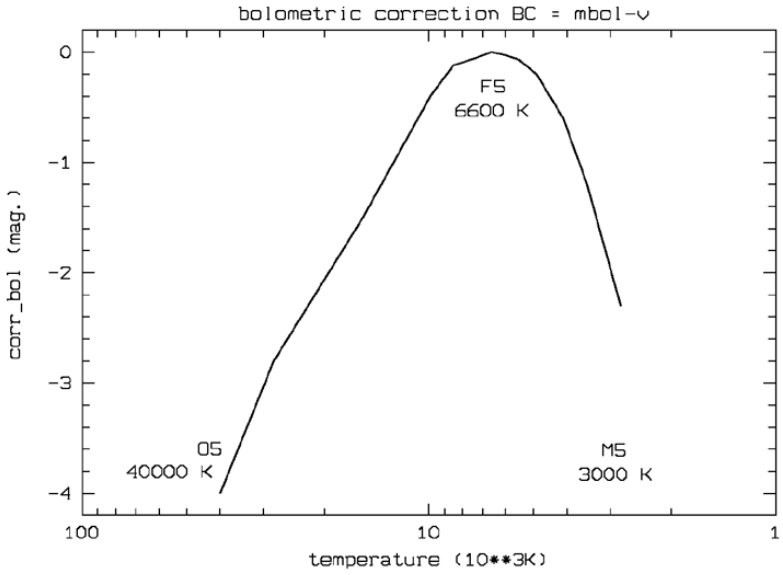


Figure 1.24: Bolometric correction.

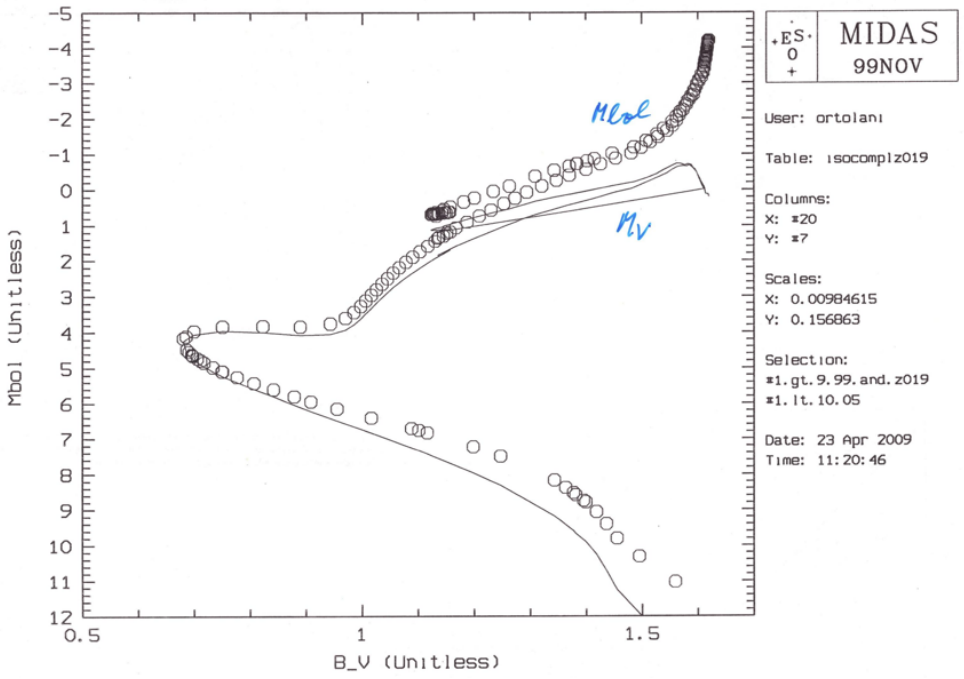


Figure 1.25: HR diagram.

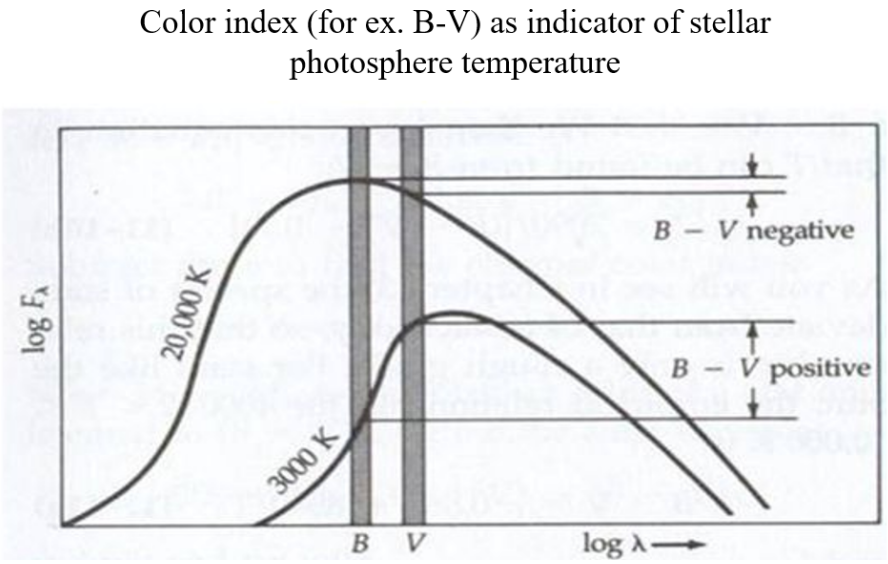


Figure 1.26: Color index as indicator of stellar temperature.

visible band. So near the giant type the luminosity in the visible band is going down because BC increases;

- the growing slope of main sequence stars with lower temperature (metal poor stars that burn H into He in the inner part in the first part of the graphic). In this area BC increases and turns down the MS line which would be instead much more linear with bolometric luminosity;
- the turn off in asymptotic branches of massive globular cluster rich in metals.

1.14 Temperature and color index relation

Now the problem is to pass from color index information to temperature information. How to get it? Consider the figure 1.26.

In figure 1.26, X-axis is the $log\lambda$ and Y-axis is $logB$ where B stands for brightness (or flux, for optical astronomers). In this graphic are visible black body profiles defined by specific temperature T and specific color index, in this case $B-V$. In particular, at lower T, the color index increases because of the scale magnitude. So color index is an indicator of temperature, until the emission peak is more or less inside the range of wavelength of filters used for photometry.

Now suppose to observe a star that is so hot that color index is on the rising branch which is the Rayleigh-Jeans approximation. Here, different black body profiles have almost parallel slopes. So, if you are outside the emission peak, color index is almost constant, even if at different temperatures T. So color index is a good indicator of temperature only close to the emission peak. If you are outside it, even on both sides of the flux distribution, the slope is almost independent from T.

So, what is the relation between T and color index, for example $B-V$? It is necessary to enter with λ_B and λ_V in Planck equation to get the difference between two planckian equations at different wavelength. Practically:

$$B - V = -2.5\frac{\int B_B T_B d\lambda_B}{\int B_V T_V d\lambda_V} + C_{BV} \tag{1.33}$$

where factor -2.5 is due to magnitude scale, B_B is the black body distribution in B band, T_B is the transmission in B band and C_{BV} is a normalisation constant. So $\int B_B T_B d\lambda_B$ is the convolution of the black body with the transmission in corresponding filter. Of course B_V and T_V concern the V band.

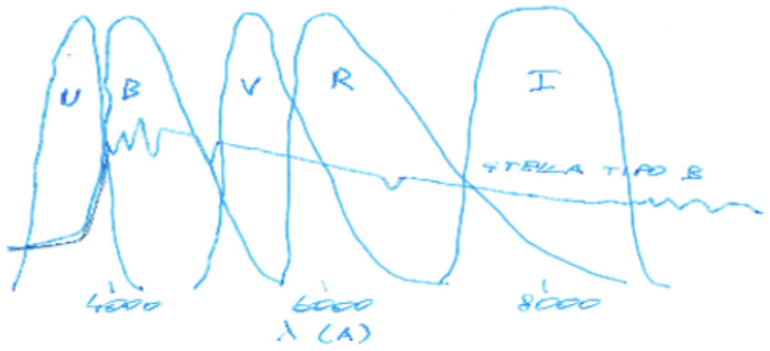


Figure 1.27: Spectrum of a B star and filters.

Then, after some normalizations and algebraic operations quite complicate, you get the relation:

$$B - V = -0.865 + \frac{8540}{T} \tag{1.34}$$

So, in principle, measuring $B - V$, you obtain the temperature T and vice versa.

When T is very high, the second term goes to zero and it is eventually negligible, compared to the constant. The consequence of this fact is that color index is no more sensitive to temperature. As see before, outside the peak there is the Rayleigh-Jeans approximation for which slopes of different black bodies are almost similar so $B-V$ is almost constant and you can distinguish anymore different temperatures of stars.

If the peak of an object is outside the filter, one possibility is to use another filter in another wavelength. In general, we use $B-V$ but if you use another color index, for example $V-I$, you extend the sensitivity of color index to temperature to cooler stars (near the infrared) or if you use $U - B$ you can analyse very hot stars.

In particular, the situation is quite complicate for lower temperatures for which there are some absorption due to the atmosphere because of some complex molecules and you don't have anymore a black body.

So a basic concept is: derivation of temperature from color index must be studied depending on the type of star.

In figure 1.27, is illustrated a B star spectrum and filter bands. It is possible to see clearly that in the case of a B star, the peak is between UV and B bands. So filters like $V - R$ or $V - I$ should not be good because in the correspond area of the graphic, the spectrum has almost a constant slope.

In figure 1.28 there are two isochrones related to color index $B - V$ and $V - I$. In this graphic it is visible that when go to very cool stars on the giant branch, having the highest color index $B-V$, after a certain point there is a sort of saturation so this color index is anymore sensitive to the temperature T. It saturates because this corresponds to a point outside the emission peak in the plankian distribution. Indeed the color index $V - I$ at this point is still close to the peak and it is still sensitive to T. This suggests that during planning observations, the choice of filters is very delicate.

For extended objects composed by different stellar populations, it is not so different and the present considerations are similar.

Remember that the majority of temperatures of stars are coming from color index measurement.

In general we need a very high photometric accuracy in order to reach a very high accuracy on T in order to get a precise age or chemical composition. Usually it is required an accuracy better than 100

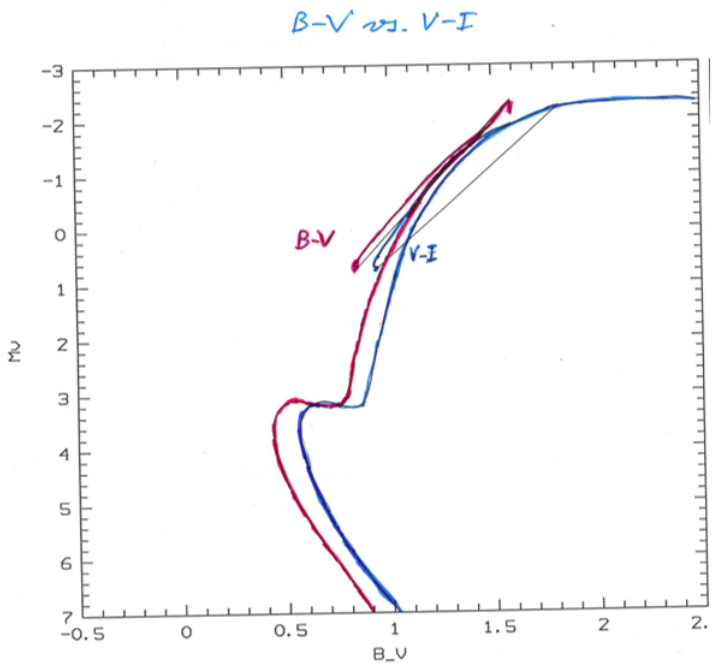


Figure 1.28: The effects on color-magnitude diagram of different color indices: $(B-V)$ vs $(V-I)$.

K, if possible 50 K that is a very small quantity for stellar temperatures.

Modern alternatives to color index There is a new different way to determine star temperature. It consists in using the ratio between absorption lines visible in the spectrum. Indeed their depth is linked to the temperature T, for Bolzmann and Saha equation. Typically, they use iron lines, because they are common in star spectrum and the temperature calculated is the *excitation temperature*. It is not the effective temperature; one comes from observing the atmosphere while color index comes from photosphere which is an opaque layer so they somehow should be different.

There is a problem: temperature from lines is derived assuming some constant values of energy at which the excitation occurs but they are not very well known. For this reason temperature from color index is still the most used.

Infrared and radio astronomy If we go to larger wavelength (infrared and radio range in the spectrum), scientists never use the color index to get the temperature. Indeed they are in a regime where the slope is almost constant (Rayleigh-Jeans approximation) so in this range color index is not sensitive to temperature. So they use other techniques.

For example, radio astronomers measure the brightness at a specific wavelength so in figure 1.26 they identify a specific point in the graphic. We know that there is only a specific black body distribution crossing this point because plankian functions never cross each other and every black body distribution correspond to a specific temperature. So it is not important the shape of the black body but the value of brightness.

The problem is to get the brightness because they need the absolute brightness. The brightness by definition is not depending on the distance so the problem is not to know the distance. It is also affected by interstellar reddening but this is low at long wavelength so it is quite negligible. So the problem here is the following: the measure of surface brightness requires measurements of the flux and the angular size of the target:

$$B \propto \frac{F}{solidangle} \tag{1.35}$$

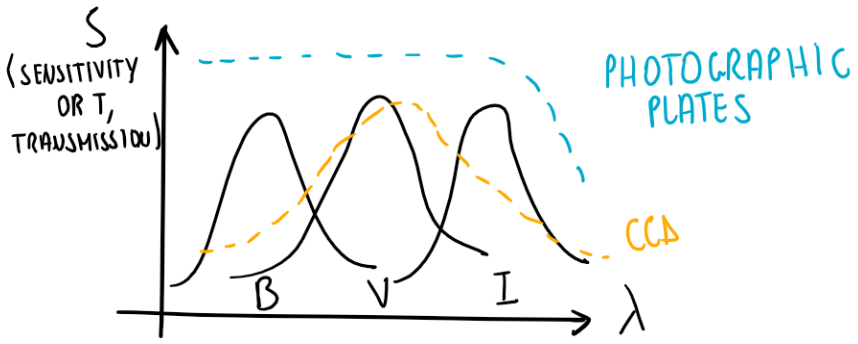


Figure 1.29: CCDs and photographic plates sensitivity.

Usually the angular size is unknown. Only with interferometry you get the angular size for giant stars. For other stars you don't have information about it.

You may argue that if we have the distance and the model of the star, so the estimated size, you can get the angular size but this is a very articulate way to get the angular size because you need some parameters, that you eventually want to derive. It is absurd!

So in radio they don't measure stars but mainly extended objects like HI clouds, stellar formation clouds, supernovae remains and so on. This is the main difference between optical and radio astronomy: radio astronomers work with resolved objects, having a good resolution thanks to interferometry between big radio telescopes.

1.15 From instrumental to international magnitude

In general they use the color index to derive temperature of the star. However there is also the interstellar absorption effect and it is important to take it in account. Indeed it changes the color index measured because of the absorption at different wavelength (Whitford law, 1.11). For this reason we have to go into fine details for both, the color index and the interstellar reddening.

In the last 50 years, since the first international photometric system was established (Johnson system), there has been an evolution of detector. A big advance was the replacement of photographic plates with solid state detectors, like CCD and CMOS. However they have a different spectral response.

As visible in figure 1.29, which has on X-axis the wavelength and on Y-axis the sensitivity S (or the transmission T), the CCDs have a more extended red response while their sensitivity goes down in the blue band. If we refer to the international system (in the graphic B, V and I bands) and we make a convolution of transmission of modern CCDs in those bands, we can notice that there is a shift of the convoluted band to longer wavelength on the red part of the spectrum. This convolution contains a bit the V band, the R band, where is located the peak of sensitivity, and the first part of the I interval, going to shorter λ. So the sensitivity of CCDs is not flat.

Instead, on old photographic plates, the function of sensitivity is much more flatter, especially on the blue side of the spectrum and it decreases at longer λ.

Therefore if we take the right filters but with different detectors, we get a different spectral sensitivity. To resolve this, there are some possibilities:

- to chose a different international system. The Johnson one is the most famous and the most used but it is not the unique;
- to change the filters. The idea is to built the filters with some new technology in order to have a different slope and to get the convolution having the same spectral sensitivity of the filters.

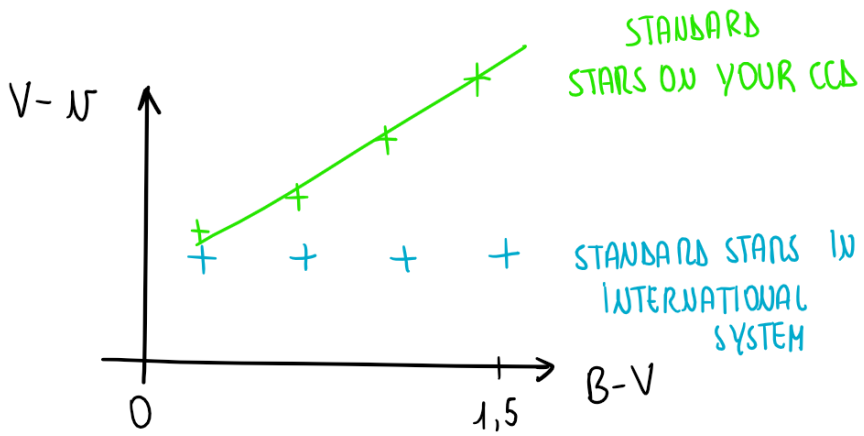


Figure 1.30: $(V - v)$ vs $(B - V)$.

Nowadays, scientists try to find a remedy of this problem observing the same object in different wavelength.

The consequence is that, when it is necessary to do the calibration of data, we have to proceed though a process which is the conversation from the instrumental magnitude to international system. Indeed, on the telescope we get values that are not in international system, depending dramatically on the specifics of the detector.

So it is necessary to convert instrumental magnitude to the international one: they can be similar but they aren't the same and the difference is relevant.

In order to make this conversion, we have to proceed with two important points.

1° step First it is necessary to observe and measure a number of standard stars with different temperatures (or color index) and create a kind of plot such as figure 1.30.

In this graphic V is the calibrated magnitude for that star (what we want to know), v is the instrumental magnitude (what we measure) and $B - V$ is the color index in international system (usually available on catalogs).

In general, the instrumental magnitude is give by:

$$v = -2.5 log C \tag{1.36}$$

where C are the counts obtained during observations with a specific detector (eventually normalized). So v is negative per definition because counts of electrons are positive and in front of the logarithm there is a sign minus, given by inverse scale used for magnitudes in Pogson formula.

Consequently $V - v$ is a positive number, generally a big one, about 25 until 40.

So for standard stars, if the system is right, the spectral sensitivity is identical to the international one, it is expected a constant plot which means a set of stars having $V - v$ constant (blue plot in figure 1.30).

In particular, $B - V$ about 0 corresponds to hot standard stars while $B - V$ about 1.4/1.5 corresponds to cooler stars. This is the saturation limit for this type of color index.

Now suppose that, instead, the sensitivity is pushing the average wavelength a bit shifted to the red (as occurs in modern CCD). In this way, more red photons are collected than international system. So, a very red star can produce more flux in red than it is expected in international system. As consequence plotted stars create a straight line with a slope, less in the blue and bigger in the red

range (green plot on figure 1.30). Mathematically, $v < 0$ because counts are growing so $V - v > 0$, growing as well.

In general this plot is not dependent on brightness; it depends on color index, so on temperature. The slope indicates that in our instrumental system we have more flux than a given standard star with a specific temperature. If the system used during observations is not strongly different from international one, the green plot can be approximate with a straight line.

Recapitulating, the first step in calibration process is to measure a number of standard stars, eventually spread in a wide temperature range. If you look at the catalogs, most of the stars are Sun-like stars because they are common around the Sun so the most of standard stars are crowded around a color index $B-V$ about $0.6-0.7$. Then you create a plot such as figure 1.30 that shows how the sensitivity range of our device is different from international photometric system.

Then, it is valid the following relation:

$$V = v + K_V(B - V) + C_V \tag{1.37}$$

where V is the calibration magnitude so what we are searching, $B-V$ is the international color index on X-axis and C_V is a constant to normalize the equation. K_V is the *angular coefficient* of the straight line and indicates how much is the slope, which is always positive. In general the order of magnitude for a relatively good system is about 0.1 magnitude in an interval of 1 magnitude.

So, in the case of a good system $V - v$ should be constant but if there are some differences from international one, $V - v$ grows and, with it, also the slop of the straight line. Suppose to observe standard stars with calibrated magnitude $V = 0$, the equation 1.36, gives the counts for $V = 0$. This is very useful because from Johnson system we perfectly know how many photons per second per surface unit are received by a star at zero point of absolute magnitude. Knowing this number, it is possible to estimate the efficiency of a specific system.

2° step Now the goal is to determine V for unknown stars which have been measured obtaining the instrumental magnitude v. Remember that $V = v + K_V(B - V) + C_V$.

K_V is known, derived on plot for standard stars: it is the angular coefficient of the straight line. Usually it is about 0.1 in 1 magnitude in color index axis so ignoring this term means an error about 10%.

C_V is the normalization constant: it is know as extrapolation to zero point.

However $B - V$, the color index calibrated on international system is NOT known. It is known for standard stars but not for unknown stars. You should know this calibrated color index, or eventually the temperature T to transform in color index, or eventually the spectrum or something telling you where to put the star on X-axis in order to get the right correction. How do it?

From algebra, it is necessary to find another independent equation in order to have a linear equation system that we can resolve. This second equation is given by using a different filter for the same star with same color index so it is exactly the equivalent of the first one. Remembering equation 1.37 and the new one, we get the linear system:

$$\begin{cases} B = b + K_B(B - V) + C_B \\ V = v + K_V(B - V) + C_V \end{cases} \tag{1.38}$$

Subtracting the first equation to the second one, we obtain:

$$B - V = \frac{b - v}{1 + K_V - K_B} + \frac{C_B - C_V}{1 + K_V - K_B} \tag{1.39}$$

where $B-V$ is the calibrated color index, $b-v$ is the instrumental color index, K_V and K_B are the slopes of the straight lines in the two bands determined from standard stars, C_B and C_V are the zero point, also determined from standard stars.

Knowing the color index calibrated on international system $B-V$ (it is a linear system so there is only one solution), you can find then V and B, the calibrated magnitude.

The main consequence is that when it is necessary to calibrate a color index, it is not just an offset of the instrumental one so it is not only a difference between constant because there is also the factor $\frac{1}{1+K_V-K_B}$. This term is very important. It makes the scale of $B-V$ different from $b-v$ scale depending if it is bigger or smaller than 1.

- $\frac{1}{1+K_V-K_B} > 1$ means that the scale of $B-V$ is expanded compared to the instrumental one. In this case $K_B > K_v$ because going to shorter wavelength, the deviation from original system is bigger (so K_B is higher) telling us that there is a relevant shift in the band. In particular the scale of $B-V$ is larger and scale of $b-v$ is compressed compared to the standard one because the two convoluted wavelength (b and v) are closer each other so it miss a bit of sensitivity of temperature or color index. When they are extremely close, you miss complete the sensitivity because the difference is so small that you can't see differences in temperatures. In this case it is necessary to go further away to increase the sensitivity;
- $\frac{1}{1+K_V-K_B} < 1$ means that the scale of $B-V$ is compacted compared to the instrumental one.

Instead $\frac{C_B-C_V}{1+K_V-K_B}$ is just the zero point, a constant, that shifts the plot.

So, analyzing color-magnitude instrumental diagram, the calibration is not just a shift; it is also a change in scale because of the **color correction term** $\frac{1}{1+K_V-K_B}$. It depends on coefficients K_V and K_B of the slopes of straight lines in the two different bands, which have been determined by plot of standard stars. In particular K_V and K_B tell us how much is the deviation of a specific system (a specific device) from standard one. In perfectly instrumental system K_V and K_B should be zero (which means horizontal lines in the plot with standard stars) so the color correction term should be 1. In this special case the $B-V$ and $b-v$ scales should be identical and there should be only a simple shift in according to the difference of the zero point.

In general, this correction is good but not perfect and it is just a model correction. However we obtained a good process to calibrate a specific system with the international one measuring standard stars. But this correction is good only for those kind of stars. Moreover now it is necessary to take in account also the interstellar reddening effect and correct it.

What it has been done here, in the 2° step of calibration process, is only the **first order approximation**; it is the easiest approach telling us that it is necessary to modify the $b-v$ instrumental color index, not only for the zero point (a shift) but also because there is a different slope in the two bands to get the calibrated one, which means a change in scales.

Process details There are several issues behind this simple process.

- It is obvious that the derivation of plot 1.30 is very delicate and depends strictly on the choice of standard stars used for calibration. In principle, standard stars should be the same kind of the stars you have to observe in order to have a similar spectral distribution, so a similar metallicity.
- Moreover stars should have also a comparable reddening. Whitford law (1.11) explains that the absorption at shorter wavelength is bigger than to longer wavelength so the reddening effect is bigger at left of the peak than in the right wing, as visible in figure 1.31.

  It this case the shape and slope are different from the intrinsic one. It is not just a scaling down, it is also distorted and the peak is quite different. Of course the absorption lines are instead exactly in the same position.

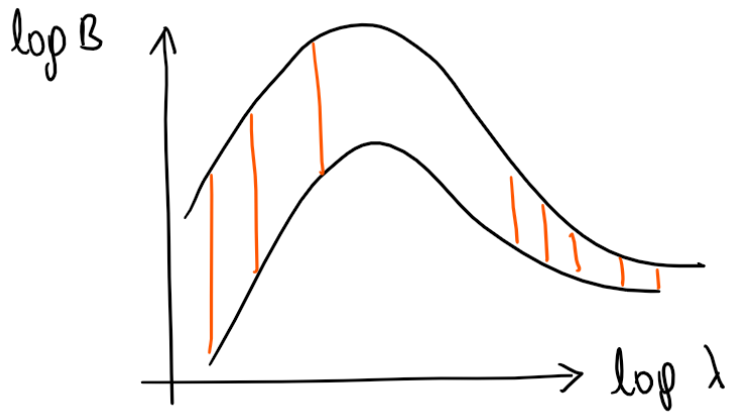


Figure 1.31: Reddening effect on a black body distribution.

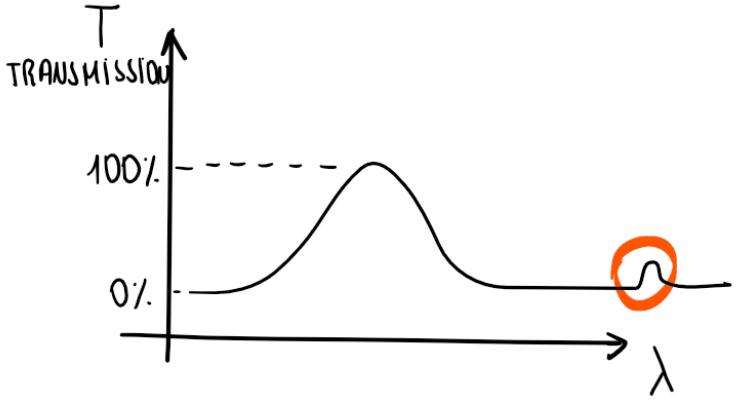


Figure 1.32: Red-Leak effect: a CCD is sensitive under 1 microns while the red jump is about 0.9 microns.

So, using a mixture of stars with different interstellar reddening and different kind of spectra you may have a spread of points, a dispersion around the straight line in figure 1.30.

Of course stars having higher $B - V$ have higher probability to be reddened stars. In general a cold or hot star but reddened have similar color index.

ATTENTION: U band is strongly affected by the Earth atmosphere because it is close to the visible limit. So, transmission of the U band is affected by atmosphere every night because of the aerosols, changing slope of calibration. Therefore the calibration in the U band is a delicate operation to repeat every night and sometimes, during night as well.

1.16 Red-Leak effects

A secondary reason for system deviation from international one is Red-Leak effect, defined as the contribution of radiation that filters at infrared emission, especially from 8000 to 9000$\mathring{A}$, where the transmission can be about $10^{-3} - 10^{-4}$ with respect to the peak of the principal band. It is a small value but not negligible when we compare photometry of stars with very different temperature. The result of this effect is to give a positive contribution to the color correction term K.

This effect is visible in figure 1.32.

From technical point of view, the Red-Leak effect derives from the difficulty to built filters able to block red-infrared radiation maintaining a good transmission in the blue band. In the past they used often the sulphate of copper ($CuSO_4$) that eliminates completely the red radiation at liquid state but in this way it is unstable and corrosive, while at solid state it absorbs water present in the atmosphere.

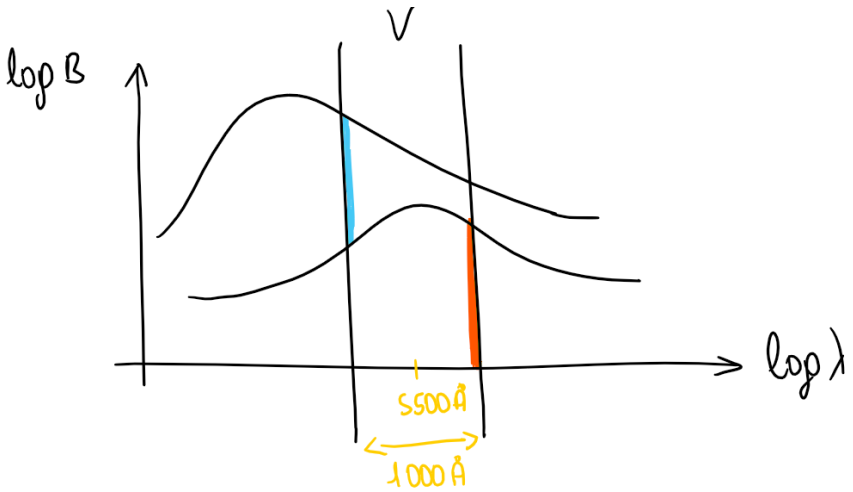


Figure 1.33: V band at different temperatures.

Nowadays there are some special filters equipped with some thermal reflectors in infrared but they are not easily reachable.

The problem is quite frequent and it is difficult to identify and quantify both on sky and in laboratory but it is extremely important for U and B band.

A qualitative test that can be done, in order to detect the presence of a significant Red-Leak effect, is to examine how flat fields appears in the considerate band compared to the extreme red. A similarity to intermediate bands can be the signal of the present of Red-Leak effect.

1.17 Second order interstellar reddening effect on photometry

Suppose to have two spectrum of two stars with different temperature, one hotter and one cooler with the peak a bit shifted. Suppose to measure brightness in V band, centered at $5500\mathring{A}$ with FWHM about $1000\mathring{A}$, so a big fraction of the entire spectrum.

It is easy to see in figure 1.33 that for the lowest temperature inside the same filter in the same system, there are more red photons (in proportion) on red side of the filter than in the blue side. For higher temperature there are more photons, in proportion, on blue side than the red one.

Considering this and applying the reddening effect we will understand what happens. Applying Whitford law (1.11), the reddening is more effective on blue stars than on red one, inside the same filter. So the reddening effect is not only dependent on the photometric band but it changes also in the same band depending on temperature of the star: this is the **SED** Spectral Energy Distribution. In the case of stars nearby the Sun, SED is just a function of temperature.

Suppose now to consider a color-magnitude diagram, an observer one in figure 1.34. What is the first order effect of the reddening?

- Considering the A_V absorption in the V band, it gives fainter apparent magnitude so the all diagram is shifted down.
- Reddening effect affects also the graphic in color axis. Remember the definition of color excess such as $E(B-V) = (B-V)-(B-V)_0$ where $(B-V)$ is the color index observed and $(B-V)_0$ is the intrinsic one (un-reddened). So the reddening effect is a vector to right, shifting the graphic to redder colors because, as we see before, reddening is more effective on the B band (shorter λ) than on the V band (longer λ).

Combining this two effects, how shown in figure 1.34, the result is a vector with a specific slope that causes a **roto-translation** of the diagram. The slope of this vector is given by the ratio:

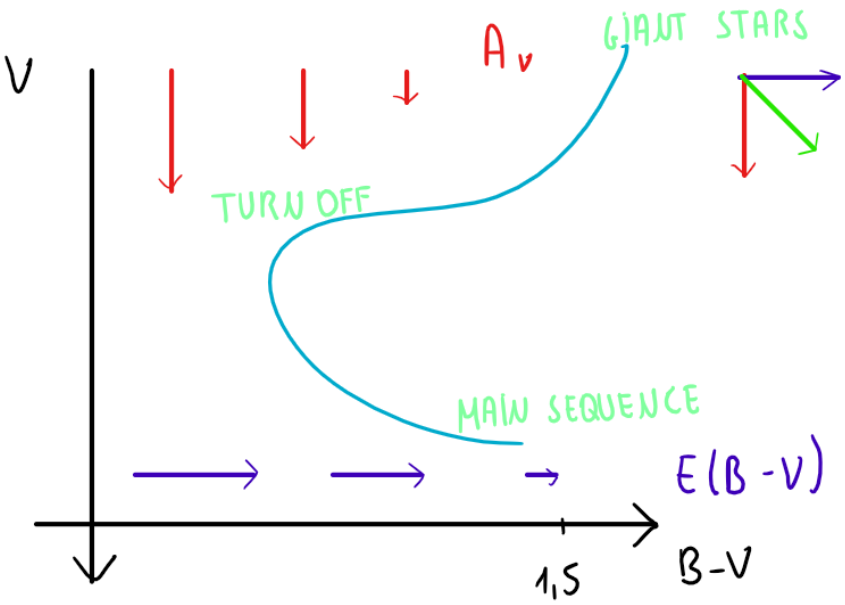


Figure 1.34: Observer color-magnitude diagram.

$$R_V = \frac{A_V}{E(B-V)} \tag{1.40}$$

that is usually about 3.1 for O and B stars. Choosing different type of stars, so different temperature, also this constant changes.

This effect tells us that the observation is not constant thought different temperatures of stars because the spectrum is different. A_V vector should be smaller for red stars than for blue stars, because red ones contain less blue photons inside the filter range. This is due to the fact that filters are not monochromatic or extremely narrow but they are usually very wide and, inside them, the slope of SED could change (changing filters, you change the sensitivity to the reddening).

So the final result, as seen before, is a color-magnitude diagram shifted and rotated in a specific direction and with a specific slope.

Actually, the situation is quite more complicate because also excess vectors change with temperature of the star and this is quite obvious: bluer stars, containing more blue photons, have a higher $B-V$ than red ones.

The consequence of this is a **compression** of the diagram scale because left points (bluer stars) are shifted more than right points, which means red stars.

Resuming: R_V increases as temperature T decreases, which means $(B-V)_0$ intrinsic color increase because $E(B-V)$ decreases much more rapid than A_V. Instead A_V decreases as function of $(B-V)_0$.

How can we correct or compensate these effects? There are some corrections linked to the color excess described by complicate mathematical treatment.

Mathematical treatment The situation is quiet complicate: simultaneously, absorption A_V and the color excess $E(B-V)$ change for stars at different temperatures so it is necessary to keep a quantity fixed in order to derive how much they change. It is possible to fix the amount of dust but we are not able to quantify the amount of dust so the absolute absorption is always relative to the source we are measuring.

Therefore, to find the analytic solution to this problem is not easy. The problem has been studied in details by Schmidt-Kaler in 1961 and then by Fitzpatrick in 1999 and Grebel and Roberts in 1995. In particular Schmidt-Kaler demonstrated numerically that, with a fixed R_V ratio, the color excess $E(B-V)$ depends on the intrinsic color of the source, decreasing for cooler stars such as consequence of the absorption dependence on wavelength. In this way, the ratio R_V, as function of absorption and the intrinsic color, is given by:

$$R_V = 3.0 + 0.14(B-V)_0 + 0.025A_V \tag{1.41}$$

derived measuring stars with different temperatures, reddening and with intrinsic color $(B-V)_0$ known (Schmidt-Kaler got $(B-V)_0$ by using spectral features so using the intensity of absorption lines of known spectral type).

So, from equation 1.41, it is clear that the ratio R_V changes with intrinsic temperature of the star, $(B-V)_0$, and the absorption, A_V. This implies a variation about 5% in R_V from a star of $B-V=0$ and a star with $B-V=1$, being equal A_V. The dependence of A_V becomes considerable, and comparable to the variation due to different spectral types, only for absorption values bigger than 5 magnitudes.

In alternative, the author gives also a different but equivalent expression:

$$R_V = 3.0 + 0.2(B-V)_0 + R_1 \cdot E(B-V) \tag{1.42}$$

where $R_1 = 0.026 - 0.007(B-V)_0$.

- So, for hot stars, for which $(B-V)_0 \simeq 0$, the ratio $R_V \simeq 3.0$.
- While, for cool stars, for which $(B-V)_0 \simeq 1.5$, the ratio $R_V \simeq 3.2$.

Is the difference of 0.2 relevant? It depends on the R_V definition, remembering that $R_V = \frac{A_V}{E(B-V)}$. In order to understand what we are doing, it is important to know the typical process used in observational astrophysics.

- First, it is necessary to derive the color excess $E(B-V)$, for example comparing theoretical models of stellar evolution.
- Then they derive the R_V and last the A_V. This is the typical order of operations.

The A_V is used then in the distance modulus formula. So, also this tiny difference affects the measure of distances with an error of many factors.

So, it is very important to correct the absorption ratio R_V for the intrinsic temperature T of the star $[(B-V)_0]$.

Schmidt-Kaler has tabulated, integrating over the spectrum, the ratio between the color excess of a specific stellar type and the color excess of a B0 star:

$$\eta = \frac{E(B-V)}{E(B-V)_{B0}} \tag{1.43}$$

where η is a complicate function of reddening and color excess of the star. To know η is very important: it allows to obtain the reddening for a specific spectral type, knowing the B0 one, used as reference. Indeed the B0 stars, very hot, are very common in literature and all reddening laws are set on this type of stars.

- Obviously, for B0 stars, used as reference, $\eta = 1$.
- For example, for M5 type, $\eta = 0.89$.

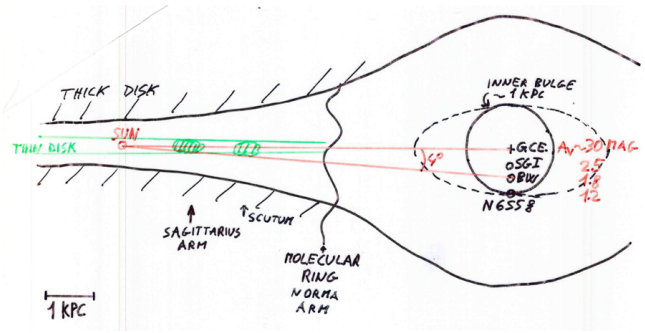


Figure 1.35: Framework of the reddening inside our galaxy.

So there is a difference about 10% for several spectral type. This means that, when you give the reddening in some direction of the galaxy, you have to clarify for which kind of stars or for which temperature the color excess was determined, in order to apply the right correction for A_V.

1.18 Interstellar reddening maps

Because interstellar reddening is a very important effect to take in account for observational data, it is fundamental to know where dust and gas are located and in which quantity in order to estimate the absorption.

In figure 1.35, it is visible a galaxy map.

In this illustration there are two sight lines, one directly to the Galactic Center (GC) and the other one from 4° from the GC, still inside the inner bulge.

As shown the absorption scale in magnitude in the figure, the absorption is very high to the GC where A_V is about 30 in magnitude. Then there are some windows showing that the reddening is dramatically decreasing at very small angles, out of the line of sight crossing exactly the thin disk. Indeed, inside the thin disk, there are different gas and dust clouds (cross section of the inner spider arms). In particular, in the area observable, most of the reddening is located inside Sagittarius Arm and Scutum. They are the reasons why we measure $A_V = 40$ in the GC direction: there is no relevant reddening at the centre of the galaxy (the bulge is basically free from interstellar matter) but is very high in the thin disk where we are located. To be specific, the reddening is confined in the thin disk, within a line (a ring in three-dimensional representations) rich in complex molecules and active stellar formation which sets, indeed, the limit of interstellar matter, so the limit of the galactic disk. Inside this area there are many kind of stars.

So the major problem is when observations are pointing the galactic disk or are crossing the nearby arms of the galaxy.

Of course, if observations are direct perpendicularly to the galactic disk the extinction is minimal but sometimes not negligible because the reddening is distributed non-uniformly and there are small clouds also at relatively high galactic latitude. In general, for any direction, reddening should be investigated.

Example: the GC The reddening effect is very important looking to the GC. In particular an image took in visible band, about 5500 $\mathring{A}$, shows only bright nearby stars and the GC is invisible. In Gunnz band, 8000 $\mathring{A}$, thanks to a bigger wavelength in Whitford law, a few clusters in the bulge appears. Finally, in infrared band, thanks to a very long λ, the reddening effect is almost complete deleted and many nuclear clusters are visible.

Sun position Observing figure 1.36, it is possible to see that the Sun is in a very lucky position because it is situated in a sort of empty zone in the galactic disk called **chimney**. It is not clear why

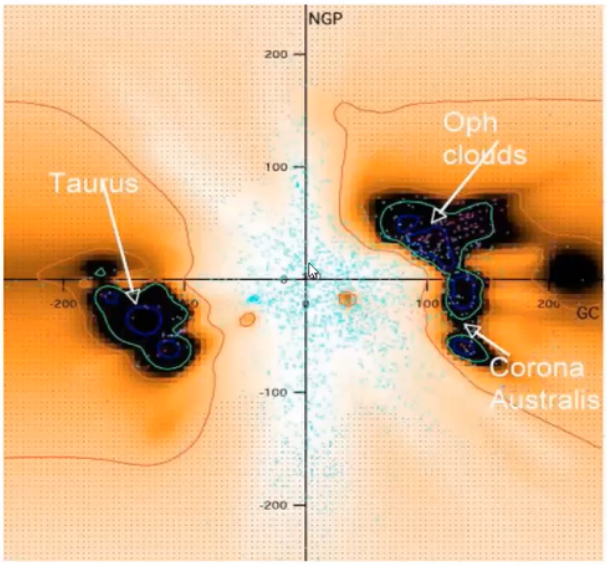


Figure 1.36: The local "chimney".

there is a sort of bubble inside the disk: maybe it has been caused by a supernovae explosion of more supernovaes which wipe out all the gas.

Is the reddening constant? During observations in a specific direction, the Sun with planets is moving across the galaxy, such as the interstellar matter, at a speed of about $20km/s$. So, is the interstellar reddening a constant effect?

There is no literature about this question but there are few papers in radio astronomy about a particular experiment. Some radio astronomers showed how the CO absorption lines in interstellar matter is changing looking at the same objects with different paths. Different paths mean different content of dust, so maybe there could be a dependence of this effect on time. However, nowadays we don't have studies or data about this issue.

Chapter 2

Measures of distance

The most used methods of measurement of distances currently used are:

- Radar
- Annual trigonometric parallax
- Group parallax
- Spectrophotometric parallax
- Nebular parallax
- Method of differential galactic rotation
- Signal dispersion (pulsar)
- RR Lyrae type stars and luminosity of the horizontal branch (clusters)
- Wilson-Bappu Effect
- Dynamic parallax
- Size of the HII regions
- Brightness / diameter for supernova remnants
- Cosmological distances (Hubble law)

These methods are divided into **direct** methods and **indirect** methods. Direct methods are based on simple geometric considerations and can be used over relatively short distances (Solar System, stars near the Sun).

In figure 2.1 it is reported a schematic illustration of the principal methods to measure the distance.

2.1 The radar

Among the direct methods, radar is the conceptually simpler tool, but with the shortest range of action. It is based on the time between the transmission of the signal and the reception of the echo. Half of the time elapsed is simply the product of the propagation speed and the distance from the celestial body that produced the echo, then it is valid $d = ct$, where c is the propagation speed of the electromagnetic signal (in the vacuum), t is half of the time elapsed and d is the object distance.

The uncertainty of the measurement is due to the shape of the celestial body responsible for the echo and to the assumption on the propagation speed of the signal that is altered by the presence of matter of different nature (atmosphere, ionospheric and interplanetary plasma ...). The refraction index $n = c/v$ indicates how light propagate in matter. For example $n > 1$ in Earth atmosphere and

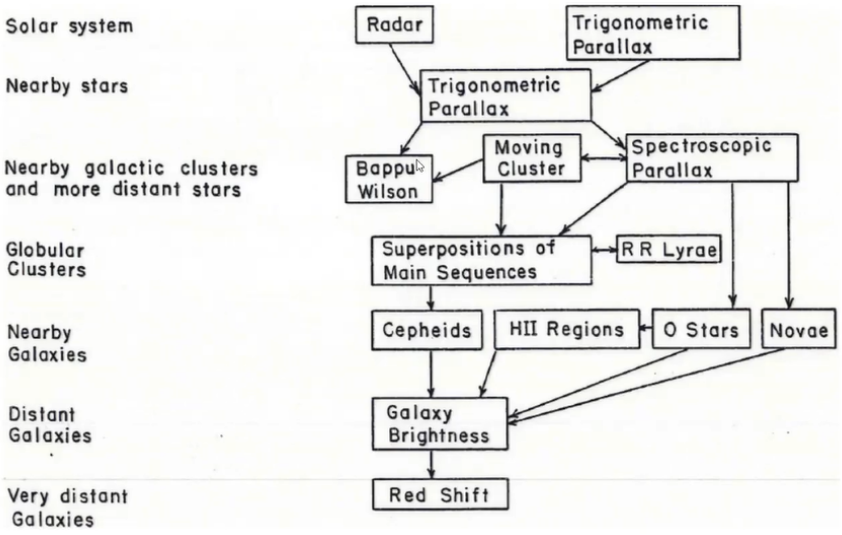


Figure 2.1: Schematic summery of methods to measure the distance.

$n = 1$ in vacuum. Moreover the radar can only be used for very short distances because the intensity of the echo decreases with the fourth power of the distance.

Earth-Sun distance Measurements of distance with respect to Venus, several small planets and, more recently with respect to Mars, have allowed us to derive the Earth-Sun distance with precision of the order of meters or less. The method is based on Kepler's third law and uses the equation expressed in the following form, using the Earth period and distance as a reference. Suppose to consider the Earth and the Sun, the third law is:

$$a^3 = const \cdot P^2 \tag{2.1}$$

where a represents the Earth-Sun distance (half axis of the elliptical orbit) and P is the Earth orbital period (a year), *const* is a constant containing the mass of the Sun, the mass of the Earth (eventually negligible) and the gravitational constant. A priori we don't know the constant because the mass of the Sun can not be easily determined directly. From observational data we only get the period of the Planet around the Sun so we will get the distance from Sun without a zero point: in this way it would be possible draw the Solar System in scale but missing the zero point of the scale. To establish the zero point it is necessary to measure distance from two planets, for example Earth-Venus, Earth-Mars, Earth-Minor planet and so on. From mathematically point of view the solution is simple because we can use another law, which is:

$$(a + x)^3 = const \cdot P_1^2 \tag{2.2}$$

where x is the measured planet-Earth distance and P_1 is the planet period, easily determined by observations. The constant is the same. The term to the left of the equation represents the distance of the planet from the Sun. Then, once the term x is known (with radars, parallaxes, etc.), the distance a is immediately obtained.

A schematic illustration is visible in figure 2.2.

The first measurements of the Earth-Sun distance date back to Aristarchus of Samos (310-230 BC) which derived the distance of the Sun from the Moon-Sun angle to the first quarter and concluded that it had to be about 20 times the Earth-in terms of Moon distance. This measure, as Kepler has shown, is far below the real distance. The cause of the error was due to the lack of knowledge of atmospheric refraction at the time which deviates the path of light. Later Cassini (1672) measured

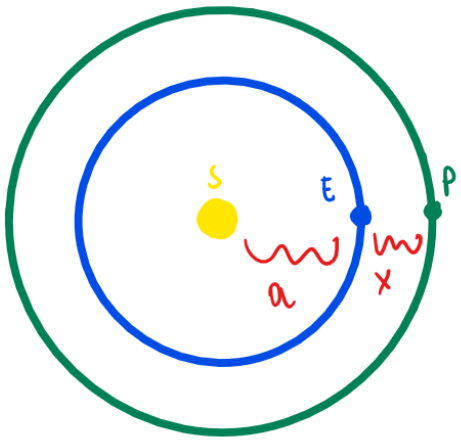


Figure 2.2: Measurement of Earth-Sun distance.

the astronomical unit from the distance of Mars obtained through the parallax. The first precise measure in modern times is that due to Jones (1932) who used the minimal approach of the Eros planet ($22 million km$) to derive the distance with the parallax (with an error of about $100000 km$). A further significant improvement was possible with the use of distance measurements from Venus, with the radar (Pettengrill, 1966) which led to an accuracy of about $1000 km$. Radar measurements on Mars, until recently, have not led to significant improvement due to the orography.

The best current determination, derived from the telemetry distance of the probes sent to Mars (Viking, Opportunity, etc.), is reported by Pitjeva (2005):

$$AU = 1.495978706960 \times 10^{11} \pm 0.1m \tag{2.3}$$

Note that due to eccentricity, the Earth-Sun distance varies from 1.02 to $0.98 AU$. Several authors have, however, found a progressive increase in the Earth-Sun distance of about $15m$ per century, significantly greater than the measurement error, which does not find an obvious explanation so $\frac{dA}{dt} > 0$.

Among the proposed hypotheses we remember the cosmic expansion, the loss of mass of the Sun (but assuming $6 \cdot 10^9 kg$ per year it would lead to only $0.3m$ per century), the variation of the constant G and the dark matter. But up to now none of these seems sufficient from the quantitative point of view. A recent attempt (Miura et al., 2009) indicates as possible cause the exchange of Earth-Sun angular momentum due to the tide, by analogy with the Earth-Moon system. Note, anyway, that these distance measurements are related to a time and the measurement of the speed of light. This point should be deeper analyzed.

2.2 The annual trigonometric parallax

It is a very simple method and it is the fundamental direct one. It is based on the well-known principle that consists in the observations of a celestial body at a distance of six months, relative to distant objects, using as base the axis of Earth revolution. So, we don't use the Earth diameter because stars are too distance: Earth diameter is too small to detect very small angles in order to determine the distance.

As visible in figure 2.3, knowing the AU, which is the half-base, it is necessary to know the angle in order to determine completely the triangle and to get the distance with very simple geometry.

In general, this angle is a fraction of arcseconds so it can not be measured with a telescope because pointing accuracy of telescopes, also the better ones, could be about 10 arcseconds, and even worst in case of pointing punctiform objects. Therefore it is measured the apparent position of the star projected on a very distance background, such as galaxies, quasars, clusters, and so on which are

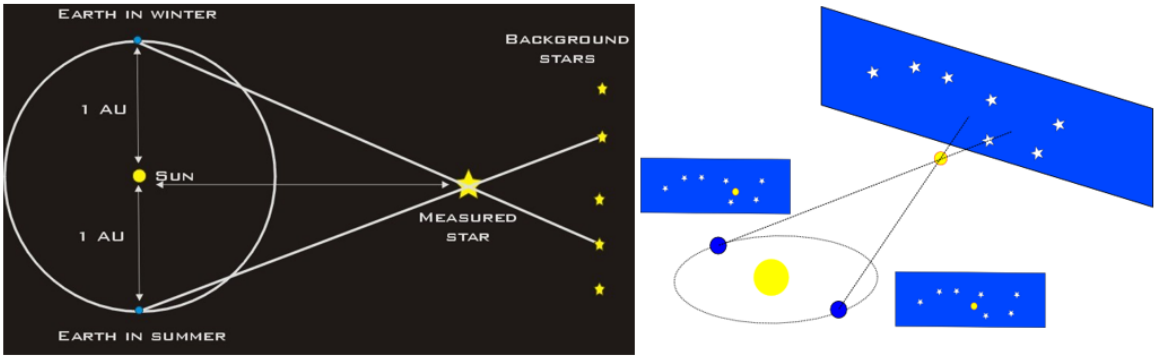


Figure 2.3: Annual trigonometric parallax.

totally independent on the parallax effect thanks to their huge distance. If the background would be not distance enough, background stars would be affected by Earth movement around the Sun and the distance measured would be smaller than the intrinsic one. So, then we measure the apparent position on the sky in celestial coordinates, which means angles and in this way the triangle is fully defined and it is possible to get the distance.

As a first approximation we can assume that the accuracy of the measurement can be expressed as:

$$\sigma = 0.5 \frac{FWHM}{S/N} \tag{2.4}$$

where σ (in arcseconds) is the error on the position of an image in arcseconds, $FWHM$ (in arcseconds) is the Full Width at Half Maximum of the gaussian luminosity profile of the source (easy measurable), S/N (a pure number) is the signal-to-noise ratio of the same image. 0.5 is an empirical coefficient, it can be different, for example about 0.7. This equation is empirical but also quite universal; it can be used also for line position in spectrum.

Taking a typical seeing of 1" and a signal-to-noise ratio around 100, we can easily see that distances up to the equivalent of about 0.01" of parallax can be measured with an accuracy of 30%, which means by definition of parallax, up to a distance of 100 parsecs (more or less the thickness of the galactic disk).

More accurate calculations show that accuracy depends on the seeing motion component and therefore also on the exposure time. It depends also on number of background stars, if they are many the errors would be smoothed down. The adaptive optics systems allow to obtain today much higher accuracy, since the $FWHM$ of the corrected stellar images, in the infrared, approach those of diffraction width which, for the K band, in a telescope of 8 m in diameter corresponds to 0.08". Thus disregarding image motion effects due to seeing, the 30% accuracy would be obtained up to distances of 1000 parsecs.

From Hipparcos space, parallaxes were obtained up to $1600pc$, while with the Space Telescope ACS, up to about the double. Indeed Gaia reaches more or less $5kpc$.

The base of parallax system can be bigger than AU using the distance between Earth and a satellite send into Solar System.

2.3 Group parallax

The group parallax can be summarized simply as the distance measurement of a cluster of stars that moves with a radial velocity known from the spectra, compared to the apparent size variation. We can write the equation $\theta = D/r$ where θ is the angle below which we see the cluster, r the distance and D its diameter, and the derivative of the angle with respect to time becomes: $d\theta/dt = -v\theta/r$ from which one immediately derives the distance r from the radial velocity v known and the measured variation

of the angular dimension with the time. In practice the method is of limited application because it assumes the isotropy of the system and then because the measurement errors make it usefull only for very close star clusters. A more complex variant is based on the measurements of the proper motion vectors and the identification of the convergence point, but the uncertainties on the method remain the same.

2.4 Spectrophotometric parallax

The spectrophotometric parallax requires knowledge of the spectral type of the star under examination and is applicable only to stars of spectral types whose absolute magnitude (main sequence stars and giants) is known. The comparison between apparent magnitude and absolute magnitude allows to directly derive the distance modulus, and the distance itself once interstellar absorption is known. This method therefore belongs to indirect methods.

2.5 Wilson-Bappu effect

It is an empirical, little known, indirect method of which no detailed physical explanation is known. It is based on the equivalent width of the weak emission that appears at the center of the absorption of the lines H and K of the $CaII$. This quantity, called W by Wilson and Bappu (1957), correlates with the absolute magnitude of G-K spectral stars. Calibrated on the Sun and on known parallax stars, this method allows to obtain the distance of stars of which we have the high dispersion spectrum in the H and K lines of calcium with an uncertainty of about 10%. The lack of knowledge of the mechanism that links the emission of the lines with the absolute magnitude constitutes a limit to the method. An attempt to explain it is that the absolute magnitude and therefore the brightness of the star would be connected with the size and motion of the convective cells and from these the existence of an emitting layer. At the moment it is not clear what the influence of the actual temperature or the spectral type might be.

The original relationship of Wilson and Bappu is:

$$M_V = -14.94 log W + 27.59 \tag{2.5}$$

The width W is measured from the ends of the line and the data is corrected by instrumental enlargement.

So, in practise from spectroscopy we measure the equivalent width of $CaII$ line (so we get W) then we enter it in the plot and we measure the corresponding absolute magnitude M_V. Then, using magnitude law with apparent magnitude, we get the distance module and from this the distance. Of course, it is important to know the interstellar reddening.

Using this method, it is possible to determine the distance of individual solar type stars, which fall in the range $1.2 < log W < 2.0$, a very common range including the vast majority of G stars.

In figure 2.4 it is visible that points are regularly distribute along a straight line.

However there are many issues connected to this method.

- It is necessary to have high resolution spectroscopy. Nowadays, using giant and modern telescopes, it is possible to reach a very high resolution, also of extra-galactic sources, but sometimes this is not enough.
- In figure 2.4, there are some point with high dispersion, also about 1 magnitude and the reason is unknown.

Therefore, in average this method is very good but sometimes, some points could have big dispersion and could move away from interpolation.

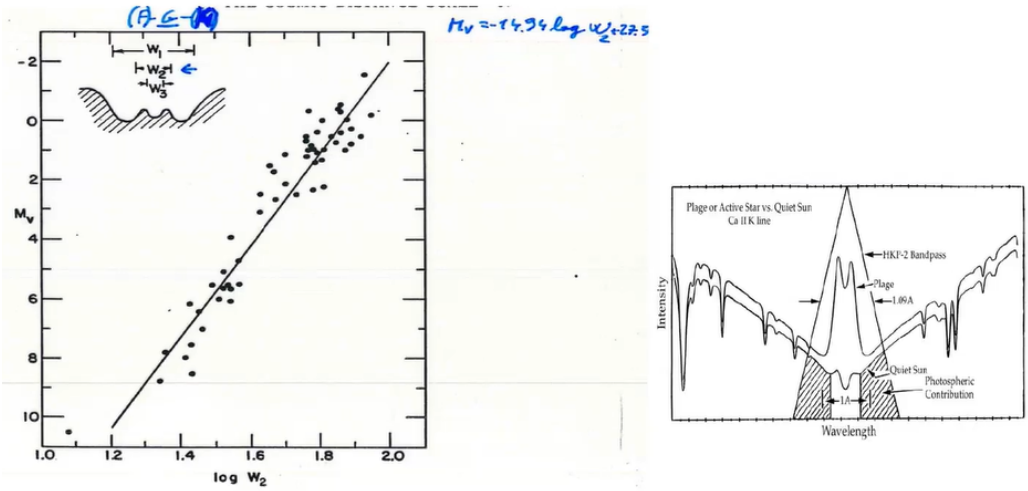


Figure 2.4: Wilson-Bappu effect.

2.6 Nebular parallax

It is applied to expanding clouds (theoretically also to shrinking clouds) and is based on the simultaneous measurement of radial and transverse velocity by proper motions. If you measure the distances in parsecs, the proper motion in arcseconds per year and the radial velocities in km/s, you get:

$$D = \frac{v}{4.74\mu} \tag{2.6}$$

where μ is the proper motion due to the expansion of the cloud.

2.7 Pulsating stars: RR Lyrae, Cefeidi, W Virginis

The method is based on the absolute magnitude of the variables once their class, period and/or metallicity is known. Absolute magnitudes have been calibrated by nearby variables with distance measured by annual parallax (for example the RR Lyrae) or when they belong to star clusters of distances known from other methods.

The RR Lyrae are characterized by periods less than a day, and by color indexes between $B-V = 0.15$ and 0.45 with $100L_{\odot}$. Instead, Cepheids have periods from a few days to tens of days. A classic example of Cepheid is the Polar Star.

Current equations for RR Lyrae calibration are:

$$M_V = 0.16[Fe/H] + 0.98(Jones, Baade - Wessenlikmethod) \tag{2.7}$$

$$M_V = 0.15[Fe/H] + 0.80(Harris, 2010) \tag{2.8}$$

$$M_V = 0.21[Fe/H] + 0.75(HSTdata, Benedictetal., 2011) \tag{2.9}$$

RR Lyrae are the most used photometric indicators, visible up to huge distances (also extra-galactic) and easily identified (without ambiguity) thanks they short and stable period of few hours. They are evolved stars, as we know from models, located on the Horizontal Branch so they burn He into C in the core and H into He in the envelope. Those stars have a color range un-reddened and very clear,

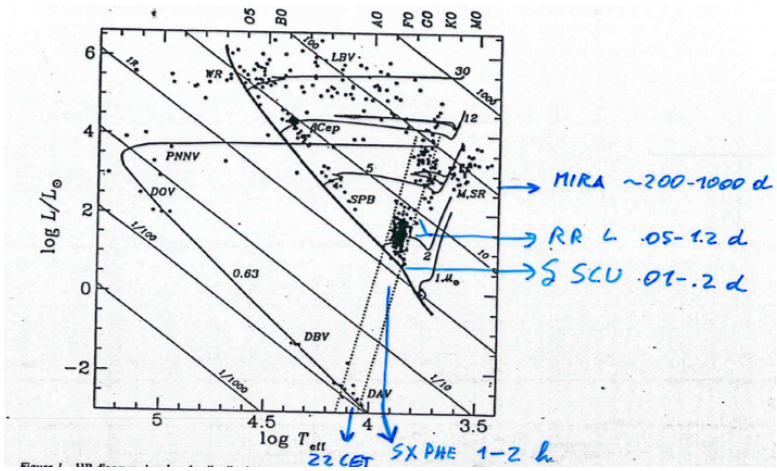


Figure 2.5: Color-magnitude diagram and instability strip.

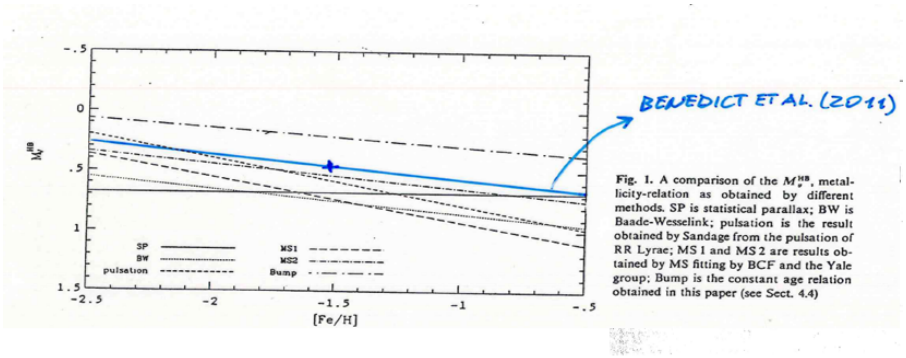


Figure 2.6: Absolute magnitude - metallicity relation for RR Lyrae.

of $0.15 < B - V < 0.45$. As all variable stars, RR Lyrae in color-magnitude diagram are located in the instability strip, a vertical section that crosses the diagram and contains:

- in the upper section, variables with long periods, about $200 - 300$ days and bigger radius, so lower density;
- in the central part of this section, RR Lyrae and more compact objects with shorter periods, about few days or few hours;
- then, in lower part of the instability strip are located all those compact objects, such as white dwarfs or pulsars extremely rapid in period and rotation.

In figure 2.5, it is visible the color-magnitude diagram and the instability strip where those variable objects are located.

But the uncertainty due to different biases (method of measuring the brightness of the RR Lyrae, range of metallicity, effects of age and evolution etc.) is wide. In particular the calibration in absolute magnitude for RR Lyrae does NOT depend on period but on **metallicity**. The metallicity coefficient ranges from 0.16, corresponding to most theoretical models, to a maximum of 0.30 (Sandage, from the period-shift method).

In particular, as visible in figure 2.6, for RR Lyrae, increasing metallicity the luminosity decreases for few percents $(0.1 - 0.2\%)$.

In particular, studies on RR Lyrae belonging to M31, so RR stars with different metallicity but same distance, have shown some peculiar dispersion of point. Maybe the relation is not a straight line as shown in figure 2.6 by Benedict et al. Indeed McNamara indicates in figure 2.7 that there is a complex trend with different slopes depending on metallicity range and the data dispersion is consequence of different range.

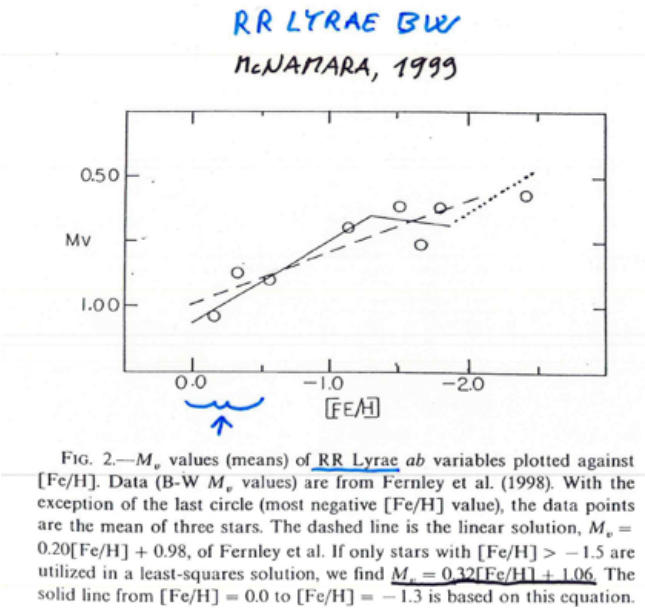


Fig. 2.—M_v values (means) of RR Lyrae *ab* variables plotted against [Fe/H]. Data (B-W M_v values) are from Fernley et al. (1998). With the exception of the last circle (most negative [Fe/H] value), the data points are the mean of three stars. The dashed line is the linear solution, $M_v = 0.20[Fe/H] + 0.98$, of Fernley et al. If only stars with [Fe/H] > -1.5 are utilized in a least-squares solution, we find $M_v = 0.32[Fe/H] + 1.06$. The solid line from [Fe/H] $= 0.0$ to [Fe/H] $= -1.3$ is based on this equation.

Figure 2.7: McNamara graphic.

This problem is still open in literature. There is still today an open apparent paradox. They measured the apparent magnitude of RR stars in Omega Centaurus, which is one of the few globular cluster with a wide range in metallicity. Since those stars belong to the globular cluster, we can assume a distance. So, we expect a straight line but we obtain a mixture of point. Is the relation not valid for this globular cluster? Nowadays we don't know. We need more spectroscopy and photometry and data.

Finally RR stars are used to measure distance of the Sun from the GC because the GC in the halo defines a sort of sphere where globular cluster with RR stars are located.

Instead the calibrations for population Cepheids I and W Virginis (or population Cepheids II) are more uncertain even if these distance indicators have the advantage of being more brilliant, and therefore observable at greater distances. Cepheids are much more bright than RR stars. They reach $100-1000\ L_\odot$ in longer periods, up to 100 days. While RR stars have constant luminosity and depend on metallicity, Cepheids are often observed in infrared range, with big excursion in luminosity.

For the classical population I Cepheids, Feast et al. (1997) give the following relation period-luminosity, calibrated through parallaxes of Hipparcos:

$$M_V = -2.81 logP - 1.43 \pm 0.1 \tag{2.10}$$

Udalski (1999) proposes for the I band the following relationship adopted for extragalactic distances:

$$M_I = -2.962(logP - 1) - 4.904 \tag{2.11}$$

which is often more convenient because the amplitude of the Cepheid variability is greater at a longer wavelength. The strong dependence on the period is evident (the RR Lyrae brightness has no appreciable link with the period). A Cepheid can reach a brightness of more than thousands of times higher than the Sun and therefore be visible up to a large distance (at $V = 25$ we get $m - M = 30$, corresponding to 10^7 parsecs, or $10 Mpc$, almost the distance of the Virgin cluster, estimated today at $18 Mpc$).

2.8 HII regions (optional)

The distance of the HII regions can be established by the size of the regions themselves which depends on the temperature of the excitation star and on the density of the medium, than the mean free path of the ionizing photons. The method can be used because the boundary of the HII regions is sharp, with an abrupt transition from the ionized hydrogen region to the outer space where neutral hydrogen dominates. The thickness of the transition region (or recombination region) is of the order of 200 astronomical units, below the resolving size of most HII regions. Known the spectral type and therefore the temperature of the ionizing star from the spectrum the size of the HII region has been tabulated by Stroemgren and can be derived from the relation that gives the limit radius where the ionization equals the recombination of the hydrogen atom. As an example, a spectral star O8V produces an ionized sphere with a radius of 140000 AU (0.63 pc) with $n_e = 1000cm^{-3}$.

The mean free path ($l = 1/\sigma n$) of ionizing photons in a neutral medium instead, with the same density ($s = 6.3 \cdot 10^{-18} cm^2$) gives 10 AU, equal to about 10^-5 of the Stroemgren radius, therefore the transition from the ionized medium to the neutral one is very sharp and the limit of the Stroemgren sphere becomes very clear. The method is limited by the uncertainty of the density of the interstellar medium, and it is in any case restricted to the HII regions for which it is possible to identify the spectral type of the ionizing star. The problem is not of secondary importance because most HII regions can be identified in radio but optical absorption is too high to obtain optical spectra.

2.9 Dynamic parallaxes

The method is based on the knowledge of the semi-axis of the orbit of the double visual stars. The method is apparently limited because it can only be used with resolved and well known double stars, but in reality it is of great interest because it is geometric, and it can constitute an independent method for measuring some stellar populations (for example, a measurement not affected by reddening or theoretical models).

The equation that is derived from the observations of double visuals, based on the third law of Kepler is:

$$a^3 = P^2 \cdot const. \cdot (M_1 + M_2) \tag{2.12}$$

where M_1 and M_2 are star masses, a is the semi-axis of the orbit movement, P is the period (obtainable from observations) and $const$ is a constant containing the gravitational constant G.

Of course, a is knows in angular scale while the two masses are assumed on spectral type and models (generally similar to solar mass). So, measuring angular distance simultaneously from comparison angle and linear scale, the triangle is fully determined so we get the distance.

With this method the distance of the Pleiades (used as standard reference of population I) from the double star Atlas was measured recently and it was found to be $d = 135pc \pm 2$, against $118pc \pm 4$ which instead results from Hipparcos. The reason for this discrepancy is not yet clear.

2.10 Novae and supernovae

The novae reach the maximum:

$$M_V = -9.96 - 2.31log(decl) \tag{2.13}$$

where $decl$ is the decline time of 2 magnitudes. Type Ia supernovae have $M_V = -19.3 \pm 0.1$.

They are candles observable from a great distance, used also to establish the expansion of the universe. The accuracy of the determinations depends on the calibration and on the distinction between subtypes

and the recognition of specific peculiarities. There is also dependence on interstellar absorption which is not always easily deducible from the same observations of novae and supernovae. Generally interstellar absorption is local, nearby the supernovae or between our galaxy and the supernovae or within out galaxy. Usually the intergalactic reddening is low.

Distances of supernovae can be determined using a photometric relation, 2.13, but also in a geometric way.

Indeed supernova is a huge explosion with ejection of external layers. From ground we measure the radial velocity, which means the ejection velocity of stellar matter that is very high, up to $10000 km/s$. At the same time it is possible to measure the angular expansion of the cloud in radio range (because we measure synchrotron effect due to the compression of medium after some days).

Therefore, the absolute distance is the ratio between absolute size and angular size:

$$d = \frac{v \cdot t}{\omega} \tag{2.14}$$

where d is the absolute distance, v is the ejection velocity and t is time of expansion.

This method can be used up to a distance of $10 Mpc$.

Chapter 3

Young stellar population

3.1 Colour-Magnitude Diagram

The colour-magnitude diagram is a useful item to study stellar populations. We have two axes.

- In X-axis: we can use, from the observational point of view, the colour index (ex. $B-V$), from the theoretical point of view, the temperature. Recall that T and colour index are related by a relation: $B-V=-0.865+\frac{8540}{T}$. In general $c=A+\frac{B}{T}$ where A and B are constant. It can be used also the spectral type of stars. The temperature, we talk about, is the effective temperature (of thousands degrees), it is the temperature of the photosphere and not that of the core (of some millions degrees).
- In Y axis = from the observational point of view we use the absolute magnitude in visual band M_V, or other bands. In theoretical model we use the luminosity.[1]

Most of the stars, near Sun, are collocate in one region of this diagram, called main sequence (MS). It is a relatively narrow region and we know from stellar evolution, stars in this region burn H into He in the inner parts. Since this nuclear reaction is very energetic, the result is that these stars can remain in MS in the same position, with stable temperature and stable luminosity for very long time. This is the reason why the majority of stars are located in MS. In the low part of the MS track, there are partially convective stars (with lower mass, fainter and cooler) while at the top of MS they are almost totally radiative stars (high mass, bright and hotter). Along MS stars are located according to the mass and they are characterized by a uniform chemical composition. Anyway the mass range of stars is quite small while the sensitivity of the position along MS from the star mass is very high. However, these stars go from $0.5M_\odot$ up to $5/6M_\odot$. In particular, going down to lower masses, they are more numerous but we don't have this perception because they are also fainter.

Above this, there is the zone of the giant and super giant stars, more and more times brighter than Sun ($L=100L_\odot$) but redder.

Between the giant zone and the MS there is the Hertzsprung gap, due a rapid evolution of the stars that exit from MS: they already burned H into He in the core and proceed rapidly in burning shells. Moreover we can not observe many stars due to a selection effect: we lose the faintest ones.

There is also a small group of stars at left and below of MS that contains white dwarfs, small and evolved stars, much dense and compact.

In CMD, stars occupy only the allow region, a sort of triangle between the MS and a vertical line called Hayashi track.

To the left side of MS is not allowed by theoretical models. In fact, this correspond to a star that

[1] how do we pass from L to M? From luminosity we obtain the bolometric magnitude thanks to the Pogson's law and with the bolometric correction (which distorts the diagram) we obtain the absolute magnitude in one band.

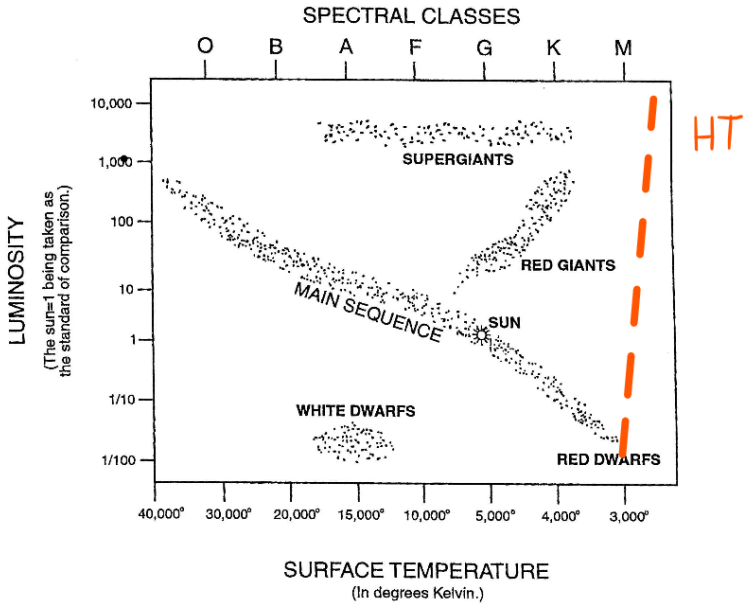


Figure 3.1

evolve completely mixed, instead star evolve by separated shells! It means that during evolution nuclear burning proceed in separated shells. During the MS stars became brighter and larger in radius and decrease surface temperature, so they move towards the giant zone, to the right. There are different reasons for MS stars to go up to giant region.

- One, not the main, is over-production of energy. When the star starts to burn H in shell, there is a He core which is unstable and contracts because there is no production of energy inside, producing thermal energy by gravitational contraction. This creates a surplus of energy and this make the star expanding.
- The most important driver to giant region is the average increase of molecular weight that to transformation of H into He. Therefore, in forces equilibrium, radius increases. Then the evolution of giant stars depends on mass.

The line that cannot be overpassed in the right side of the diagram is the **Hayashi track**, that correspond to the track of full convection. Indeed, moving to higher color index, which means lower surface temperature, the surface opacity increases in a strong function of temperature. So if temperature decreases, opacity increases and vice versa. When the layers are getting opaque then the energy transfer can not be anymore radiative and becomes convective. So, as T decreases, convection stars to be the first mechanism to transport energy in external layers and the star gets more cooler. At the same time convection goes deeper in the star until it is fully convective along the Hayashi track. Stars cannot be more than fully convective so there are no stable configuration on the right side of Hayashi track.

The main sequence models are known precisely and verified on the Sun and in a large sample of stars located near the Sun.

PAY ATTENTION: Do not confuse stellar evolution in CMD with stellar evolution on central temperature-central density diagram. In CMD there is the effective temperature, of the order of thousands degrees on surface, so observable on the photosphere while in the other graphic there is the central temperature (core T) which is about of million degrees. This last one is due to thermal nuclear reactions that occurs in a very limited area in the inner part of the star because there is a strong gradient in density inside the star.

Sun age A key issue is to know age stars. In particular it is very important to know with great accuracy the Sun age because the Sun is a reference for stellar evolution models and many of them are calibrated on it (not the opposite).

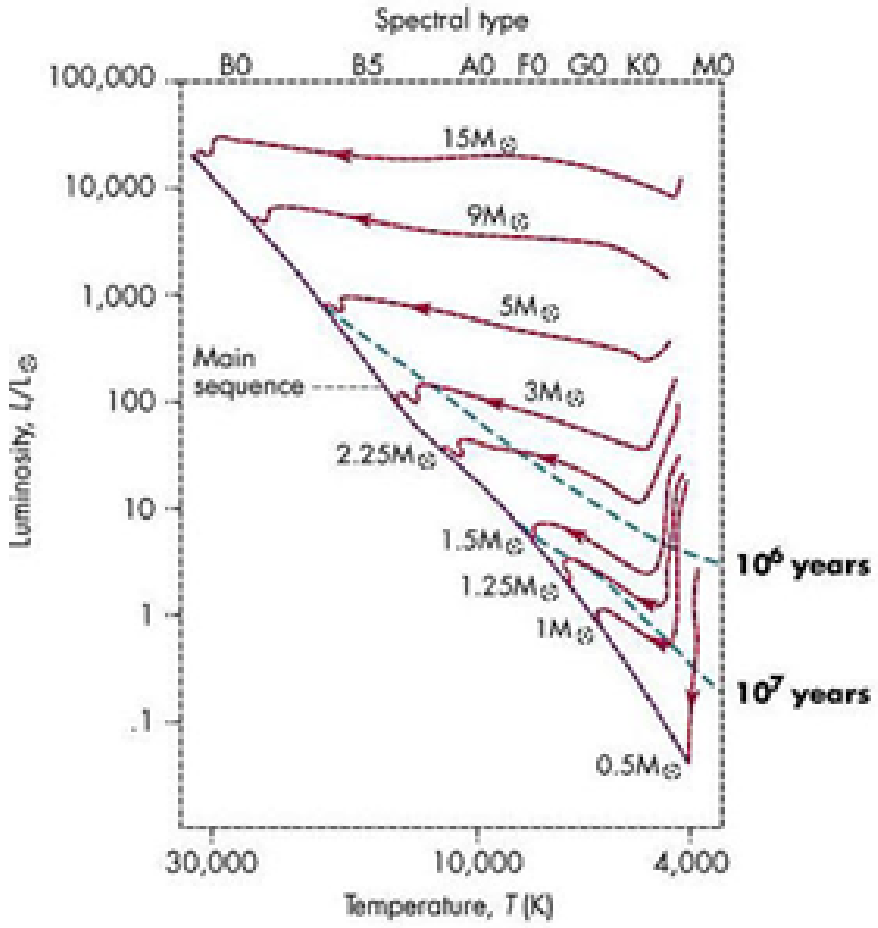


Figure 3.2: CMD from theoretical point of view, with temperature on X-axis and luminosity on Y-axis.

To calculate age of Sun, we use the radioactive isotopes, in particular, assuming that the Solar System formed together with Sun, it has been derived from the isotopic ratios present in the rich inclusions of calcium and aluminum of the most ancient meteorites, of the chondritic type, collected on the ground. Amelin and collaborators (2002), estimate 4567.2 ± 0.6 million years, a very high accuracy. This computing is possible because the radioactive decay is independent of environment.

CMD of young population

The young population correspond to stars that are not arrived to MS, as visible in figure 3.2. The blue dotted line is the isochrone curve that represent a simple stellar population with same age, so constant in time. Therefore isochrones are not evolutionary paths and they are useful in CMD as reference of observational plot, eventually more tilted due to bolometric correction. The red solid line represent the evolution track of stars with different masses.

From the phenomenological point of view there are limits to the study pre-sequence evolution on HR diagrams (or analogous color-magnitude diagrams). These are due to at least two factors:

- the position on the HR diagrams is influenced by the presence of excess infrared color due to the presence of discs, or excess of ultraviolet radiation due to processes of growth or in any case to alterations of the spectrum due to emission lines;
- the regions of recent star formation are rich in dust which cause strong reddening with consequent difficulties in deriving the intrinsic color temperatures.

Moreover, from the point of view of the models, in the main sequence we have the **Vogt-Russell's theorem** that says that the position of a star on the main sequence does not depend on its pre-

sequence history but only on its mass (with a fixed age and chemical composition), like saying that for a given mass there is only one equilibrium position for stars that burn hydrogen in the core. So, the treatment of pre-sequence models is instead more complex.

The color magnitude diagrams of young populations are normally based on the near-infrared (JHK) bands to minimize the problem of interstellar absorption. They are characterized by a main sequence where the most massive stars are found, while going down to smaller masses, from a certain point, the sequence appears truncated because the stars are still in the pre-sequence phase, on the right of the diagram.

The evolution towards the MS

Stars form in a big cloud, eventually thought fragmentation, at very low density and low temperature by gravitational contraction. Although temperature is low, luminosity emitted by the cloud is relatively high because it has a very large radius so the star starts from the upper part of the diagram. Then it reaches the upper limit of Hayashi line; now the protostar is formed and it go down through the Hayashi track. Going down, the star contracts, reducing radius and reducing the emitting surface. There is not many observations of this phase because:

- This is a fast evolutionary phase, fast means less than 1 million years.
- In this phase there is too many dusts, so stars in this phase are not observable in visible band, eventually in infrared or radio range.

Evolution proceeds in fully convective phase contracting and heating up. In this phase the energy comes from the gravitational contraction and not from nuclear reactions. According to the Virial theorem, half of the energy is radiated while the other half goes to increase the internal energy and therefore the internal temperature.

The temperature increase, of the inner regions, decreases the opacity, and therefore the temperature gradient below the convective shell and the inner part of the star becomes radiative with increasing radius over time. The evolution slows down, and the star increases the surface temperature with a small increase in luminosity, which follows the decrease in the opacity of the entire star. So, star left the Hayashi track and reaches the main sequence and triggers the thermonuclear reactions of hydrogen burning. It is when the star leave the Hayashi line, that the gas surrounding the star is cool enough to condensate into grains and it becomes a flat rotating disk which is the protoplanetary disk. It is in this phase that planets form in a timescale of few million years, usually only one.

Time-scale of evolution

For a solar-type the time scale of the phase of pre-sequence is around 10^7 years. In particular the time scale of the evolution along the Hayashi line is around 10^5 years. It should be noted that the evolution time along the Hayashi line, during the entirely convective phase, is very short, about 2 orders of magnitude lower and therefore difficult to observe. The stars that are generally observed in the young clusters are already found in the phase of rising of luminosity that precedes the entry in the main sequence.

3.2 The ages of young population

The age can be estimated from the position and shape of the pre-main sequence turnover. The diagrams often appear to be dispersed due to high field stars contamination, reddening and differential absorption and the presence of stars with infrared excess due to protoplanetary disks.

- Contamination: different techniques allow to clean up the diagrams from the contamination, for example with the proper motions or with the typical X emissions of young stars.

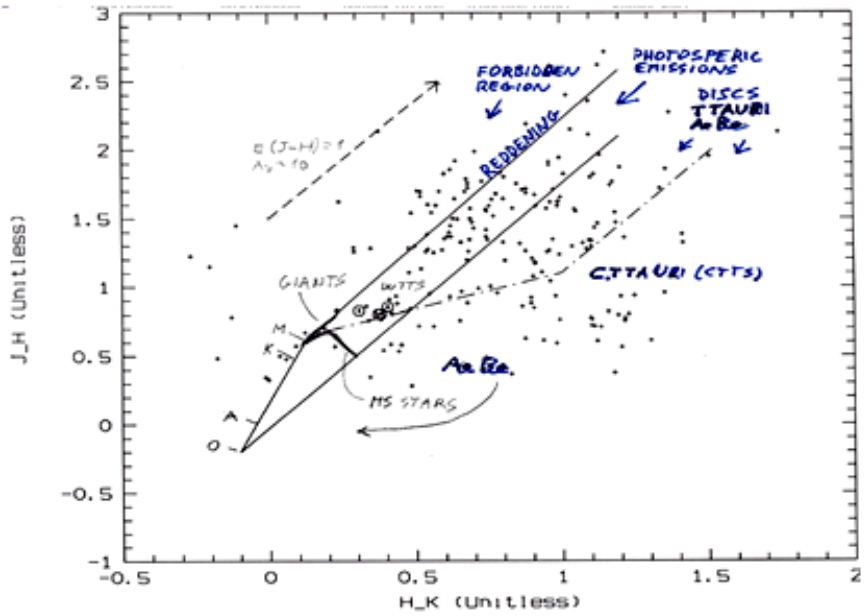


Figure 3.3: Standard JHK diagram.

- The reddening and the differential absorption instead can be corrected by obtaining for each star the reddening in the two-color diagram $J - H$-$H - K$, with respect to the position of the unreddened points, and then correcting each point on the CMD diagram.
- The stars with infrared excess are identified in the two-color diagrams with the technique described below and treated separately.

Two colour diagram

For the study of young populations, in order to minimize the problem of interstellar absorption and to over-plot the isochrones to get the age, it is fundamental the preliminary study in the two-color infrared diagram (generally $J - H$ vs. $H - K$).

In this diagram the place of the points occupied by the stellar photosphere models, both of MS stars and giant stars, is well defined and limited between about $J - H$ and $H - K$ about 0.0 for hot stars, type OB, up to about $J - H = 0.8$ and $H - K = 0.4$ where the first extreme is reached by the giants, the second by the dwarfs. In particular, we can see the track from O star to M star, at this point the track split: the giant one goes upper and the dwarf one goes bottom. The black solid straight lines correspond to the reddening limit for all the stars so all stars are shifted up by reddening effect within these two parallel lines having the slope of the reddening vector. As visible in the figure, the reddening vector corresponds to $E(J - H) = 1$ that, from tabulated values, is equal to an absorption A about 10. Moreover points are moved up not by the same reddening because it is a differential one. The reddening lines in this diagram can be easily calculated, we have, from literature, that:

$$\frac{E(J - H)}{E(H - K)} = 1.94 \tag{3.1}$$

Ratios between different color excess ate tabulated in literature but pay attention: they vast majority of them have been calculated for very hoot stars like O or B.

Above the upper black line there is the **forbidden area**; we would have any star above the reddening line because it is the limit to color index of M stars but however there are some points. They are all measurements errors; the stars here are affected by a larger error because they are faint stars at

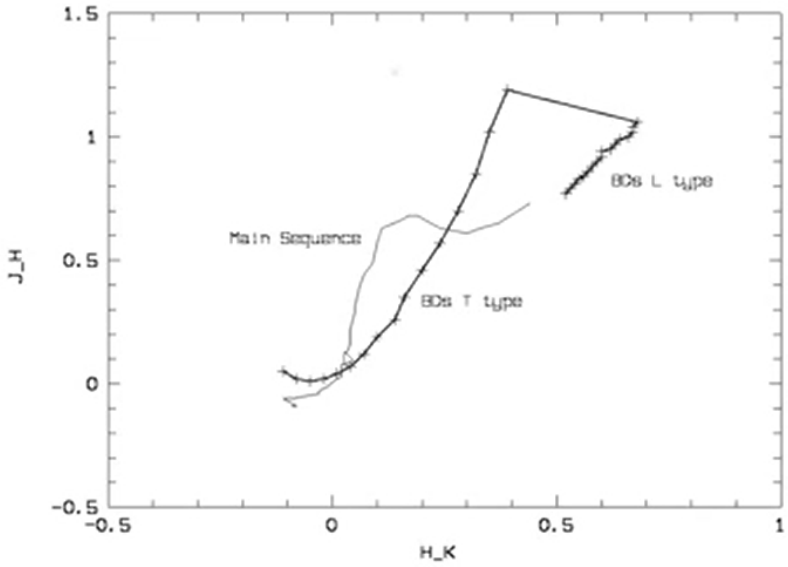


Figure 3.4: JHK diagram for brown dwarfs.

the limit of photometry. Between two black lines there is the **permitted area** where the majority of point is located. Under the bottom black line there is the **disk region**: here the stars present a protoplanetary disk. This region is populated by T.Tauri stars (a class of irregular variable stars that are less than about 10 million years old), Ae and Be stars (having less than 10 million years and coming from A and B types). All these points are real stars (not measurement errors) located below the permitted area due to the present of the disk with different slope or extension. It causes a $H - K$ excess so it moves the points to right. In additional there is the reddening effect that shifts up the points. The result is a large distribution of point below the parallel lines that delimit the permitted area.

From the two-color diagram it is possible to individually correct the stars for reddening and in some cases also for the infrared excess due to the disk. The photometry thus obtained can then be used in the infrared magnitude color diagram to obtain age and distances from the fit with the pre-sequence isochrones. The accuracy of the age from the fit with the isochrones in the pre-sequence phase is limited by the accuracy of the isochrones themselves.

One exception Of course, there are some exceptions. For example in figure 3.4, it is reported diagram for a peculiar class of stars, brown dwarfs (L and T types), very peculiar objects characterized by a very low temperature and small mass. The gray line below represents standard black body emission in MS while the black one represent the plot of brown dwarfs. It is completely out of the standard one because at very low temperature there are a lot of lines in near infrared so it is not anymore a BB emission.

JHK diagram of clusters Figure 3.5 is a graphic created on observational data of a cluster, G287, very reach in points. The circled crosses represents the most probable membership of the cluster. We see that the points are clustered along the two dotted lines which are the isochrones for age of 1 and of 2 million years (green circle). There also some crosses under the MS track that should not be under the isochrones. Indeed under them there should be nothing however we see some points (orange circle). These crosses are contaminating stars, not from the cluster. They are disk stars, nearby the Sun and seen in projection in front of the cluster.

The point of the plot are genuine data, not corrected for reddening, not for infrared excess indeed all stars inside red circle, which are reach in infrared excess due to the present of the disk around, are out of the MS going to the right side of the diagram. Indeed 1 or 2 million years are very young ages so there are still disks.

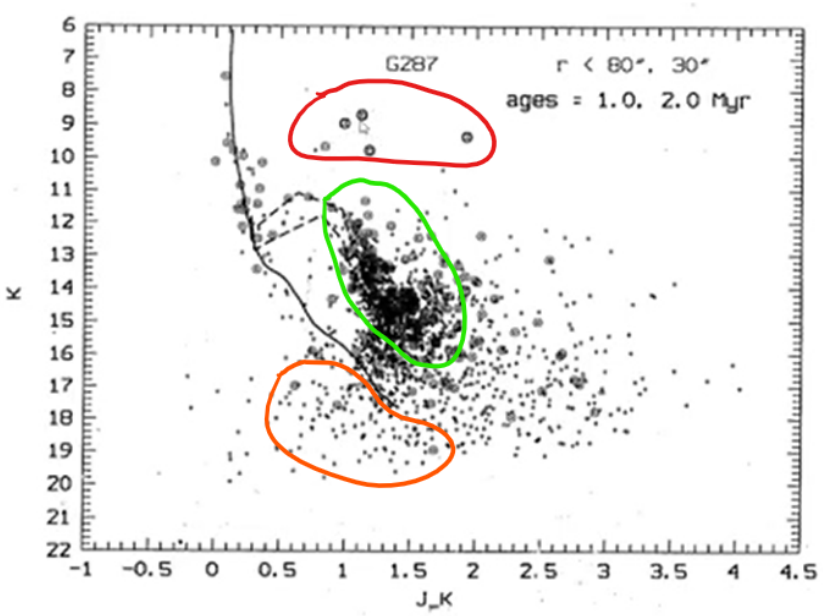


Figure 3.5: K - $J-K$ diagram for cluster.

Other methods

For young stars there are four other different age estimation methods (Mamajek, 2007):

- **Lithium content** - The decrease of photospheric lithium over time is considered one of the most reliable methods of dating the pre-sequence stars. Observations are made on the Li line at $6707\mathring{A}$. The method is based on the rapid burning of lithium at temperatures lower than those of the PP hydrogen cycle (2.5 vs. 10 million degrees) which can then take place in pre-sequence stars. More precisely the ignition temperature, for solar-type stars, occurs at the base of the the Hayashi line, where the star is still entirely convective and therefore the mixing between the photospheric material of the star can occur, at the age of just over a million years. The reaction takes place directly by capturing a proton or an alpha particle and can lead to total exhaustion of lithium in a relatively short time (about a hundred million years).

  The exhaustion of lithium obviously depends on the temperature of the star. For very low temperatures (so small masses) the reaction is obviously slower. But, at high masses and temperatures, the decrease in lithium can be limited by the rapid development of the radiative core that pushes the base of external convection towards greater distances, bringing it to a region where the temperature is too low for the reaction of lithium. This explains why the models expect greater abundance of lithium at the extremes of very low and very high temperatures. In any case we observe a rapidly decreasing of Li in stars according to their ages so this method is one of the most used to calculate the age. The calibration was done in the age range between 1 and 30 million years, using atmosphere and isochrone models. This method, strong for same temperature, is often used for the study of age of individual stars.

  In particular our Sun have a lower Li abundance than other Sun-like stars (less than 1 order of magnitude) and we don't know exactly why. Moreover the "sensitivity" to age bigger than the solar one is less in the sense that Li abundance may be the same for wide range of ages. Instead for age lower than the solar one, the connection Li abundance-age is great.

  Figure 3.6 shows Li abundance as function of age (in Gyr). For stars in MS phase, there is a well defined mathematical decrease for stars from 1 to 10 Giga-years.

  This method is easily used with Li because it presents very well defined and strong lines at the center of efficient range in wavelength of our sensors. We can use also deuterium, more difficult so observe due to H lines nearby, or also Beryllium, with a line less strong.

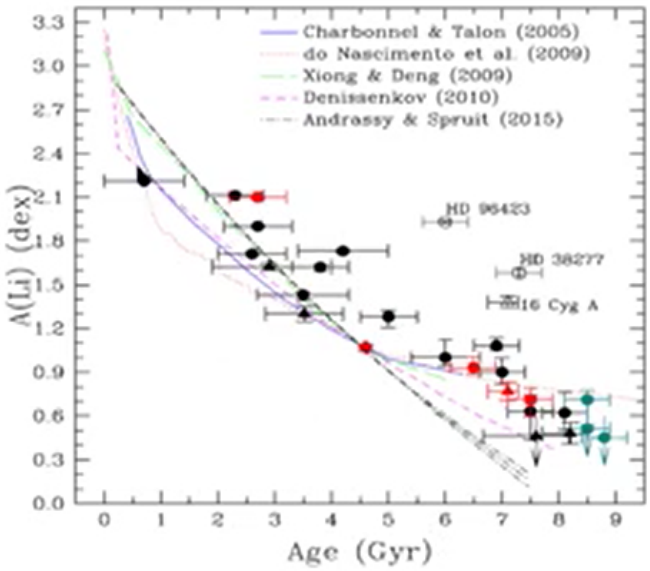


Fig. 5. Connection between stellar ages and NLTE lithium abundances for our current sample (circles) and some previous results (triangles) referenced in the text. Teal blue circles indicate alpha-enhanced stars and red symbols refer to stars hosting planets. The models of Li depletion were normalized to the solar Li abundance. In some cases, the lithium abundance errors are smaller than the points.

Figure 3.6: Li abundance against age.

A peculiar aspect connected to this method must be studied more: what is the influence of rotational speed of the star? It prevents the full development of convection because the Coriolis forces distorts the convection and decrease the efficiency of the transport of material from center to surface. However, nowadays we don't have much information about it.

- **Disk fraction against age** - Figure 3.7, shows, for different clusters at different ages, the fraction of them having disk (in per cent) against the age (in million years) so every point is not a star but a cluster of stars at a given age. In particular, the identification of stars with disks is obtained by studying the two-color infrared diagram, $J - H$ and $H - K$, as explained before. More or less, the 50% have a disk and 50% don't have it.

  Turning to the plot, we plot the diagram for different ages, known from isochrones, *Li* or other methods, against the fraction of clusters with stars having a disk. This method can be applied only for groups of homogeneous star in term of age.

  Conclusion: it is found that stars with ages less than about one million years are almost always surrounded by a disk, instead stars with more years are not surrounded by disk indeed the fraction of stars with a disk decreases with the age. In particular at age of about 1 million years, the vast majority of the stars, up to 90% are surrounded by proto-planetary disks. At the age of 2 million years, the fraction of stars with a disk is more or less 50% and finally to an age of 10 million years, the fraction goes below 10%. At 100 million years, the fraction is about 0% which means that the proto-planetary disk disappears and there is a pure photometric emission. This means that proto-planetary disks disappear in a time scale of about a couple of million years so at the base of the Hayashi track.

  The reasons why there is this gradual reduction of stars with a disk, going from the youngest to the oldest are the following:

  – planets, that form in 10 million years which means in the time scale of this plot, have collected all the material from the disk;

  – disks are removed partially by stellar winds.

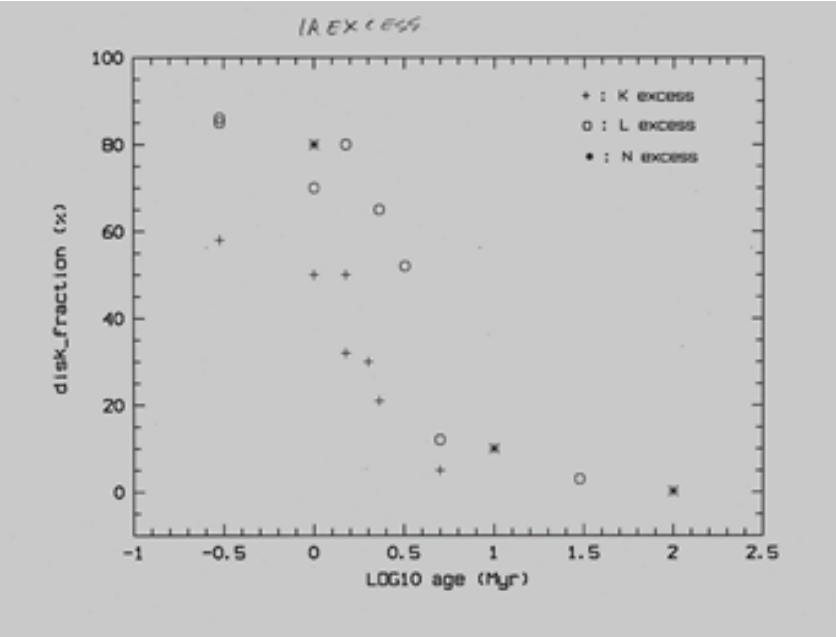


Figure 3.7

In figure 3.7, we can observe also that more or less, the points and the time scale are consistent (the dispersion is about half a million year). This means that since this is in order of increasing λ, their temperature decreases so this means that disks simultaneously disappear at all temperatures. Therefore the external part of the disk, which emits more in L or N bands than in K, disappears at the same time scale so planet forming time scale should be quite homogeneous on all over the proto-planetary disks.

- **Asteroseismology** - Stellar oscillations are an effective means of deriving the ages of isolated stars. The fundamental point is that the oscillations give direct information on the stellar structure. What is called the small frequency separation represents the observable speed of sound where thermonuclear reactions are generated. The limit of the method is given both by the availability, for now limited, of data, and by the models involved for the interpretation of the observations.

- **Gyrochronology** - The rotation speed of young stars decreases with time (Skumanich, 1972). It is known that for T Tauri stars the rotation period is around 10 days, while it is known that the Sun rotates with a period of about 25 days. There is therefore a braking mechanism not yet well identified with correlates the decreasing in rotational speed with the age.

  The method is applicable to cold stars, of the F-M type, with convective envelope and age of over 100 million years. In these limits the period of rotation can be expressed as: $P = f(B-V) \cdot g(t)$, where f and g are two functions dependent on the color index and age. The calibration is adapted in order to reproduce the solar parameters.

  Unfortunately, this observable relation between rotational speed and age has a wide dispersion due to the different nature of stars. There are the **rapid rotators** (upper line in figure 3.8), in hydrodynamic collapse, and the **slow rotators** (lower line in figure 3.8).

Calibration of isochrone age Figure 3.9 is a plot of ages in scale of Gigayears. In particular HK are lines of $CaII$ with emission in the extreme blue of optical spectrum. They represents emission core indicating stellar activity which decreases with age. So this plot it the calibration of isochrone age against the activity age. This plot is somehow good; more or less the trend is monotonic.

Recap Talking about age measurement, there are different methods, as we seen before:

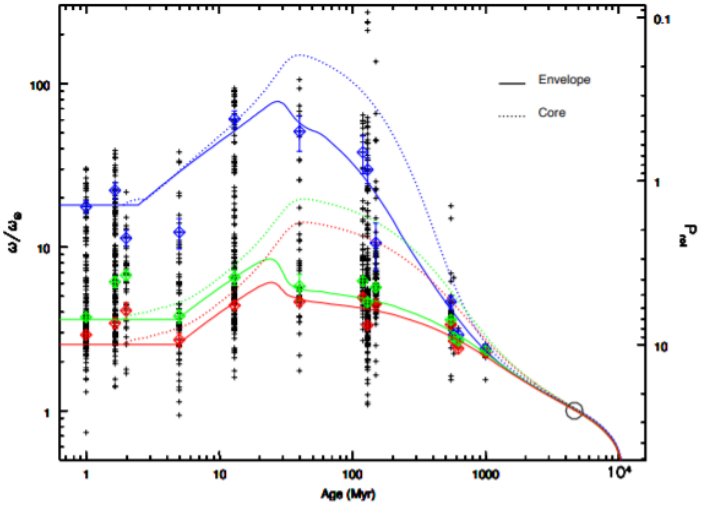


Fig. 7.— Observed rotation rates of stars near one solar mass in young clusters (*Gallet and Bouvier*, 2013). (See also the chapter in this volume by Bouvier et al.) The colored lines show several models that are not of relevance here. The plus symbols show observed rotation periods. The blue, red, and green diamonds represent the 90th percentile, 25th percentile, and the median for each cluster's distribution of $P_{\rm rot}$. Note the very large spread (1.5 to 2.5 dex at any one age) and the lack of a clear trend for at least the first $\sim$ 100 Myr.

Figure 3.8

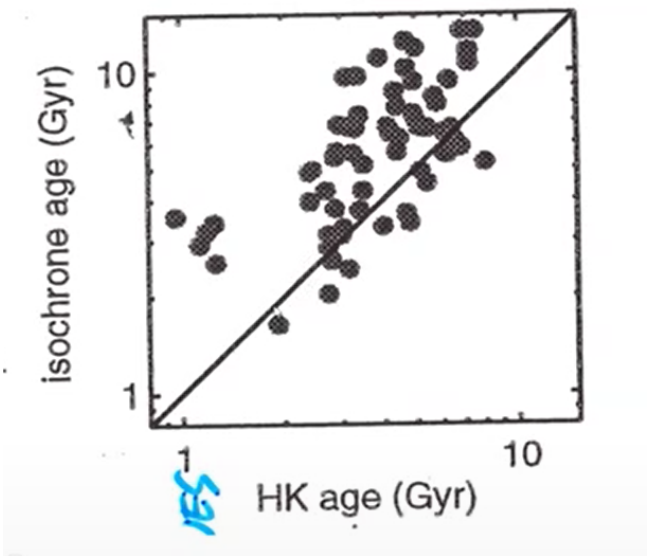


Figure 3.9

- methods only for single stars, measured one by one. In this case we use:
  - *Li* abundance;
  - rotational speed of the star and activity.
- methods for group of stars:
  - isochrones;
  - fraction of stars with a proto-planetary disk.

Chapter 4

Physics of planets

4.1 The atmosphere

We start this chapter showing the composition of our atmosphere.

- Nitrogen (N) $\sim 78\%$ in molecular form, which is very stable with high bonding energy;
- oxygen (O) $\sim 21\%$ in molecular form, very reactive;
- argon (Ar) $\sim 0.93\%$ in single atoms because it is a noble gas, less stable than N;
- carbon dioxide (CO_2) $\sim 0.04\%$;
- traces of other elements;
- water vapor (H_2O), very variable depending on whether conditions.

In general oxygen is not very abundant in nature in free form because it is very reactive, as said before, for example with carbon C. Inside the stars occurs the CNO cycle so reactions between oxygen and carbon are very common, producing first carbon monoxide (CO) and second carbon dioxide (CO_2).

Where does argon come from? It is the production of decay, in 1.6 $Gyrs$, of the ^{40}K, potassium isotope: this process products of the ^{40}Ar. On Earth ^{40}Ar is about 99.6% vs. 0.34% of ^{36}Ar, while about 84.6% of ^{36}Ar is in the Sun. This tells us that on Earth original argon was completely lost.

Except for the water vapor, the Earth atmosphere is composed by **perfect gases** and it is a unique chemical composition among all planets in Solar System.

Other planets have different composition: Mars and Venus have most of the carbon dioxide and monoxide, and only traces of argon and nitrogen, and lower quantities of water. So it is very weird

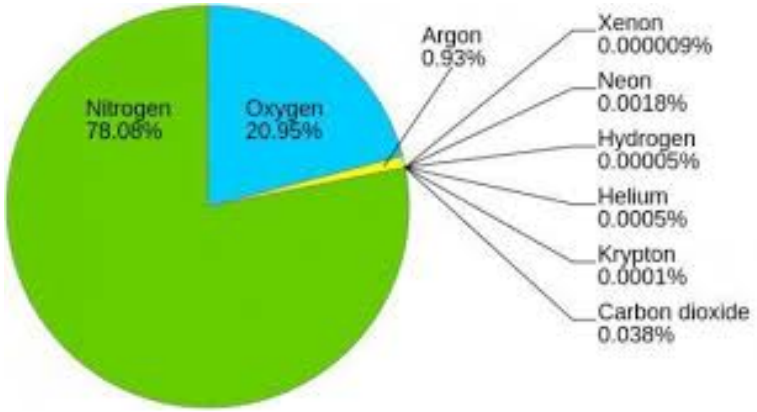


Figure 4.1

seeing these quantities of nitrogen and oxygen in our atmosphere and a low quantity of carbon dioxide. To explain this big difference in chemical composition, in the past it was proposed that on Earth carbon dioxide is low, so there had been in the past a loss of this molecule, due to the photosynthesis of chlorophyll by plants and it is fixed, storing it on biomass. However it is partially wrong, as we will see.

Now, Mars and Venus are dead planets (on these planets reactions occurs slowly, due to the absence of a liquid chemical compound), so their atmosphere can be assumed as a primordial (initial) atmosphere.

How to explain the nitrogen quantities on Earth? N is not an obvious element because usually it is under abundant. Mars is not good to comparison because it lost most of its atmosphere. Venus, instead, has a very rich atmosphere that cause at ground 90 times the atmospheric pressure of Earth. So N is a few percent of total, but it is comparable to absolute quantity of our atmosphere. Despite all these observational data, nowadays we don't know why on Earth there is a so high abundance of nitrogen. Maybe it was fixed in some meteorites but there are no solid evidences of it.

Where the oxygen came from? There are a few reactions to obtain oxygen; one of this is the photoionization of the water, that is a common molecule, indeed in nature H is the most abundant atom, also the O is much abundant. This is confirmed on Venus by some recent resources. On Earth there is also the action of plants, but the most producer of oxygen is the phytoplankton in the oceans.

It is interesting to know that in the last 3/4 Gigayears the nitrogen abundance has been mainly constant, due to the fact that reactions with this element on ground are very slow. Instead oxygen has been varied. For example 250 million years ago oxygen went from 35% to 17%, maybe due to volcanic eruptions, testing life survival (96% of life disappeared). Then the oxygen stabilized at 21%. We can know this studying evidences on rocks which show that atmospheric pressure was almost the same now. So, with a good accuracy, we can deduce that pressure has been almost constant until today.

4.2 Basic equations to study the atmosphere

As said before, gasses in Earth atmosphere can be treated as perfect gasses so they are described by the equation of perfect gas:

$$PV = nRT \tag{4.1}$$

Where P is the pressure, V the volume of gas, n the mole number, R the constant of gas and T the absolute temperature.

An equivalent expression is:

$$PV = NKT \tag{4.2}$$

where N is the molecular number and K is Boltzmann constant.

Other fundamental equation is the following, to derive kinematic energy of gas particles.

$$\frac{1}{2}mv^2 = \frac{3}{2}KT \tag{4.3}$$

where m is the mass particle (depending on gas type), v is the velocity of particles and T is the temperature, as seen before. This equation means that kinematic energy is equal to the energy of motion of particles due to temperature. So, from the above equation, we derive:

$$v = \sqrt{\frac{3KT}{2m}} \propto \sqrt{\frac{T}{m}} \tag{4.4}$$

Therefore we can derive the velocity due to thermal motion, that depends on type of atom or molecules we considered. For example, at $T = 18°C$ the velocity for H_2 is $2km/s$ while for O_2 $v = 0.5km/s$.

Relating the kinetic energy with the gravitation energy we can compute the escape velocity:

$$\frac{1}{2}mv^2 = \frac{GmM}{r^2} \tag{4.5}$$

where M is the mass of the planet and r is the distance between the centre of the planet and the position of the particle with mass m. In this way we obtain the escape velocity, equal to:

$$v_{escape} = \sqrt{\frac{2GM}{r^2}} \tag{4.6}$$

Therefore, if we want to keep bound the atmosphere on a planet the escape velocity must be larger than the thermal velocity. If the escape velocity is lower than gas velocity, particles are lost in space and with them an atmospheric fraction. The process of loss of the atmosphere is very important, because the presence of atmosphere is fundamental for life. E.g the cycle of temperature on a planet depends on atmosphere. For example on Earth the temperature excursion between night and day is quite small (about 10 or 20 degrees) but on the Moon, where there is a very thin atmosphere, the temperature excursion is very big (from 127 to -240 degrees).

4.3 The processes of atmospheric loss

Thermal processes

The loss of atmosphere due to thermal motions (**Jeans mechanism**) becomes significant when the thermal velocity of the molecules is comparable with the escape velocity.

Useful numbers: the mass of the Earth is $5.9 \cdot 10^27g$. Mass of the atmosphere $5 \cdot 10^18g$, total mass of the hydrosphere $1.3 \cdot 10^24g$, v_{esc} of Earth $11.2km/s$, $4km/s$ of Mercury, $5km/s$ of Mars.

On Earth $v_{esc} = 11km/s$, and for H_2 $v_{gas} = 2km/s$, so the thermal velocity hydrogen molecules is about 20% of the Earth escape velocity at room temperature. Therefore hydrogen, which is the most abundant element on Earth, should be in the atmosphere but it doesn't.

Question: why is not the hydrogen present in atmosphere? In the higher atmosphere the temperature increases to about hundred degrees, but this is not enough to overcome the gravity, we could have temperature of the order of thousand degrees to escape. To solve this problem, we have to consider that the particles follow the Maxwellian distribution of velocity (similar to a Gaussian but asymmetric). So, the Maxwellian distribution has a long tail towards the high velocities and this means that some particles reach and overcome the escape velocity and escape from the Earth. This explanation however involves only a small fraction of the all particle of H_2! So, with this process, atmosphere loses a fraction of the H_2, then, with time, the distribution is re-normalized so another small fraction will lose. This process happens many times until a larger fraction is lost. To obtain the same quantity that we measured about 4 Gyrs are needed.

So in general, if the thermal speed of a gas is about 20% of the escape velocity, this kind of gas is almost completely lost in 4 Gyrs. Of course, at the end it will be still a remainders, eventually taken off by other processes.

People think that this is the dominant remove process in planetary atmospheres but this is not true. Indeed there are other removal effects, explained in the following subsections.

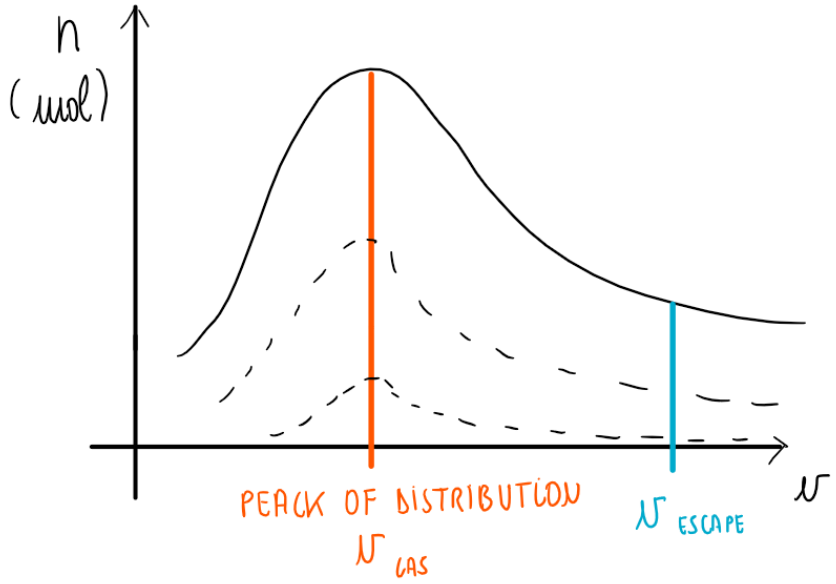


Figure 4.2: Maxwellian distribution.

Non-thermal processes: chemical-physical capture ("sequestration")

The capture of gas by the crust is important. It consists in chemical fixation of atmospheric gasses and on Earth is the main effective mechanism, still working. Also it has a relevant impact on climate and its changes.

The capture mechanism is based on the Urey reaction which can be simplified in the following form:

$$CO_2 + CaSiO_3 \rightarrow CaCO_3 + SiO_2 \tag{4.7}$$

where $CaSiO_3$ is calcium silicate, $CaCO_3$ is calcium carbonate and SiO_2 is quartz. This reaction is very slow combining a gas (CO_2) with rocks ($CaSiO_3$) and inefficient so liquid water is a fundamental catalyst for the capture of carbon dioxide and the chemical attack of calcium silicates. This happen thank to rainfall, that goes through atmosphere capturing CO_2 and brings it on the ground, in particular on young exposed rocks (common, for example, on mountain chains or regions with volcanic activities). The carbonates thus formed, taken deep in the Earth by the motions of the crust, under conditions of high pressures and temperatures, can again release carbon dioxide in the form of volcanic eruptions, so the reaction can be done back to front. By the way, there are also some captures of water vapor though other complex combinations, indeed water is stored also inside the crust and erupted by volcanos.

This mechanism is critical for maintaining the thermal balance on the Earth's surface. This is the chemical cycle on Earth; a delicate equilibrium. The time scale of the reaction is very uncertain, recent estimates give between 100000 years and some millions of years. Factors related to uncertainty are several: we don't know which role have plants that attack rocks, the regions where rainfall is more effective are equatorial regions where there aren't many exposed rocks.

It is believed that on Earth the carbon fixed on the rocks constitutes about 250 times that of the atmosphere. Given that, the current carbon dioxide content in the atmosphere is of the order of 380 parts per million; the release of all carbon in the form of carbon dioxide could make the atmospheric pressure comparable to that of Venus, where this process does not happen. This gives an idea of the importance of the mechanism of chemical capture of atmospheric gases by rocks.

In conclusion, the balance of this process is not exactly 0; there is also a fraction not irrelevant of CO_2 inside rocks, in deep in the crust. Also there is a loss of mass though atmosphere, despite this process.

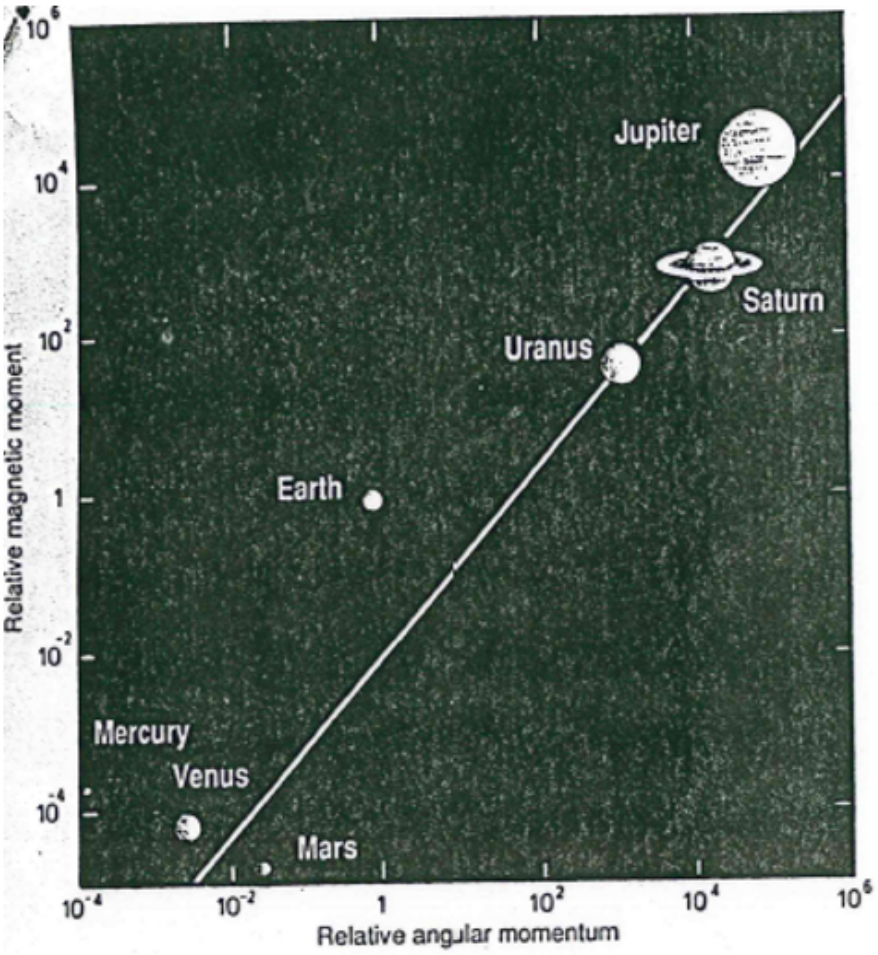


Figure 4.3: SOLAR WIND EFFECT: In this graphic there is the angular momentum on X-axis (rotation of the planet) and the magnetic momentum on Y-axis (magnetic field). We can observe that Mars has a so low magnetic field while Earth has a strong one, the strongest in terrestrial planets. Also Venus has a low angular momentum. We don't know why and in the future we must to answer.

Solar winds removal

The importance of solar winds removal of the atmosphere was discovered recently, by MAVEN mission on Mars.

Solar winds are composed by electrons, protons and other energetic particles that travel at a speed of $400km/s$ with very high energy and the impact of them on atmosphere is not irrelevant. Indeed the impact can transfer kinetic energy to atmospheric particles sufficient to allow them to escape. So solar winds can rip the outermost layer of the atmosphere off. On Earth this process is not efficient due the magnetic fields that deviate the wind. Indeed the rotational and the magnetic axis are perpendicular to the direction of solar winds, so particles deviates thanks to Lorentz force.

Therefore magnetic fields, that are fundamental for life survival, are very extended (up to the Moon so astronauts can have partially protection) but don't protect Earth atmosphere completely. In particular Earth, despite its size, has strong magnetic fields given generated by electric currents due to the motion of convection currents of a mixture of molten iron and nickel in the Earth's outer core. Instead Venus and Mars now have very small and weak magnetic fields. In particular on Mars, solar winds removal has been very effective.

However, the loss for solar wind does not seem to be significant at least on Earth.

Figure 4.4

UV effect

Resent resources show that also UV component of radiation emitted by the Sun can cause an atmospheric loss but it is necessary to study more this effect.

Bombardment erosion

This effect is caused by impact of the Earth with asteroids, meteorites, etc.. The impact causes a removal of some fraction of gasses due to mechanical and thermal reasons. However, this mechanism was effective at the early beginning of our planet, not now.

Calculation of the surface temperature

The most important parameter to study of a planet is the surface temperature. The effective temperature, temperature at the surface, depends on the energy receives from the Sun e not from the Earth's core: it is the balance temperature obtained by the received energy from the Sun and the emitted energy from the planet (that emits like a black body). This because the crust is an optimal insulation, so the thermal energy is enclosed inside the Earth. We will see how to calculate this fundamental parameter.

4.4 Atmospheric pressure distribution in hydrostatic equilibrium

It is known that in conditions of hydrostatic equilibrium, if we consider a perfect gas, the equation 4.1 is valid, and also the hydrostatic equilibrium equation:

$$\frac{dP}{dz} = -g\rho \tag{4.8}$$

From these two we can obtain:

$$P_z = P_0 \exp(-z/H) \tag{4.9}$$

Where $H = KT/mg$ is called the scale height. The scale height therefore depends on the temperature, on composition of the atmosphere and on the gravity of the planet. It decreases with gravity as it increases with temperature (for example, it grows from the poles to the equator as an increase in surface temperature).

4.5 Tidal Force

Consider now tidal forces. In a simple configuration, they are defined as:

$$F_{tidal} = G\left(\frac{Mm_1}{(D-r)^2} - \frac{Mm_2}{(D+r)^2}\right) = GMm\left(\frac{1}{(D-r)^2} - \frac{1}{(D+r)^2}\right) \tag{4.10}$$

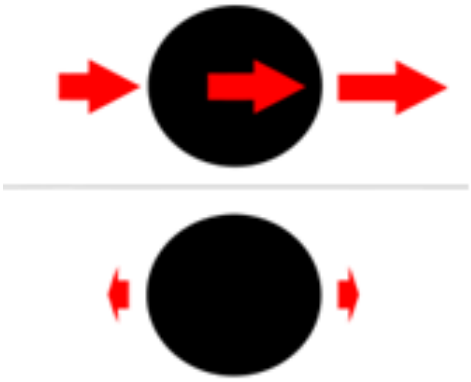


Figure 4.5

Using the approximation derived from Maclaurin series development we get:

$$F_{tidal} = GMm\frac{4r}{D^3} \tag{4.11}$$

A body with mass $2m$ (composed of two identical masses m) is in equilibrium conditions if the internal force of gravity exceeds one of tide, that is:

$$F_g = \frac{Gm^2}{(2r)^2} > F_{tidal} \tag{4.12}$$

then:

$$\rho > 4M/D^3 \tag{4.13}$$

Often the value 10 is used as the coefficient instead of 4 to take into account other effects including, fro example, rotation. If the above condition is respected, a static tidal force has only the effect of an elastic deformation. However, in the much more frequent case of a variable force, then dissipation energy is produced inside the bodies, with heat production. In the case of the Earth we have a total of Sun and Moon tidal effects giving about $4TW$ of power, or about $7mW/m^2$ (for comparison the total geothermal flux is about $80mW/m^2$, so about 10 times higher). The estimated tidal heating (due to the Earth) inside the Moon is actually about four order of magnitudes lower.

4.6 Electromagnetic emission from the planets, effective temperature, greenhouse effect

The electromagnetic emission of the planets includes the diffused radiation of solar origin, which dominates in the visible, and the emission of thermal radiation. It is easy to see that the thermal emission has a peak in the far infrared, between about 10 microns of Mercury and the 100 microns of Pluto, while in optical the radiation is largely dominated by the diffused solar component with a peak at about 0.5 microns.

The bodies of the solar system are therefore visible only because solar radiation is largely concentrated between 0.4 and 0.8 microns. Part of this radiation is absorbed, part is diffused back to the space. The relative proportion depends on the absorption coefficient, or rather on the quantity defined as the **Bond albedo**, which is a measurement of the amount of light reflected from the surface of a celestial object, such as a planet, satellite, comet or asteroid. The albedo is the ratio of the reflected light to the incident light. The maximum value of the albedo is 1 for a perfectly reflective surface, while it

tends to zero for a black body. The average albedo of the Earth is between 0.32 and 0.36. Said A the albedo of Bond, the quantity of relative radiation absorbed is therefore $1 - A$.

An effective temperature of a body of the Solar System can be defined as the equilibrium temperature of a sphere that emits as a black body, located at a distance D from the Sun, of diameter r and albedo A. We can write a relation between the Sun energy absorbed by the planet and the irradiated one, in the hypothesis that the thermal emission is like a black body:

$$S(1 - A)\pi r^2 = 4\pi r^2 \sigma T^4 \tag{4.14}$$

where S is the solar flux at the distance of the planet, A the albedo, T the effective temperature, r the radius of the planet, σ the constant of Stefan-Boltzmann. The equation is valid if the object is in rapid rotation, otherwise the term in the second member must be divided by two, to take into account that the radiation is dominated by a single hemisphere. Then T, so defined, is the effective temperature.

We see that T depends on the albedo and the heliocentric distance, while it is independent of the size of the object. This is true, however, up to the size of particles of the same order as the wavelength, where the law of Mie must be taken into account.

In the opposite case of very massive bodies we will have to instead take into account the sources of internal energy that can, as in the case of Jupiter and Saturn, be comparable to the amount of energy received by the Sun.

The actual temperature of the Earth thus calculated, assumed an albedo of 0.33, gives $263K$, about 30 degrees lower than the average surface Earth temperature.

The presence of an atmosphere around a planetary body modifies its temperature in the sense that it acts as a filter for the infrared radiation sent back towards the space. In the latter case, part of the radiation diffused by the soil is reabsorbed by the atmosphere which in turn will radiate both towards the space and towards the ground. It is the case of **greenhouse effect**, which on Earth is due to the absorbing action of water vapor and carbon dioxide of the infrared radiation emitted from the ground.

The greenhouse process

The basic concept of green house effect is that we are treating the atmosphere like a region of the space above the ground, comparable to a thin screen due to its very low extension compared to the Earth radius, which absorbs part of the IR radiation and re-emit it again into two directions: to space and to ground. In particular the IR radiation hit the molecules that can absorbed this wavelength and they re-emit the radiation. So it is a normal ionization and recombination.

There are special components of atmosphere absorbing IR radiation. The dominant gasses, N and O, are neutral, completely transparent to IR radiation. They have not absorption line or bands but only more complex molecules, composed at least by three ions, are suitable for this effect, for example water vapor, carbon dioxide, eventually methane or other more complex organic compound. Indeed they have more degrees of freedom in oscillation between ions, so they can create much more absorption lines and bands. For example methane (CH_4) has 5 ions and a power which is 100 times more the carbon dioxide absorbing IR radiation.

Ones the IR radiation is absorbed by molecules, the absorbed energy is usually used by ions to do an electronic jump to higher levels (a quantum jump). Then the go in a exited status that could be rotational, electronic etc. This status is unstable so the day re-emitting. In this case the green house effect could not occur. In order to have an heated atmosphere, it is necessary a transfer of this amount of energy to the dominant gasses, N and O, by collision. In particular, this process can heat up the temperature of a planet only if the the time of decay is bigger than the time of collision, because the energy absorbed must be transferred to other molecules.

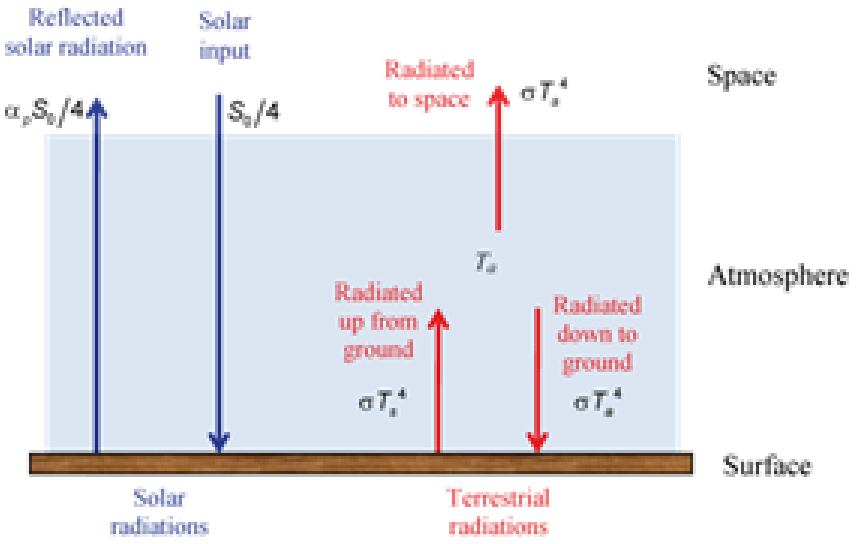


Figure 4.6

F.g. for CO_2 the decay time is of the order of 10^{-3} *sec* and the time of collision depends on the density of the atmosphere; on the ground this time is very short, of the order of 10^{-7} *sec*. So time collision is smaller than decay time.

Instead the atmosphere is almost completely transparent in visible band. In general the atmosphere is transparent at shorter wavelengths and it is opaque at longer wavelengths.

The greenhouse effect increases the temperature from almost 400 degrees in the case of Venus to about 30 degrees for the Earth, to a few degrees for the rarefied Martian atmosphere.

To quantify the greenhouse effect we can use a simplified model where the atmosphere is seen as a partially opaque body, with an emissivity coefficient that is currently estimated around 0.7 for Earth atmosphere. The emissivity can vary from a minimum of 0 (for a perfectly transparent or reflective body) to a maximum of 1 (for a black body).

This additional radiant element has a temperature similar to that of the soil and radiates infrared radiation both towards the outer space and towards the ground in equal proportion. It is therefore a question of adding an additional component to the first member of equation 4.14 which takes into account this additional energy contribution in thermal equilibrium:

$$\frac{1}{2}\epsilon 4\pi r^2 \sigma T^4 + S(1-A)\pi r^2 = 4\pi r^2 \sigma T^4 \tag{4.15}$$

We then easily obtain:

$$T^4 \cdot (4\sigma - 2\sigma\epsilon) = S(1-A) \tag{4.16}$$

Please note that these equations are based on the energy equilibrium.

An alternative way to express the correct temperature for the greenhouse effect is to evaluate the ratio between actual temperature (T_e) and temperature with greenhouse effect (T_s):

$$T_s^4/T_e^4 = \frac{1}{(1-\epsilon/2)} \tag{4.17}$$

or

$$T_s = 1.106 T_e \tag{4.18}$$

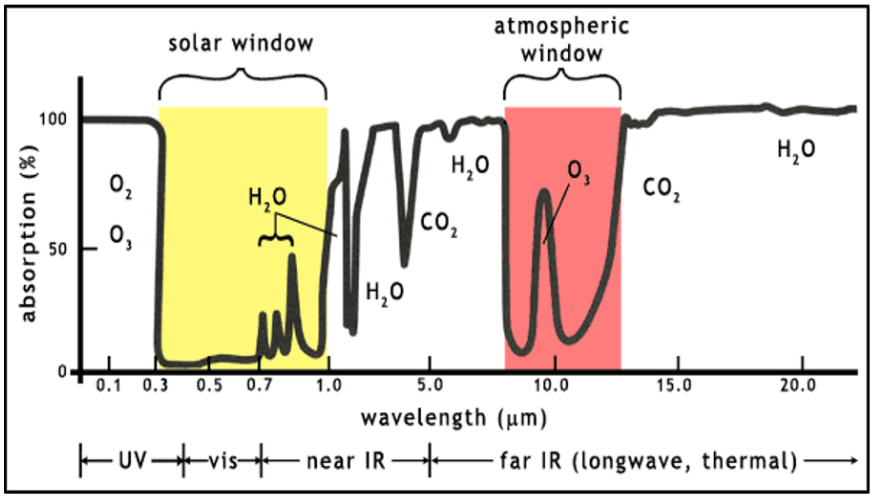


Figure 4.7: Earth spectrum concentrated on IR range.

This result indicates a contribution of the greenhouse effect of about 30 degrees, so $T_s = 293K$. From the equations we see that the greenhouse effect is determined by the emissivity coefficient (if we put it at zero, $T_s = T_e$). The theoretical estimate of ϵ as a function of the content of gases in the atmosphere is a crucial problem of climatology. In general it can be assumed that the emissivity coefficient is given by the sum of the contributions of the gases that have absorption lines in the mid-infrared, in particular water vapor (at 70%) and carbon dioxide, in addition to other gases present in traces like ozone and methane. The water vapor content grows rapidly with the temperature and therefore also increases the produced greenhouse effect, while the carbon dioxide reacts slowly to the temperature, not necessarily in a positive way, with a more complex mechanism, basically following Urey's reaction. From equation 4.17 it is evident that a significant increase of the emission coefficient of the atmosphere, up to a maximum of 1.0, would increase the temperature a few degrees ($263 \cdot 1.19 = 312K$). However the above calculation do not takes into account the temperature difference between the surface and the atmospheric layer.

A more accurate calculation takes into account the equilibrium temperature of the atmosphere and that of the soil.[1].

Assuming that the surface temperature is different than that of atmosphere, input and output radiation including an atmospheric "thin screen", with emissivity ϵ, can be written as:

$$S(1-A)\pi r^2 = 4\pi r^2 \sigma T_s^4 (1-\epsilon) + 4\pi r^2 \epsilon \sigma T_a^4 \tag{4.19}$$

where $S(1-A)$ is the fraction of absorbed energy per surface unit, T_s is surface temperature and T_a is atmospheric temperature. σT_s^4 is the black body emission. Doing the balance energy from space, the first term on right represents the fraction of radiation emitted by ground and going thought atmosphere and the second is an additional flux coming from the thin layer and emitted to the space; this last one is a purely emitted radiation.

The emissivity ϵ, the transmission coefficient, is given by 70% of water vapour, and about 30% of CO_2 (there is also a contribute of methane). As visible in figure 4.7, which represents part of Earth spectrum, around 10 microns, peak of IR radiation, it is absorbed CO_2 bands on longer λ and water vapor at shorter λ. This define the opacity of absorbing layers.

The thermal equilibrium of the atmospheric "thin screen" is:

$$\epsilon \sigma T_s^4 = 2\epsilon \sigma T_a^4 \tag{4.20}$$

[1]For more details see at `http://acmg.seas.harvard.edu/people/faculty/djj/book/bookchap7.html`

The latter gives:

$$T_a = T_s 2^{-1/4} \tag{4.21}$$

so $T_a = 241K$ or $-32C$: this is the temperature of a high altitude layer. So, knowing ϵ and T_a (energy balance of the layer, equaling the radiation emitted by the layer and the radiation received and absorbed by the layer), we can get T_s:

Replacing T_a in equation 4.19, we get:

$$T_s^4 = S(1-A)/4\sigma(1-\epsilon/2) \tag{4.22}$$

For $\epsilon = 0.77$ and $A = 0.33$ we get $T_s = 288K$, just a bit higher than the real temperature of about $287.9K$ (reference of 2017). This calculation is not perfect because "thin screen" is an assumption. Atmosphere is actually an extended layer, composed by many different layers at different temperature. For example, layers nearer to the ground receive more radiation from soil than the upper ones. In general it is necessary to derive another equilibrium equation in order to take this structure in account. A **multi-layer calculation** can give a more accurate result, for example putting an infinite number of layers and integrate the equations. However this approach is mathematically correct but it is not so easy because layers are not perfectly parallel, due to complex meteorology, there are clouds changing the layer structures, there is a strong convection on changing gradient of temperature etc. This theoretical method as been checked some times but they failed because we need more details on this theory.

Effective temperature

The effective temperature of a body such as a star or planet is the temperature of a black body that would emit the same total amount of electromagnetic radiation. Effective temperature is often used as an estimate of a body's surface temperature when the body's emissivity curve (as a function of wavelength) is not known.

When the planet's net emissivity in the relevant wavelength band is less than unity (less than that of a black body), the actual temperature of the body will be higher than the effective temperature. The net emissivity may be low due to surface or atmospheric properties, including greenhouse effect.

The area of the planet that absorbs the power from the star is A_{abs} which is some fraction of the total surface area $A_{total} = 4\pi r^2$, where r is the radius of the planet. This area intercepts some of the power which is spread over the surface of a sphere of radius D. We also allow the planet to reflect some of the incoming radiation by incorporating the parameter a, the albedo. Remember that an albedo of 1 means that all the radiation is reflected, an albedo of 0 means all of it is absorbed. The expression for absorbed power is then:

$$P_{\text{abs}} = \frac{LA_{\text{abs}}(1-a)}{4\pi D^2} \tag{4.23}$$

The next assumption we can make is that although the entire planet is not at the same temperature, it will radiate as if it had a temperature T over an area A_{rad} which is again some fraction of the total area of the planet. There is also the emissivity factor ϵ that represents atmospheric effects. The Stefan–Boltzmann law gives an expression for the power radiated by the planet:

$$P_{\text{rad}} = A_{\text{rad}}\epsilon\sigma T^4 \tag{4.24}$$

Equating these two expressions and rearranging gives an expression for the surface temperature:

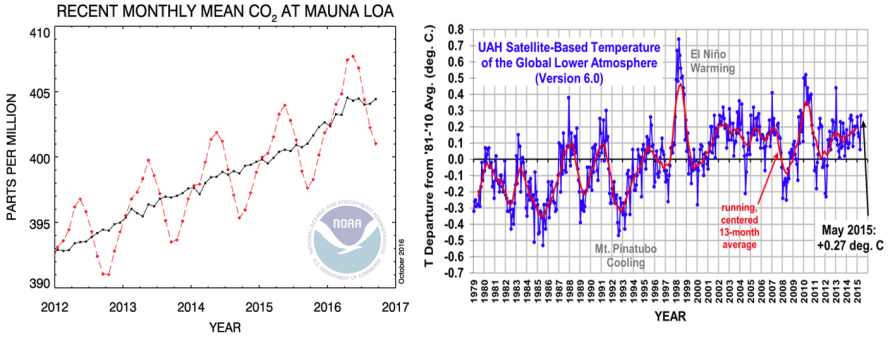


Figure 4.8: On left, CO_2 increase. On right, T increase.

$$T = \sqrt[4]{\frac{A_{\text{abs}}}{A_{\text{rad}}} \frac{L(1-a)}{4\pi\sigma\epsilon D^2}} \tag{4.25}$$

Note the ratio of the two areas. Common assumptions for this ratio are 1/4 for a rapidly rotating body and 1/2 for a slowly rotating body, or a tidally locked body on the sunlit side. This ratio would be 1 for the subsolar point, the point on the planet directly below the Sun and gives the maximum temperature of the planet.

Also note here that this equation does not take into account any effects from internal heating of the planet.

Remembering that power is the amount of energy transferred per unit time, the energy absorbed by the planet per unit time is:

$$E_{absorbed} = E_p \pi r^2 (1-a) \tag{4.26}$$

where E_p is the energy emitted by planet, r the radius of planet and a the albedo.

Increase of CO_2 and of temperature

Different studies state that CO_2 is rapidly increasing in the last decades, now about 400 part per million, and also they discovered a temperature increase in the last years (from the second half of twentieth century). We have many data (for temperature coming from satellites and weather-station) buy we have also to study more. Is Earth an exception? Is this phenomena linear? Increasing CO_2 and water vapor in the atmosphere, we will have deeper and deeper lines until the saturation point so we can assume that this process is not linear but we have to study more.

In particular form satellites, it is much easier to measure the radiation and get the temperature, measuring the emitted radiation from the ground, called also *skin* emission. Anyway the temperature measured increases with a strongly dependence on the day while CO_2 is increasing constantly, mainly due to anthropic activities. We are pretty sure of this origin because photosynthesis (and life in general) prefer the lighter isotopes, the ^{12}C with respect to ^{13}C. Therefore we conclude that the ^{12}C increase is the result of burning organic processed C. In inverted scale, it indicates a ^{13C} relative depletion.

In figure 4.8, it is visible the constant instrumentation of CO_2 and the growth of temperature in the last years. Is it visible that they are not perfect parallel due to some additional forces or effects. In particular on temperature increase they have been defined the minimums due to volcanic eruptions with ejection in the atmosphere of dust and aerosol (creating anti-greenhouse effect, as we will see in the next section).

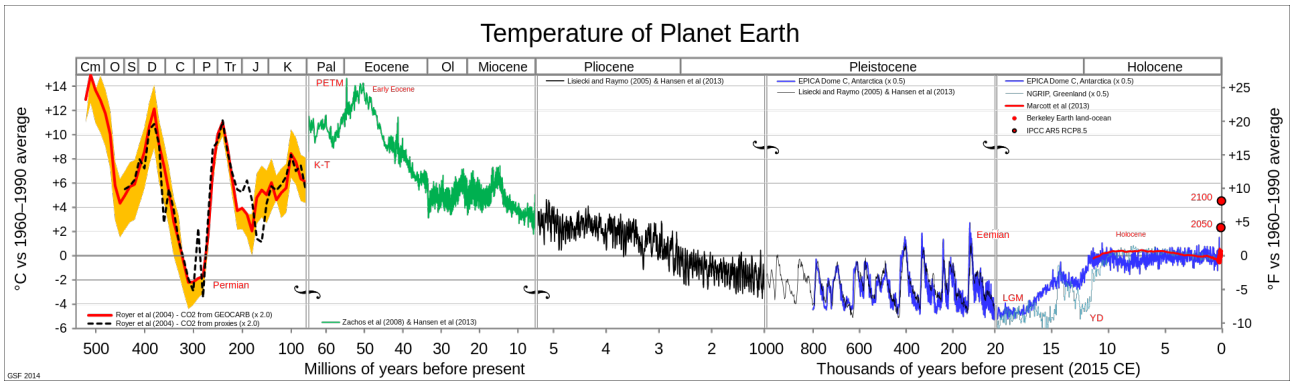


Figure 4.9: Temperature changes in time.

Another interesting feature is the Antarctic plateau, the central region of Antarctic, a very extended area about 2000 km of radius characterized by a constant temperature in the last 50 years. This is surprising because temperature due to water vapor is negligible, this region is dominated by CO_2 which is increasing there. Why temperature is no increasing? Maybe due to whether conditions but we don't know.

Moreover, due to temperature increase, seasons are changing, in particular the peak of temperature is shifted; it is anticipated of 4 or 10 days. The consequences are extreme whether episodes more and more frequent.

Figure 4.9 it is represented the temperature change from Earth formation until now in logarithmic scale. Good and reliable measurements are made from 500 million years until now, due to information recorded inside arctic ice. The last and the most stable increase of temperature is happening now, from some decades until now.

Anti-green house effect

The above calculations don't take into account neither the effects of aerosols neither the uneven distribution of the temperature across the Earth surface. Recent calculations give a lower "effective" temperature, then the need of a higher greenhouse contribution. The aerosol contribution has been discussed by Turco et al., (1991). They calculated the temperature, including aerosols (they call them "smoke"), after the pollution due to a nuclear explosion, which has an effect very similar to an asteroid collision. The complete equation is:

$$\frac{T_g}{T_0} = \left\{ \frac{f}{2-\epsilon_s} \left[1 + e^{\tau_s/\mu_0} \left(\frac{2+\epsilon_a(1-\epsilon_s)}{2-\epsilon_a} \right) \right] \right\}^{1/4} \tag{4.27}$$

Here f is a coefficient related to the two albedos (ground and aerosols, or smoke) and it is near to 1, s means smoke and a means clean atmosphere, T_0 is the effective temperature. Limit cases to zero aerosols (smoke) contribution and to a maximum of $\tau = 1$ (for aerosols), give respectively:

$$T_g/T_0 = (2/2-\epsilon_a)^{1/4} \tag{4.28}$$

and:

$$T_g/T_0 = 1 \tag{4.29}$$

The latter shows that a heavy aerosol (smoke) effect can counterbalance the greenhouse effect. In particular when $\tau = 1$, they completely counter balance the greenhouse effect.

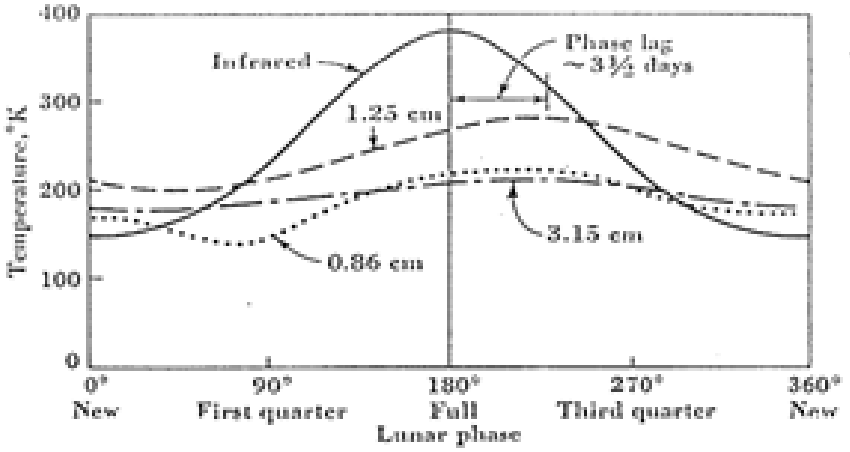


Figure 4.10: Moon emission in IR and radio range.

On Mars and Moon

On Mars the greenhouse effect is almost negligible so the effective temperature is almost identical to real one so atmosphere doesn't count. We can say that this is obvious due to ground pressure on this planet about 10 millibars, 100 times less than Earth but this is not the right reason. On Mars greenhouse effect is negligible due to pure CO_2 in the atmosphere. Relative pressure on Earth by CO_2 is much lesser on our planter than on Mars so we would expect a stronger effect on Mars but lower total pressure and lower temperature make this effect negligible. Indeed we know from spectroscopy that at lower pressure and lower total temperature, the line are narrower so the absorption effect is substantially lower.

Instead on Moon, during Apollo missions, astronauts put some thermometers on the ground at different depth and they remained there for many years, collecting data. They registered that temperature is increasing with time on the ground but nothing to do with CO_2. Why? Since this increasing effect is stronger near surface than in depth, it is something connected to surface: astronauts walked around to install devices and they changed the local albedo of the ground, of regolith, which is a bit darker than the original one so it is absorbed more energy and T increases.

Greenhouses

Greenhouses are based on a different principal because they are structures with roof, transparent to shorter λ but cutting thermal convection so warm bubbles created inside are stopped here, increasing local temperature. So there is nothing to do with absorption of infrared radiation, even the name is the same.

Insight into measurements

The thermal and radio millimetric and centimetric emission is currently used to measure the surface temperatures of the bodies of the Solar System. Therefore identified the peak which depends on temperature (the λ of the peak is much longer as T decreases, black body law), it is not so difficult to measure the temperature of a planet.

The radio emission of the Moon has been detected at centimetric wavelengths several decades ago. Observations conducted throughout the lunar cycle show systematic variations in brightness temperature. However the amplitude of the variations is smaller than that observed in the infrared between 8 and 14 microns. Furthermore, a phase lag of nearly 45 degrees was noted. The infrared curve of Petitt and Nicholson is shown in comparison, with centimetric radio measurements in figure 4.10.

The infrared temperature is symmetrical with respect to the full moon phase, with a maximum in full moon and a minimum in new moon. In radio instead there is not always symmetry with respect to the maximum temperature, while the amplitude decreases with the wavelength. At 75 cm the width is less than 10 percent. The small temperature variation compared to the infrared indicates that the centimetric radiation is originated in depth below the surface, while the infrared radiation comes from the surface. A proof of this explanation comes from the absence of radio variation during lunar eclipses when solar radiation is blocked for about an hour, while the infrared component shows a drop of about 200 degrees. The average temperature of about 250 K corresponds to the actual temperature for a body located at an astronomical unit from the Sun with the typical albedo of the Moon. The coefficients of the curve, together with the phase shift, form the basis for the quantitative interpretation from which it is possible to deduce indications on the thermal inertia and the dielectric constant of surface materials.

The observations of Venus are instead limited to the lower conjunction, where a brightness temperature of about 600 K is measured at centimetric wavelengths, while the temperature drops to around 350 K in the millimeter. It is evident that the millimetric measurements (which coincide with the infrared results) come from the top of the clouds, while in the centimeters we observe the emission from the Venusian soil or subsoil. At the lower conjunction Venus appears at the minimum distance from the Earth, with the dark part visible from Earth. There is therefore to expect that the temperatures on the opposite side are higher. Observations around the conjunction phase, at less than 3.2 cm, although they do not extend for the whole period as in the case of the Moon, nevertheless they show a temperature variation of almost 200 K and a shift of 12 degrees of the minimum from the lower conjunction. The direction of the shift must be interpreted as rotation of the planet in a retrograde sense. Observations at 10 cm confirm these results, with a variation in temperature reduced to less than 100 K, while the phase difference from the minimum to the lower conjunction increases to 17 degrees. Also in this case the data are interpreted as emission from the subsoil at longer wavelengths. It should be noted that the measured temperatures are lower limits, as the real value depends on the emissivity of the soil.

The radio emission of Mars is complex and gives systematically lower temperature values than the infrared ones. At 3 cm this turns out to be around 200 K, against 240 K in the thermal infrared at 20 microns. The observations of Mars refer to the day side, given that it is difficult to observe the planet far from the opposition. The highest infrared temperature can therefore be explained in terms of thermal inertia and conductivity of the surface layers as explained above. Furthermore the temperature variation with the longitude of the central meridian will still depend on the different thermal inertia in different regions. The regions of low inertia will reach a high temperature in the infrared, but not necessarily in radio. The temperature curves can therefore also appear in anti-phase.

The dominant radiation from the giant planets, Jupiter in particular, differs considerably from the thermal one, mostly at decametric wavelengths where it reaches brightness temperatures of millions of degrees. It is a non-thermal synchrotron radiation by relativistic electrons. The observed emission also comes from a region that extends almost 3 times beyond the optical image. At centimetric wavelengths, instead, radiation is thermal, coincides with the optical region and gives a brightness temperature of about 140 K, coherently with the infrared temperature that measures the emission of the surface of the ammonia clouds.

Sometimes plots of emitted radiation by a planet can have a strange shape at minor λ due to radiation of the Sun reflected.

4.7 Generalities of the Solar System

The Solar System is constituted by the Sun and by the bodies rotating around it. There are 8 planets, Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus and Neptune (Pluto is a dwarf planet) and a set of minor bodies such as asteroids, comets, interplanetary dust and a solar wind composed of nuclear particles, mainly helium and hydrogen nuclei. The mass spectrum of these minor bodies covers an

interval of as many as 40 orders of magnitude in mass.

The Sun, together with the Solar System, is on the periphery of the celestial system to which it belongs, the Galaxy or Milky Way, at a distance of about 8 kpc (more precisely the recent data give 8.2 kpc) from the galactic center, to only a dozen parsecs from the symmetry plane, inside the fragment of a spiral arm known as the Orion arm.

In astronomy the rotation motion of one body around another is called revolution, while rotation means the rotation of a celestial body on itself. The orbits of the planets all lie on the **same plane** with small deviations, in particular of the planet closest to the Sun, Mercury. The **ecliptic** is defined as the maximum circle that describes the Sun on the sky in the course of the year and it corresponds physically to the plane of the Earth's orbit around the Sun.

Revolutions and rotations

The Sun rotates on itself with an average period around twenty days, corresponding to an equatorial tangential speed of about 2 km/s. The orbital characteristics of the planets are largely dominated by the gravity of the Sun.

The planets rotate around the Sun in the same direction and the orbits are almost strictly circular, with the exception of Mercury, the planet closest to the Sun, and Pluto, which is the farthest. Mars also has an orbit with an appreciable ellipticity ($e = 0.09$ vs. 0.017 of the Earth) a property that, together with its proximity to the Earth has allowed Kepler, from the study of the motion of Mars, to derive the three laws that regulate the motions of the planets around the Sun. Kepler's laws are empirical, approximate laws, which are often used in the study of the solar system due to their simplicity and immediate application which express some important relationships between the orbital elements.

The planets in turn rotate around their own axis (axis of rotation) which, for most of the planets, is almost perpendicular to the plane of the orbit. Important exceptions, only partially explained, are that of Venus, which has a retrograde motion, and of Uranus, which has the axis of rotation almost parallel to that of the orbit.

The planetary orbits are characterized by parameters called the orbital elements, the most important of which are the **eccentricity**, the **period**, the **inclination** (with respect to the plane of the Earth's orbit around the Sun, called the plane of the ecliptic when it is seen projected in the sky) and the position of the **line of the nodes** (the line of the nodes is the intersection of the planet's orbit with the reference plane of the Earth's orbit).

4.7.1 Regularity and properties of the solar system

Distribution of the distance

The Solar System has important regularities, only partly explained in terms of general formation processes of planetary systems. An important regularity in the solar system is the clear separation of the planets into two main families with distances from the Sun, and distinct physical properties. The 4 planets closest to the Sun (Mercury, Venus, Earth and Mars) are all within about 250 million kilometers from the Sun (within 1.5 AU), while the next, Jupiter, is at a distance of almost one billion kilometers. The outer planets (Jupiter, Saturn, Uranus, Neptune) are also called Jovian planets and are distinguished not only for the distance (from 5.2 to more 30 AU) but also for the lower temperature, for the large size and for the density. Finally, the farthest planet or minor planet, Pluto, is not easily classifiable, being very distant, of small size and mass.

The surface temperature values of the planets that can be measured from the Earth in the range from about 300 C of Mercury, in the part facing the Sun, to around -250 C of Pluto. While the interpretation of temperatures is fairly obvious, there is still no consistent explanation regarding the distribution of the masses. It should be noted in this regard that the planetary systems around other

stars so far discovered show several cases of very massive planets even at very small distances with respect to the star around which they rotate.

Mass distribution and density

Sun contains the 99% of total mass of Solar System. The distribution of the masses of the planets follows a characteristic pattern, with small masses, comparable to the terrestrial one, for the 4 planets closest to the Sun while the most distant planets all have masses of several orders of magnitude higher, with the sole exception of Pluto, which instead has a mass lower than that of the Moon (0.178 Moon masses). For these common properties (proximity to the Sun, mass and size), the 4 planets closest to the Sun are called terrestrial, the other Jovians.

Another important difference between the two families is the density. Very high for terrestrial ones (between 3 and 5 times the density of water), to that of the Jovian planets which is comparable to the water. Of course, as we will see, difference in density means difference in chemical composition. We thing that high density of Mercury, for example, is due to iron while we are sure that low density of Jovians is due to the lighter elements they are composted, like H and He mainly.

Rotation

The planets of the Solar System rotate around their own axis, in the same direction, with two important exceptions: Venus rotating in a retrograde direction, and Uranus which has the rotation axis almost parallel to the orbital plane. While for the case of Uranus the magnetic axis forms an angle of almost 90 degrees with respect to the ecliptic, and therefore we can explain the anomaly with a violent impact, the problem of Venus remains to be explained. According to Colombo the retrograde rotation of Venus is due to the secular effect of the attraction of the Earth with the result of an Earth-Venus resonance.

The rotation periods are of the order of days, or several months, for the terrestrial planets, while they go down to a fraction of day for the Jovian planets.

Satellites

Most planets in the Solar System have natural satellites. In the case of the terrestrial planets, the Earth possesses only one (the Moon) and Mars two (Fobos and Deimos), while Mercury and Venus, the closest to the Sun, do not have any. It is not clear if this is linked to the process of planet formation or if the gravitational field of the Sun has prevented their formation.

The giant planets, on the other hand, have many satellites and of considerable dimensions. Jupiter has 4 main satellites (Galilean satellites), of dimensions comparable to Mercury, and a few dozen smaller satellites. There are also numerous Saturn satellites. The largest satellite, Titan, is the only satellite in the Solar System to have an extended atmosphere. There are also numerous satellites around Uranus and Neptune. Pluto also has 5 satellites. One of these, Charon, has a size comparable to Pluto itself. In this context the Earth is still a rare case (similar to Pluto) because the Moon has a size of almost a third compared to the Earth. The properties of the Earth-Moon system will be dealt with in a subsequent chapter.

In figures 4.11 and 4.12 there is a schematic summary of main planet properties.

4.8 Minor bodies of the solar system: meteorites, asteroids and comets

Asteroids and meteorites

The Solar System contains numerous bodies with dimensions smaller than one hundred kilometers up to a fraction of a millimeter. Many of these are found between the orbit of Mars and Jupiter (**asteroid belt**) while others, recently discovered, are very far, beyond the orbit of Neptune (**transnettunians**).

PLANETARY PROPERTIES

ORBITAL PROPERTIES

Planet	Mean Distance From The Sun ($\times 10^6$ km)	Mean Distance From The Sun (Earth = 1)	Period To Revolve Around The Sun	Mean Orbital Velocity (km s^{-1})	Orbital Inclination (Earth = 0)	Orbital Eccentricity
Mercury	57.91	0.387	88.0 days	47.87	7°	0.2056
Venus	108.21	0.723	224.7 days	35.02	3.394°	0.0067
Earth	149.6	1.0	365.25 days	29.78	0°	0.0167
Mars	227.92	1.524	687.0 days	24.13	1.850°	0.0935
Jupiter	778.57	5.204	11.75 years	13.07	1.304°	0.0489
Saturn	1,433.53	9.582	29.5 years	9.69	2.485°	0.0565
Uranus	2,872.46	19.201	84 years	6.81	0.772°	0.0457
Neptune	4,495.06	30.047	165 years	5.43	1.769°	0.0113
Pluto	5,869.66	39.236	248 years	4.72	17.16°	0.2444

PHYSICAL PROPERTIES 1

Planet	Diameter (km)	Diameter (Earth = 1)	Rotational Period	Oblateness	Axial Tilt
Mercury	4,879	0.38	58.65 days	0.0	2.0°
Venus	12,104	0.95	-243.02 days	0.0	177.4°
Earth	12,742	1.0	23 hrs 56 mins	0.0034	23.45°
Mars	6,780	0.53	24 hrs 37 mins	0.005	25.19°
Jupiter	139,822	10.97	9 hrs 55 mins	0.065	3.12°
Saturn	116,464	9.14	10 hrs 40 mins	0.108	26.73°
Uranus	50,724	3.98	-17.24 hours	0.03	97.86°
Neptune	49,248	3.87	16.11 hours	0.02	29.56°
Pluto	2,390	0.19	-6.38 days	0.0	119.6°

PHYSICAL PROPERTIES 2

Planet	Mass (Earth = 1)	Density ($\times 10^3$ kg m^{-3})	Surface Gravity (Earth = 1)	Escape Velocity (km s^{-1})	Escape Velocity (Earth = 1)
Mercury	0.0553	5.43	0.378	4.3	0.384
Venus	0.815	5.25	0.907	10.36	0.926
Earth	1.0	5.52	1.000	11.19	1.0
Mars	0.107	3.95	0.377	5.03	0.450
Jupiter	317.83	1.33	2.364	59.5	5.32
Saturn	95.159	0.69	0.916	35.5	3.172
Uranus	14.536	1.29	0.889	21.3	1.903
Neptune	17.147	1.64	1.120	23.5	2.10
Pluto	0.002	2.03	0.059	1.1	0.0983

Figure 4.11: Summery of Solar System properties.

THERMAL PROPERTIES

Planet	Solar Irradiance (W m^{-2})	Solar Irradiance (Earth = 1)	Albedo (%)	Surface Temperature (° C)
Mercury	9126.6	6.673	11	467° to -183°
Venus	2613.9	1.911	65	465°
Earth	1367.6	1	37	45° to -60°
Mars	589.2	0.431	15	0° to -100°
Jupiter	50.50	0.037	52	-148°
Saturn	14.90	0.011	47	-178°
Uranus	3.71	0.0027	51	-213°
Neptune	1.51	0.0011	41	-216°
Pluto	0.89	0.0007	30	-223°

OBSERVATIONAL PROPERTIES

Planet	Synodic Period (days)	Apparent Diameter (seconds of arc)	Maximum Apparent Magnitude	Color
Mercury	115.88	4.5 - 13	-1.9	Silvery
Venus	583.92	9.7 - 66	-4.4	White
Earth	-	-	-	Bluish White
Mars	779.94	3.5 - 25.7	-2.8	Red
Jupiter	398.88	29.8 - 59	-2.6	Pale Yellow
Saturn	378.09	14.5 - 20.1	-0.5	Yellow
Uranus	369.66	3.3 - 4.1	+5.7	Green
Neptune	367.49	2.2 - 2.4	+8.2	Blue
Pluto	366.73	0.06 - 0.11	+13.7	Yellow

MISCELLANEOUS PROPERTIES

Planet	Composition of Atmosphere	Discovery	Number Of Moons
Mercury	H_2 He (trace amounts)	Mesopotamia	-
Venus	CO_2 (96%) N_2 (3%) H_2O (0.1%) - Surface pressure: 90 atm - Clouds: H_2SO_4	Mesopotamia	-
Earth	N_2 (78%) O_2 (21%) Ar (1%) - Surface pressure: 1 atm - Clouds: H_2O	-	1
Mars	CO_2 (95%) N_2 (3%) Ar (1.6%) - Surface pressure: 0.02 atm	Mesopotamia	2
Jupiter	H_2 (90%) He (10%) CH_4 (0.7%)	Mesopotamia	63
Saturn	H_2 (97%) He (3%) CH_4 (0.05%)	Mesopotamia	60
Uranus	H_2 (83%) He (15%) CH_4 (2%)	England (1781)	27
Neptune	H_2 (74%) He (25%) CH_4 (1%)	Europe (1846)	13
Pluto	CH_4 N_2 (trace amounts)	USA (1930)	5

Figure 4.12: Summery of Solar System properties.

Depending on the orbital characteristics, these objects are divided into families. The most important families are those called Apollo-Amor, which, in their orbit around the Sun, intersect the Earth and can therefore collide with the Earth. Traces of impacts of these objects are still visible on the earth's surface (eg the Meteor Crater in Arizona, or Monturaqui in Chile).

Meteorites

The smaller bodies are very numerous and often enter the Earth's atmosphere. Depending on their mass and composition they are destroyed at high altitude (50 km) or can reach the Earth's surface. The phenomenon known as shooting stars refers to the high-altitude destruction of small-sized meteorites (around the millimeter), usually of cometary origin, as shown by their periodicity due to the passage of the Earth in a cometary orbit.

How to distinguish a meteorite? The meteorites collected are distinguished from the terrestrial samples by the evidence of signs of superficial fusion produced during the passage through the atmosphere.

Why so important? Meteorites that reach the ground are important residual samples from the formation of the solar system. Unlike the planets, in fact, they did not undergo those processes of fusion and chemical-physical modifications that occurred instead on the planets.

Meteorites are divided into two main categories: ferrous and non-ferrous.

- The ferrous ones are made of a high purity iron-nickel alloy, compact, and of high specific weight, and constitute most of the collected meteorites. They are also the most easily identifiable ones. Iron meteorites should derive from differentiation of already formed asteroids: a body of some kilometers can contain radioactive material, high enough in order to melt rocks so heavier elements are concentrated in the centre and the lighter ones are on surface. Then some destructive effects (collisions) they expose the nucleus to fragmentation.
- Non-ferrous meteorites can be basaltic or carbonaceous and can appear compact or as chondrites (the name "condrule" indicates inclusions of small spheres).

Therefore there are many different types of meteorites.

- Basaltic meteorites are uniform material made by basalt mainly so they are very rare to distinguish from terrestrial basaltic. They are also called shergottites, from the Indian town where they were founded the first time. Comparing isotopic analyses on stable isotopes of those meteorites, such as the ones of O or N, to isotopic analyses of Mars, we conclude that they are samples of Martian rocks ejected by Mars due to impacts. In this way we can study Mars composition.
- The carbonaceous chondrites, on the other hand, are more easily distinguishable and often appear as friable agglomerates of small particles. By their nature they are the meteorites that can hardly survive the impact with the Earth's atmosphere. It is therefore believed, despite their scarcity compared to ferrous ones, that they constitute the most numerous samples in space. We think they are the most interesting because, due to our actual models, they had formed planets during formation phase.
  - A subfamily of chondrites are a type rich in carbon and rocky materials.
- Tectic meterorites are instead a melted of glass and they are found in sparse areas. They are from the Moon, ejected after impacts on Moon surface.

Meteorites cross the atmosphere at speeds around $50 - 70$ km/s. It should be remembered that the Earth rotates around the Sun at a speed of 40 km/s, while the acceleration that an external body undergoes to the solar system that reaches the Earth, can give it a speed of 30 km/s. Taking into account also the acceleration suffered by the Earth's gravity, we see that the maximum combined speed of a meteorite is just over 70 km/s, while the minimum speed is around $30 - 40$ km/s, in excellent agreement with the observations. An exception are some meteorites that reach $100 - 120$ km/s, according to measurements made with the triangulation by radar, of doubtful interpretation.

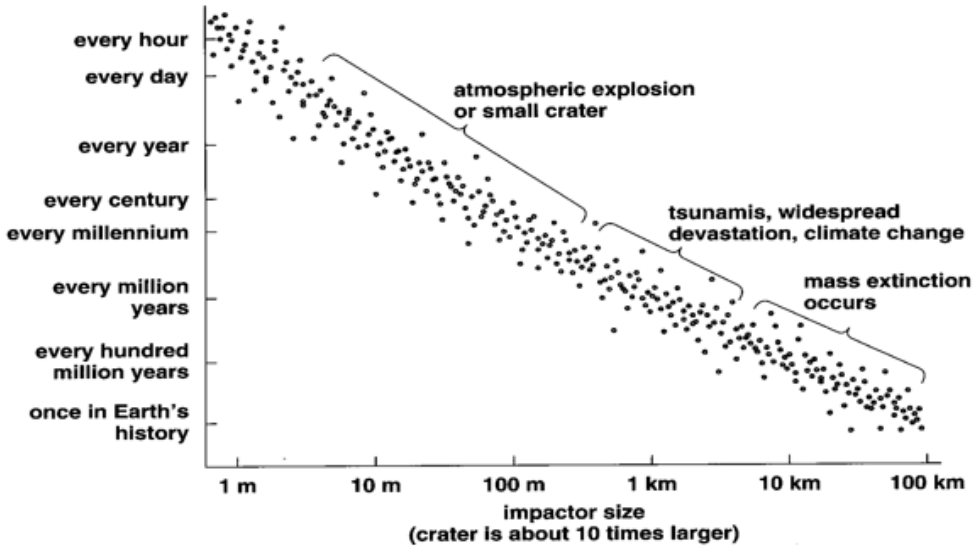


Figure 4.13

The speed of the meteorites is very high and, in terms of kinetic energy the meteorites have an energy, per unit of mass orders, of several orders of magnitude higher than the most known energetic chemical compounds. A gram of a meteorite has an energy of 10^6 J/g.

The dating performed by the radioactive isotope ratios with respect to their derivatives (eg uranium 238 with respect to lead 206) allows the samples to be dated. It is shown that carbonaceous chondrites are the oldest relicts of the solar system and their upper ages converge towards 4.6 billion years, much older than the oldest terrestrial rocks (except for some zircons recently discovered in Australia). This dating, together with some lunar samples, constitutes the basis for the estimation of the age of the Solar System. It is also used by astronomers for models of stellar evolution that are calibrated on the age of the Sun (estimated equal to that of the Solar System).

In figure 4.13 is visible the relation between size per meteorites and the frequency of falls. Objects small than 1 m fall every hour. From 10 to 100 m they fall every millennium of every century. From 1 km to 10000 km every 100 million to 1 million years.

Comets

There is no clear distinction between comets and asteroids, although the first are constituted by volatile materials which, when the comet approaches the Sun (within the orbit of Jupiter), evaporate and give rise to the coma and the characteristic cometary tail.

The theory on the formation of comets derives from the original idea of Wipple, subsequently refined according to the dirty snowball model. The nucleus (about 10 km of radius) of a comet would consist of a set of ice and dust. Specifically it is composed by an inner part, a mixure of water vapor, carbon monoxide, carbon dioxide, and then a thin upper layer of dust. With the evaporation of the ice (mainly ice water) dusts of various sizes are also dragged into the dust tail and leave a long trail in the comet's trajectory (hence the meteoric rains when the Earth passes through these trails). The coma created around the nucleus is about 10^6 km almost spherical and sometimes there is ion tail; both tails are opposite to solar wind direction.

Figure 4.14 represents schematic functioning of comets.

Figure 4.15 reports the comet spectrum.

It should however be noted that in ground spectra no water molecules are observed, which should instead constitute the dominant fraction of the molecules, in the inner part of the coma. Most of the

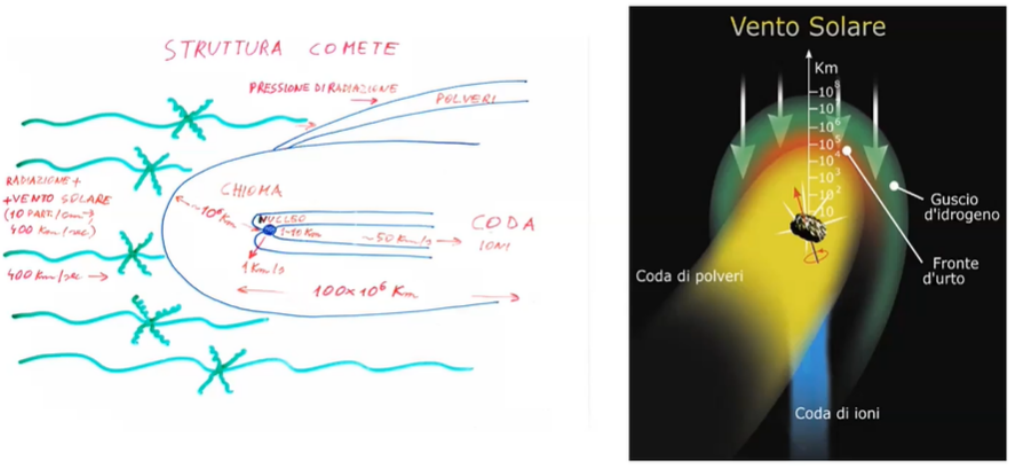


Figure 4.14: Comets

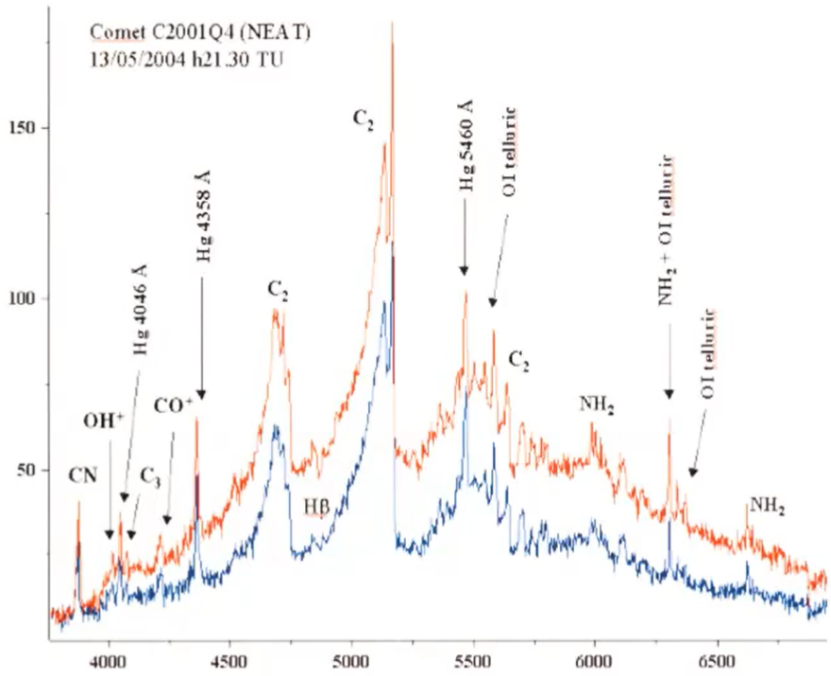


Figure 4.15: Spectrum of a comet measured on the Earth

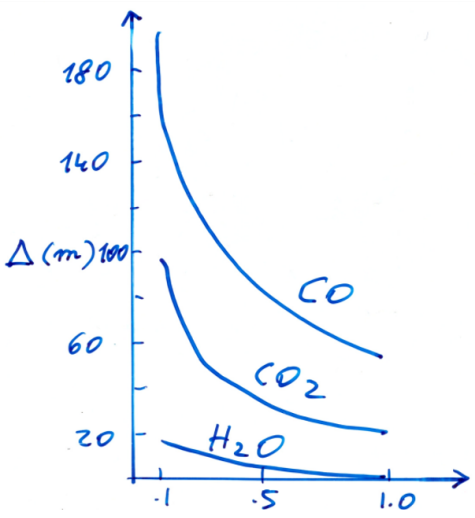


Figure 4.16: Evaporated gasses.

molecules visible in the spectra instead, are those of the CN, CH, Na, C_2 etc. which result from the decomposition of the primary molecules (proteins, carbohydrates, lipids and nucleic acids) by solar radiation, in the immediate vicinity of the nucleus, therefore in a region that is not resolved from the ground. In particular water vapor is easily decomposed by solar radiation so in the spectrum we don't see the original gas; we see the already processed gasses which are on the tail comet.

According to studies from Oort, comets are present in large numbers in the form of inert nuclei, at a great distance from the Sun, well beyond the orbit of Pluto (at about $50.000-100.000$ AU), with orbits characterized by very long periods. As shown by the images of the Halley nucleus, the appearance of the cometary nuclei is of very dark objects, with a diameter of a few kilometers. They have a very eccentric orbit and therefore penetrate the inner part of the Solar System, until they reach the Sun.

Also the comets constitute a sample that has undergone few chemical-physical changes since their formation. Comets can be the main source of interplanetary dust that reaches Earth. It is interesting to note that, according to some theories, on the comets there would be suitable conditions (solar radiation, graphite particles, ice), for the synthesis of complex prebiotic organic substances.

It is clear that at each step the cometary nucleus loses mainly the ice of carbon dioxide and carbon monoxide, while the water ice persists longer, with an estimated loss of around ten meters for each passage to the perihelion (within an astronomical unit). It is difficult to estimate the mass of cometary nuclei: from gas and dust losses they are estimated between 10^{14} and 10^{17} g, while the most precise estimate, by the Giotto probe on Halley's comet, gave 10^{17} g, corresponding to about 10 orders of magnitude less than the mass of the Earth. It is therefore easy to conclude that the total mass of comets in the solar system does not exceed the mass of the Earth. How much materials are evaporated from nucleus of comets are visible in figure 4.16 in which on X-axis ther is the perihelion of the comet in AU, usually between 0.1 and 0.5. The evaporation for every passage at perihelium are 100 meters in CO, 40 meters in CO_2 and almost 10 meters in H_2O, since water vapor is the dominant one.

A fundamental plot to understand which and when gasses are produced by solar radiation hitting the comet, is the one in figure4.17.It shows how a nucleus of a comet, approaching from very distance space and going in the Sun direction, is increasing the production of these gasses by evaporation. The first to evaporate are CO and CO_2 because they are more volatilise, produced almost constantly several tens of AU far away from the Sun. Instead water vapor increases from very low production at 7 or 8 AU, with a jump of 2 order of magnitude at 4 and then it stabilizes at 1 AU. This means that at about 4 AU the ice is strongly evaporating under Sun radiation so within this distance we can not have solid ice. At least we have to go at double distance to reduce a factor 4 the solar radiation to have somehow stable ice.

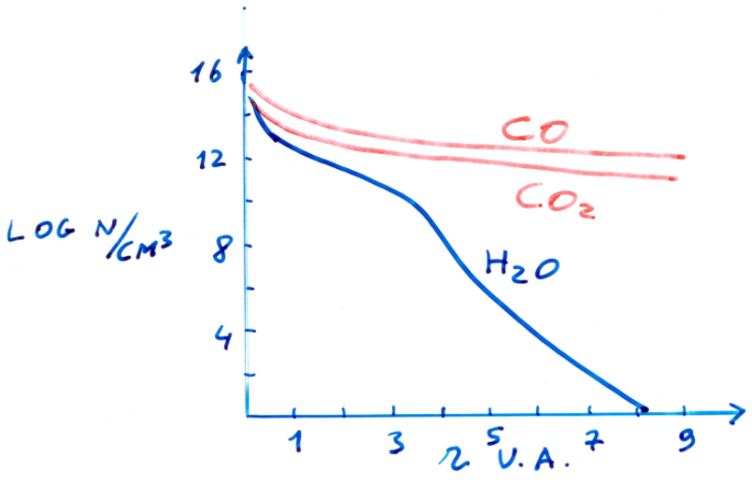


Figure 4.17: Production of gasses.

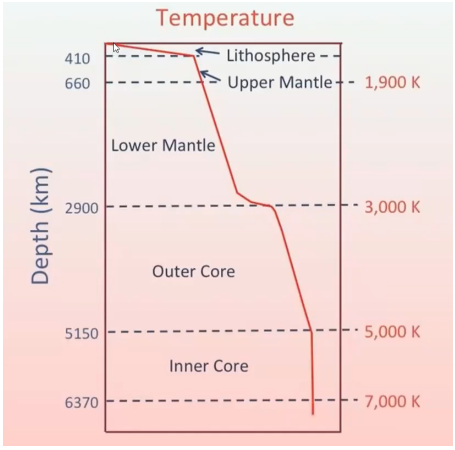


Figure 4.18

Where comes from energy on Earth?

As visible on figure 4.18, temperature inside Earth increases going toward the core. In particular the temperature increases of two degrees every hundred meters. Then, at the Lithosphere, there is a dramatic drop pf temperature due to the fact that the crust is a strong thermal insulator. Indeed there is no relevant heat on Earth surface coming form inner part of the planet: the vast majority comes from the Sun.

Mainly there are three energy sources inside terrestrial planets:

- **accretion**: gravitational potential energy converted into kinetic energy and then kinetic energy is converted into thermal energy;
- **differentiation**: dense materials fall to the core converting gravitational potential energy into thermal energy. The lighter ones rise to the surface. In particular to have differentiation we need 1800 degrees in temperature. Since the surface is just above 0 C, this means that we need about 1500 C to have rocks melted and a viscosity low enough to separate heavier elements from the lighter ones;
- **radioactivity**: In the nucleus heat comes for half from "fossil" heat (primordial heat left over from the formation of Earth) and for half from nuclear energy which is converted into thermal energy. It occurs at the centre of the Earth with radioactive isotopes. Figure 4.19 shows heat flow produced by different isotopes and the total production. The main isotopes are:

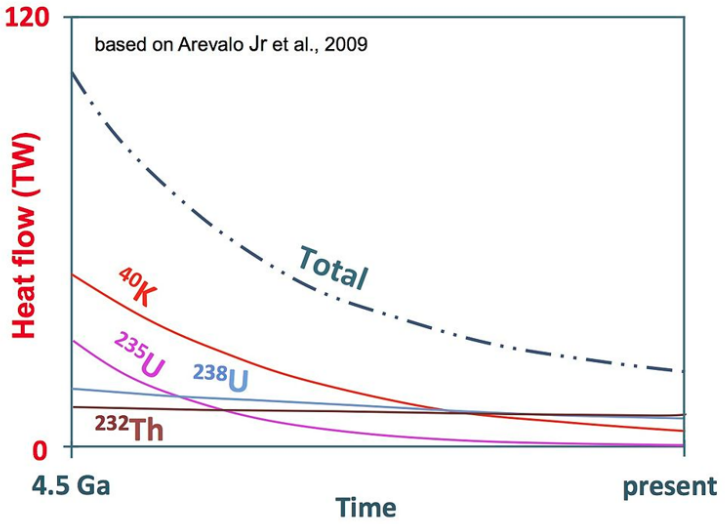


Figure 4.19: Flux heat produced by decay of isotopes.

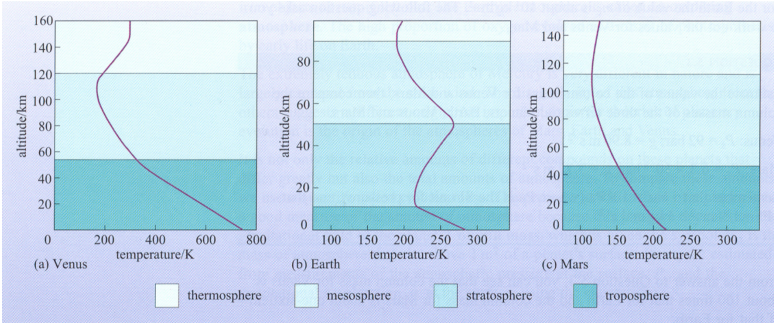


Figure 4.20: Vertical temperature gradient for Venus, Earth and Mars.

- ^{40}K, potassium, necessary to produce ^{40}Ar, aluminium;
- ^{235}U, uranium, very active so it decays in very short time indeed nowadays it is quite rare on Earth;
- ^{238}U, much more stable indeed it is dominant;
- ^{232}Th, thorium, which has a very long time to decay, about 12 $Gyrs$.

Of course the shorter living isotopes are realising more energy. As visible in the total line, the production was much higher in the past than now so we should expect that in the future the release of energy by these isotopes at some point will be so low that heat will not be produced anymore.

Nowadays heat produced inside the inner part of Earth and emitted thought surface is 85 mW/m^2. Moon emits only 18 mW/m^2. For Mars we don't know the values but we expect it should be around 20 mW/m^2 because there are not relevant signatures of geological activity.

Temperature inversions on planets

In figure 4.20 is visible the typical trend of temperature inside atmosphere of Venus, Earth and Mars. Comparing vertical temperature gradient, we can see that only on our planet there is a thermal inversion above 10 km, in the troposphere. This occurs because stratosphere is hotter than troposphere. It is fundamental for preservation of water vapor in atmosphere because at this point of inversion, convection is stopped and H_2O can be conserved and protected above 10 km by the upper layers against UV radiation which would be able to destroy water molecules separating H from O.

Last studies about Pluto

Usually terrestrial planets like Earth have a reactive geology and they present a melted inner part and a floating crust. Going to objects very small, of course they are cooler because ratio between surface and volume decreases so they irradiate more easily compared to the volume and the heat they contain. They cool and evolve more rapidly than bigger objects in the direction of dead bodies in term of geology. We assume that objects with size below $1000 - 2000\ km$ (typically asteroids) could be only geologically dead at age of 4.5 $Gyrs$. However this is not true for Pluto. Pluto and its moon Charon (with a size between Mercury and the Moon) are geologically active. From data registered by Horizon space probe, the inner part of the planet is warm, melted and it is doing some internal processes. In particular Pluto has a very tiny atmosphere and it is covered by ice (mainly N) which is relatively transparent. Therefore some radiation from Sun can arrive and absorbed while IR radiation is trapped by the upper layers. This constitutes a peculiar greenhouse effect at solid state.

We remember also that orbit of Pluto is one of the most elliptic. To be specific, it is located in the external part as result of the perturbation inside the system due to migration of giant planets like Jupiter and Saturn (maybe it was captured in the inner orbits).

Moreover it ha 5 satellites, including Charon, locked in resonant orbits. In particular, the ratio between size of Charon and Pluto is about 1/3, similar to Earth-Moon system. Various features indicate that Charon was formed by a big impact with an object with size comparable to satellites at which Pluto was subjected in the past at very low speed. This process remembers the formation of the Moon so studying Pluto, we can have information about our system.

4.9 Solar System formation

It is now believed that the stars are formed by the gravitational contraction of large clouds of gas of a sufficiently low temperature to allow the prevalence of the gravitational force over that of the internal gas pressure (**Jeans theory**). The process, once started (ex. from the passage of the density waves of the spiral arms of the Galaxy or from shock waves due to supernovae explosions), proceeds with increasing speed, supersonic, with the gas in free fall towards the center.

The equilibrium is reached when the density of the gas and its temperature are sufficient to oppose the gravitational force. A steady-state equilibrium situation is not completely reached, however, because the gas continues to lose energy by radiation and therefore the contraction continues, but in much larger scale times, through states that can be considered as quasi-stationary. A spherical central body of high temperature is thus formed that will give rise to the Sun, while in the external part the contraction leads to a high rotation speed (due to the conservation of the momentum) and therefore to a flat disk. From the current dispersion of the planets orbits with respect to the ecliptic plane, it is estimated that the thickness of the protoplanetary disk is of the order of one tenth (1/10) of an astronomical unit.

In the disk the average minimum density of the material to resist to dissipating forces, which are basically tidal forces coming from galactic field, is about $\rho \simeq 4 \cdot 10^{-9}\ g/cm^3$. Multiplying ρ with the volume where planets formed, we get the minimum mass of original Solar System that is about $M_{min} \sim 0.1 M_\odot$. Of course, some mass has been lost because, for example, H escaped during formation of planets due to the high temperatures and due to the fact the gravity was very low at the time. The actual mass collected into the planets is about $M_{act} \sim 0.001 M_\odot$, 1000 times less than the original one.

The same result comes from the chemical analysis of the planets and from the estimate of the lost mass of volatile elements during the planet formation. However there is no common agreement about the temperature of this disk.

Beckwith and Sargent have recently shown that many young stars from the region of Taurus and Ofiucus are surrounded by disks (detected in IR emission) with masses (measured in millimetric range) of the same order of magnitude. It follows that the process of forming a protolanetary disk must be very common during star formation.

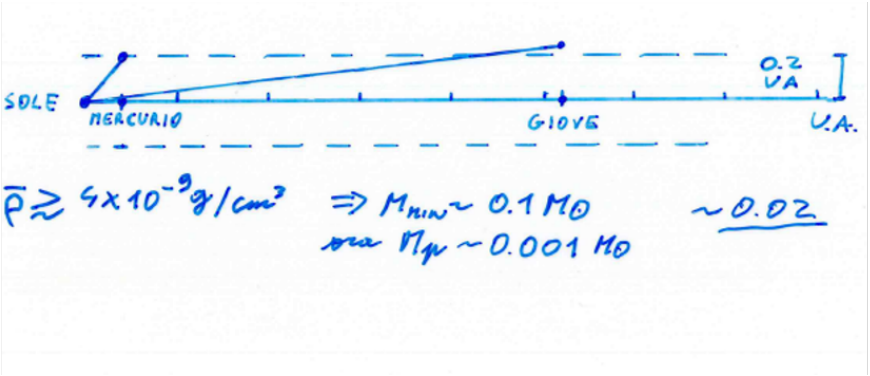


Figure 4.21: in the x-axis there is the distance in linear scale, every notch is one AU. the solid horizontal line represents the ecliptic plane, the dotted horizontal line represents the highest value of distance that Jupiter reach in its orbit, that is a bit tilted. The same value is reached also by Mercury because its orbit is more tilted but smaller. This shows that planet from from a disk of thickness of 0.2 AU.

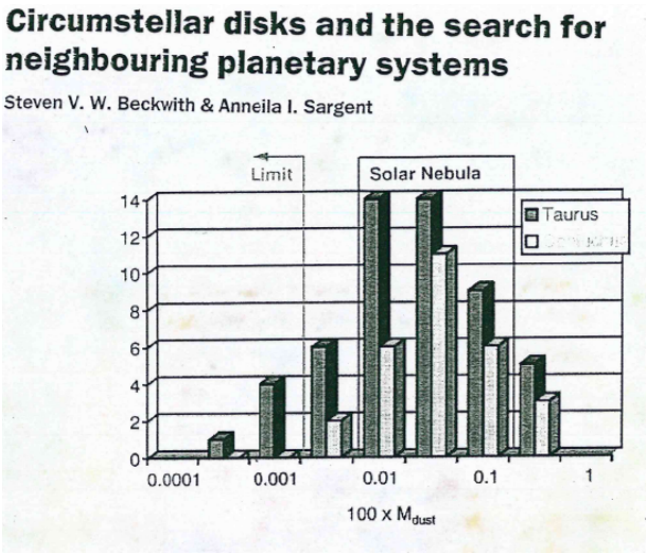


Figure 4.22: Steven at al. implement a way to derive the mass of the planetary disk thanks to the millimetric and radio photometry, that reveals the property of the dust. The detect of a young star with a disk is due the infrared photometry (see the previous lectures). The results of their work is this histogram. They find that the solar nebula's mass is ranged between 0.02 and 0.1, the same of the graphic 4.21.

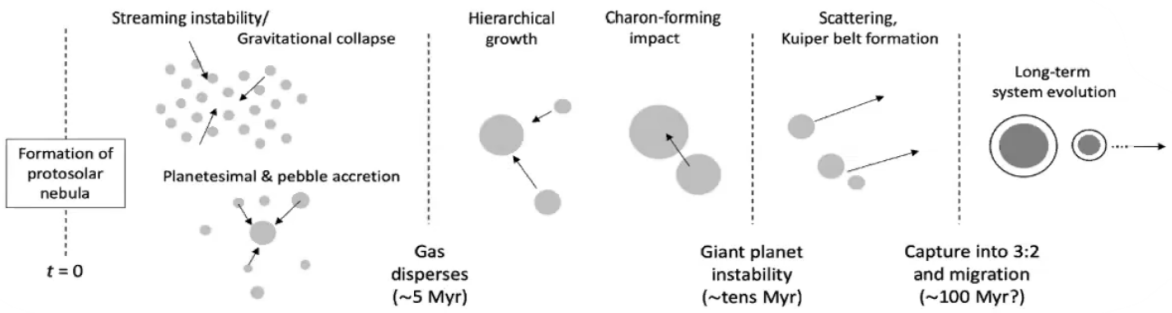


Figure 4.23: Phases of system formation

The temperature of the disc in its innermost part was higher than in the outer part. On the basis of the thermal equilibrium, between energy received from the Sun and lost by radiation, some authors have estimated a temperature between about 300 K up to 600 K for the gas that should have been 4.6 billion years ago at the distance of the Earth from the Sun. The uncertainty is due not only to the difficulty in estimating the opacity of the interplanetary medium, but also to the brightness of the Sun during the first evolutionary phases. The temperature on the outer regions of the disk, near the orbit of Pluto, would have instead reached a few degrees above the absolute zero.

4.9.1 Formation of planetesimal

The process that leads from the protoplanetary disk (represented schematically in figure 4.23) to the formation of the planets consists of three main phases:

1. **condensation** of the gas (dominant on the dust) contained in the disk into solid particles (grains) when gas cools down. At the beginning particles interact with Van Der Waals forces, growing to bigger ones of size about few millimeters;
2. **aggregation** of grains to solid particles up to the formation of planetesimals of size around one kilometer. At this size they interact with gravitational forces that are strongly enough to proceed with hierarchical growing;
3. **agglomeration** of planetesimals in planets.

In the end we have a process of intense solar wind that removes from the Solar System gas and residual particles, after 5 million years.

One of the most critical point in the formation of the Solar System is the aggregation of the grains up to the size of the planetesimals. Although the density of solid particles was high enough to result in a mean free path around the centimeter, it is necessary to postulate a physical mechanism such as to favor aggregation rather than destruction following impacts. This phase is still not completely clear today in its quantitative aspects.

Computer simulations show that a rotating disk of planetesimals tends to break in a series of rings and these in a series of clouds. Moreover, the growth of planetesimals until the formation of protoplanets requires special collision conditions. The speed of impact must be low, much lower than the escape velocity of the growing planet, and the materials must have cohesive properties.

Planets are mainly, not completely form via accretion of minor particle which is the accretion model, but we have two different models:

- the **completely accretion model**: just solid particle accreating and eventually with a thin layer of atmosphere captured by gravitational force;
- the other is an extreme case: very massive part of the disk could collapse due a gravitational instability and this contribute to the formation of the giant planets.

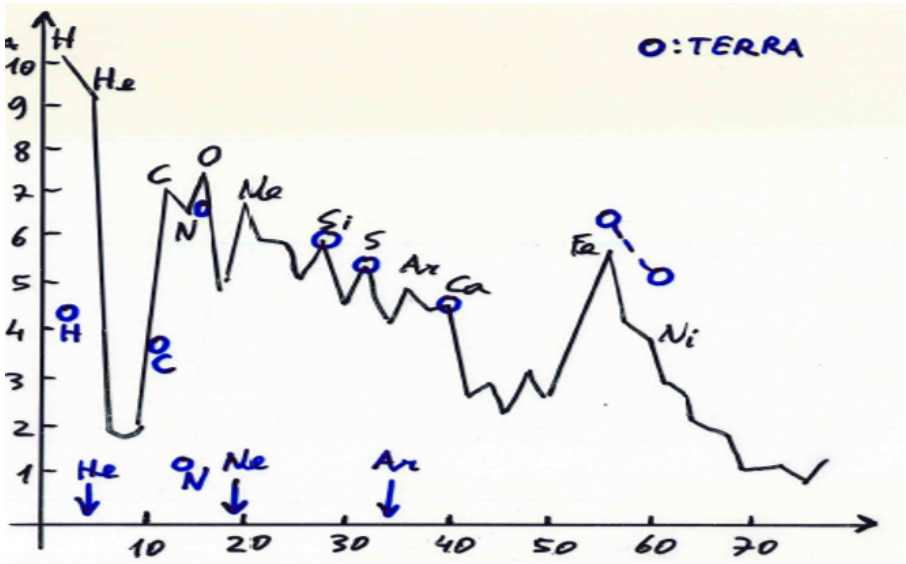


Figure 4.24: Abundances in the universe and on Earth.

However the key role in formation is the accretion. In particular accretion between grains of dust works with Van der Waals up to 1 cm size. Instead particles that arrive from an elliptical orbit with very high velocity, the impact can destroy the particles and the kinetic energy is transferred into thermal energy and then particles break up or evaporate. Moreover we don't know exactly how the process works from centimeter size to 1/2 meters. In this case the gravity is so low that it doesn't work so materials stick together with some forces that are not very obvious. We think that it should be a peculiar material, more porous and similar to snow in terms of sticky properties.

Element abundances on Earth

Figure 4.24 represents the atomic abundances versus the atomic mass. The black solid line is representing the universe, while the blue spheres are the abundances on Earth.

Universe In the universe there is a monotonic decrease from H until Ca, so from lighter to heavier elements, as result of thermonuclear reactions inside stars. Pay attention: the scale is a logarithmic scale so the jump from H to Si, for example, is very high. Then there is the iron peak, an anomaly produced by supernovae.

Earth Instead on Earth we see that there is o strong under abundance of H, up to 6 orders of magnitude. He is lower than in the universe for 11 orders of magnitude and there is a very low abundance also for Ne and Ar. This is because these atoms were present in the primary atmosphere that is lost: they escaped with the process we saw. In particular there is a low abundance of these element and not the other elements because the other ones were fixed in the planet, captured in complex compounds. Also C is less abundance than Si due to the composition of Sun that privilege Si than C (other stars instead are more abundance in C). Someone could expect that life could be based on Si, but it is base on C.

Urey and Lewis theory

Plot in figure 4.25 is a fundamental one to understand how Solar System has formed as we see today.

This is the **Urey-Lewis equilibrium condensation theory**. Urey and Lewis have listed the compounds that can condense (the inverse of sublimate) into solid particles at various distances from the Sun in the protoplanetary disk under physical-chemical equilibrium conditions.

The temperature considered on X-axis is the effective temperature after the condensation of the planets. It is also a sort of distance scale because temperature reflects distance from the Sun, so at

higher temperatures we are much closer to the Sun.

On Y-axis there is the pressure that changes in different part of the Solar System.

On the black line there are planets. We see that Earth at formation time had about 500 K, well above the boiling point of water indeed water evaporates at this distance. As the protoplanetary disk cooled down, at the position of inner planets only the most refractory compounds condensate. So we expect they are rich in these elements, like Fe, iron, and $CaTiO_3$, calcium titanate in the case of Mercury. Venus, in addiction to the previous ones, has also $NaAlSi_3O_3$. On Earth and Mars there are the previous ones, but also FeS, iron sulfur,very common in rocks which contains Fe and S in equal quantities. In particular Sulfur is a very stable element since the formation proceed slowly in chemical equilibrium. Also serpentine elements (rocks rich in magnesium and water) contain some water, present on Earth and Mars.

Going further away, up to 4 AU, Jupiter is the first which contains solid water plus all the molecules on the right side, so it has a inner core with these other molecules and ice water an ammonia NH_3 in the outer part. Going away from Jupiter, near Pluto we can find CH_4IDR (idrato).

Therefore at this distance, about Jupiter's orbit, condensation of lighter elements, such as water vapor, can occur. This explains different things:

- why Jupiter, Saturn, Uranus and Neptune are giant planets: at time formation they had a huge abundance of solid material after condensation;
- why inner planets have higher densities than the outer ones. The outer ones have the density of water, more or less, while the inner ones are intermediate between iron (about 7 g/cm^3) and aluminium (about 3 g/cm^3). Moreover there is a gradient density inside planets. So it is nothing to do with gravitational forces. The material is completely sprinkled so there is no hydrostatic equilibrium in the direction of the Sun;
- it explains abundances on planets.

To be more specific, it is evident that only the most refractory compounds such as iron oxides and titanium compounds can condense at the distance of Mercury, while in the terrestrial orbit they can also condense iron and metallic nickel in addition to silicates, feldspars (silicate oxide with Al, K, Na, Ca, Fe ...) and troilite (iron sulfide, FeS) where the last is the result of chemical reactions between the iron and the sulfur, still in the gas phase. The presence of troilite both on land and on many meteorites demonstrates the validity of the equilibrium condensation model.

It is important to note that water ice cannot condense in the Earth's orbit. This process can only take place at temperatures of about $170 - 200$ K, much further from the Sun, near the orbit of Jupiter, in a region called **the snow line** or **frost line**. This is a line over which ices, mainly water ice, are condensated. Today this limit is located between Mars and Jupiter orbiter (about 4 AU).

It is too difficult to understand where this line was located in the past. The temperature is the key to understand this. Indeed T was settled by different things:

- the thermal energy from Sun, from the gravitational contraction disk due to the Virial Theorem (compression of material and flattening of the disk);
- irradiation of heat outside the disk;
- irradiation of heat by the Sun. However it is difficult to estimate due to the absorption of the innermost part of the disk in the past given by big quantities of dust during planetary formation.

Nowadays we have some measurements from young protoplanetary disks and scientists think that in the past, during planetary formation, the snow line was settled about 2 AU. To be more specific, in the literature astronomers talk about two snow lines, one for water ice and one for CO_2, the second most abundant element at the time. However this secondary line was settled several times away from the first one.

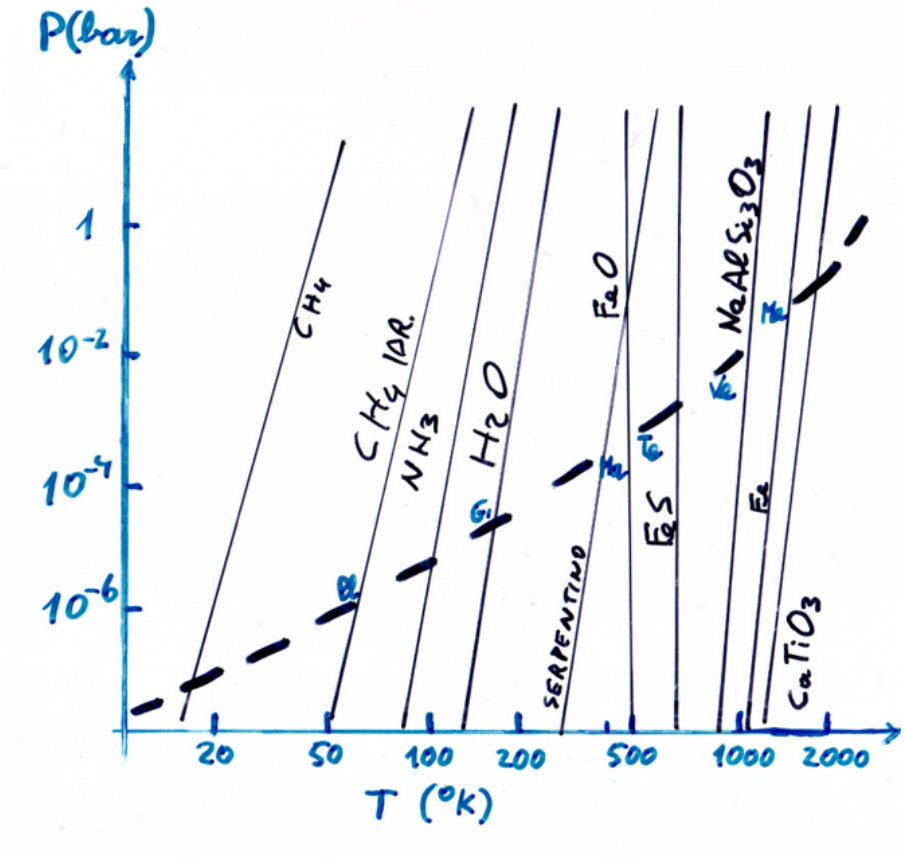


Figure 4.25: Urey and Lewis equilibrium theory.

A fundamental proof that the condensation model is correct, at least in first approximation, is the decreasing density of the planets of the Solar System as they are farther from the Sun. Iron oxides and metallic iron, in fact, have a specific weight higher than silicates, and their specific weights, in turn, are higher than those of water or methane ice.

Water on Earth Recent works that keep into account the opacity of the disk and of the outer "flaring", give the limit of condensation of the water around 2.5 astronomical units. Consequently the accumulation of water on the Earth can occur either indirectly from hydrated compounds of higher condensation point (such as serpentine) or from later processes of meteoric bombardment of bodies rich in ice like nuclei of comets (**late bombardment**).

Indeed, if we study the ratio Deuterium/Hydrogen on planets we can see that Earth has a value different from the other planets, but similar to comets. So, this implies that water on Earth comes from far objects as comets. This bombardment happens when the Earth surface was cooler enough to keep in the atmosphere the resulting water vapor.

4.9.2 Elimination of the residual cloud

Immediately after the formation of the planets, the solar nebula still contained a significant amount of dust and gas. This material must have been expelled, not only because today it is not observed but also because its presence in the long time would have led to the destruction of the planets themselves making them spiral towards the Sun by a braking mechanism. It is possible that this happened at first and that a not easily estimable number of planets has been incorporated into the Sun. Certainly the cloud has been cleaned up in a very distant era.

One of the proposed mechanisms is the solar wind during the so-called T Tauri phase. The stars in phase of T Tauri, with an age of around $10^6 - 10^7$ years, of similar mass to the solar one, introduce an intense stellar wind, with speed of the order of the $400 - 500$ km/sec and intensity around 10^8

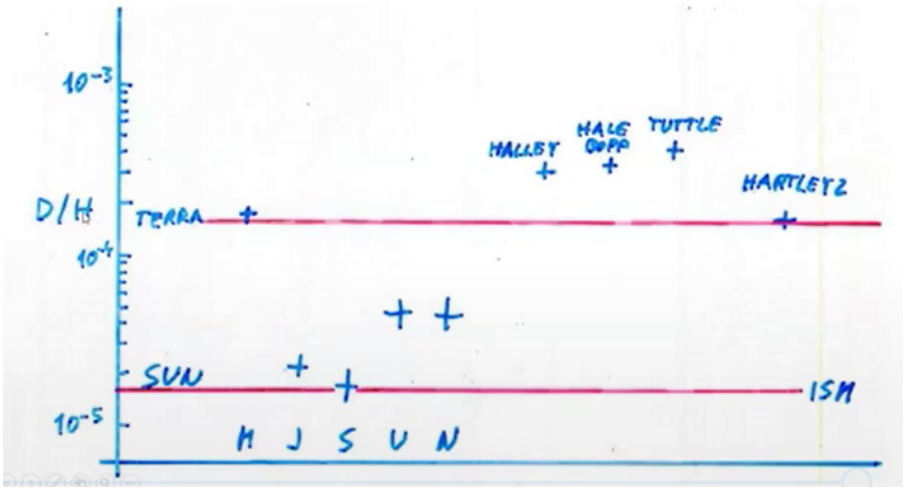


Figure 4.26: D/H ratio.

sometimes the solar one (which is currently around 10 particles per cm^{-3}). This wind is believed to be generated by the intense convection of the outer layers, and it is able to rapidly eliminate all the gas from a protoplanetary system, and also the particles with a diameter of less than ten centimeters. For bodies of higher dimensions, the attraction of the solar gravitational field prevails instead.

An important event in a system formation is the migration of the planets towards the Sun. This happens due the conservation of the angular momentum and due the presence of gas in the planetary plane. Assuming planetary migrations, Urey and Lewis equilibrium theory doesn't work exactly. In general a planetary system is a unstable system; if the gas is removal the system became a bit less unstable. Of course, the instability is due the interaction between all the bodies in the system.

Migrations of planets are described in the Nice model in the following subsection.

Nice model

The Nice model is a scenario for the dynamical evolution of the Solar System. It proposes the migration of the giant planets from an initial compact configuration into their present positions, long after the dissipation of the initial protoplanetary disk. In this way, it differs from earlier models of the Solar System's formation.

This planetary migration is used in dynamical simulations of the Solar System to explain historical events including the Late Heavy Bombardment of the inner Solar System (which allowed to deliver water to Earth by comets), the formation of the Oort cloud, and the existence of populations of small Solar System bodies such as the Kuiper belt, the Neptune and Jupiter trojans, and the numerous resonant trans-Neptunian objects dominated by Neptune.

Planets close to the Sun

We know that Mercury is located at 0.4 AU. Why don't we see another planet in the inner part of Mercury orbit, more close to the Sun? Indeed there are some refractory elements that can condensate also between 0.3 and 0.4 AU but here there are no planets. To explain this, it has been found from laboratory experiments that dust at 1400 C, or more, the properties are a bit different and at this temperature dust is not sticky enough to grow and produce a new planet. However it is very strange because we know many exoplanetary systems were there are planets very close to their star, also closer than 0.4 AU.

4.9.3 Evolution of a planets: focus on Earth

After the formation of the planets, there had been a bombardment. It is very intensive due the high number of residuals in the system. The bombardment produces a degassing and the surface temperature on Earth was of the order of 1500 C. The surface was total melted indeed it was a magma ocean. So the condition were: very high temperature and new gas liberated in the atmosphere (mainly CO_2). At the same time the original gases were collected by gravitational forces on Earth, mainly H and He. So the originally atmosphere was composed by H, He, noble gases, CO_2. In particular the planet was very rich in H and He. However, these light gases were then removed by Jeans escape due to impacts or due to the very high temperature. The Earth remained spoil by primary atmosphere. Then a secondary atmosphere was generated by other impact and volcano eruptions, when the surface started to cooling down, by radiation.

As seen in the previous subsection, an important problem, and not easy to solve, is the origin of water on Earth, given that the condensation temperature was much lower than what was in the disk at the distance of the Earth.

Collisions with planetesimals coming from the external part, at least in the initial phase, would be excluded since the elliptical orbits would lead to too high impact velocities to hold the ice.

One possibility is that the water has come indirectly through rocks containing hydrates, which are present in carbonaceous chondrites. A proof is that the Deuterium/Hydrogen ratio in the seas of the Earth (10^{-4}) is similar to that of the chondrites and slightly less than that of the comets, while it is clearly greater than the ratio measured in the atmosphere of Jupiter and in the Sun (about $2 \cdot 10^{-5}$).

According to the model of Nice, instead, shortly after its formation, after the most volatile compounds of the solar nebula are condensed, the Earth would have been intensely bombarded by planetesimals similar to cometary nuclei rich in ice, in an episode called the "late bombardment". This occurred following the instability of the orbit of Jupiter and Saturn, around 200 million years after the formation of the planets. This process of intense late bombardment could have lasted about half a billion years and would have played a decisive role in enriching the Earth with water.

So, with the arrival of the water, the Urey reaction started and fixed the CO_2, depleting the atmosphere of the greenhouse molecules. At the same time ammonia was decomposed by UV radiation of the Sun into nitrogen, enriching the atmosphere, and hydrogen, which was lost. A minor presence of greenhouse molecules in the atmosphere implies that the temperature started to cooling down.

A dramatic decreasing of the temperature occurred when the oceans were formed. The Earth entered in the **snowball phase**, in which Earth was covered by ice. The albedo dramatic increased. So, it is difficult to exit from this phase. It is not reversible. In addiction, at that time the Sun was less luminosity (now it is more brighter of 40%), a reason more to think that Earth entered this phase.

How the Earth exit from this phase is a mystery. We think that very strong and efficient volcanic eruptions increased the CO_2 in the atmosphere and, due the greenhouse effect, the Earth left this phase. However, it is difficult to understand how climate on Earth has been so stable, at least in the last 2 $Gyrs$, in spite of minor luminosity of the young Sun at that time. This problem is called **the young Sun paradox**.

Source of heat and the differentiation What is certain is that planets larger than a hundred kilometers collect enough material to produce a considerable heating of the rocks due to the decay of radioactive isotopes.

The most relevant isotopes, in this phase, are ^{235}U (which produces $4.6 \cdot 10^{-12}$ W/kg of rocky material with half-lives of $7 \cdot 10^8$ years), ^{40}K (which produces $3.7 \cdot 10^{-11}$ W/kg of rocky material with half-lives of 10^9 years) and thorium.

At the time there was another one, the Aluminium, which is very abundant in the rocks. In particular ^{26}Al decays into ^{26}Mg, which is very stable. The isotope ^{26}Al has the fastest decay time (about

$7 \cdot 10^5$ years) and produces by far more energy in a short time (4 or 5 order of magnitudes more than other elements), therefore it is responsible for the fusion process in the first evolutionary phases of the terrestrial planets. But where does ^{26}Al come from? It is the result of capture of protons by Ca during nuclear reactions, mainly in Novae and sometimes in Supernovae. However, as we seen, it decays in a very shot time, not enough to reach the Solar System. This means that our planetary system has been formed nearby or as product of the explosion of a Nova or SN.

Therefore, heating can melt the inner rocks of the planet and maintain the state of fusion for a period sufficient to allow the differentiation of the material, i.e. the gravitational separation between heavy materials (iron and nickel) towards the interior and light materials, such as silicates, on the surface.

The condition for differentiation is expressed by the relationship between the surface temperatures $T(R)$ and the central temperature $T(0)$:

$$T(0) = T(R) + (a/6c)R^2 \tag{4.30}$$

Where a is the production of energy per unit of volume and c the thermal conductivity. In order for differentiation to occur, the difference in temperatures must exceed 1800 K for rocky bodies composed of silicates and about 150 K in the case of ice. The calculation with the typical values of the planets gives a radius, in both cases, of a few kilometers. It is therefore concluded that the process of differentiation was very common both between the planets and between the asteroids. This can explain why we found some iron meteorites "older" than the Solar System. We can understand this only if they are completely melted and differentiated in a very short time. Only smaller bodies may have retained the original composition.

The result of differentiation is that the superficial part of the Earth is constituted by a thin crust with density around 2.7 g/cm^3 (continental plates) that float on the denser layers of the mantle (3.3 g/cm^3). Then the core as a density about 7 g/cm^3. Making an average on all the planet, we obtain an average density of 5.5 g/cm^3. One of the proof that this theory is correct is that on Earth the maximum height reachable from underwater is about 12/14 km. Instead on Mars it is about 20 km. The movements of the continental plates can lead to collapses of the crust inside the mantle and to lifts due to the hydrostatic force. So a mountain (of density about 2.7 g/cm^3) is floating on a fluid of density about 3.3 g/cm^3. The crust features cannot sink more than 85% of its thickness since at the depth of about 70 km they find such a temperature to melt with the mantle. For this reason the terrestrial reliefs reach maximum elevations of about 12 km, corresponding to $0.15/0.85 = 17\%$ of 70 km.

Unstable systems

It is important to know that planetary systems are the most unstable systems in space. How about the future of Solar System?

In one of Batygin and Laughlin's simulations, Mercury was throwed into the Sun in 1.3 billion years from now. In another, Mars was flung out of the Solar System after 820 million years later than Mercury and Venus collide. So dynamical changes occur in a very short time compared to Sun evolution.

4.9.4 Dating the surface of terrestrial planets

Dating planetary surface is fundamental in order to understand planetary evolution. The dating of the surface through the counting of meteoric craters represents a very important technique given the variety and uniqueness of the surface structures: the continuous meteoric bombardment leads to a progressive increase of craters on more and more ancient surfaces.

The meteoric impact on a planetary surface forms a crater with dimensions that can be expressed empirically according to the kinetic energy of the body that causes the impact:

$$D = (498.7dr^3v^2)^{1/3} \tag{4.31}$$

where D the diameter is expressed in km, the speed v in km/s, d the density in g/cm^3. 498.7 is a constant to normalize the relation. Usually big craters are rare and surface is more uniform going to smaller diameters.

In the absence of atmosphere or other erosive agents meteoric craters grow in number with time and therefore their surface density increases.

In older lands, the falling of meteorites can lead to the limit such a density of craters to destroy the oldest ones and, therefore, their number can no longer grow over time. In this situation the ground is said to be **saturated**. In the Solar System, for planets without erosion sources, near the Earth, saturation occurs in about 4.3 billion years.

The method of dating through the density of the craters is based simply on the assumption of a progressive increase in density of the craters with age, up to the limit of saturation. The calibration of the production rate of the craters with the age has been fixed definitively through the counts on lunar terrains of known age by dating the rock samples brought on Earth by Apollo missions.

It is also noted that the number of craters is inversely proportional to the square of their diameter:

$$N = cost.D^{-2} \tag{4.32}$$

In a logarithmic diagram, figure 4.27, where the density of the craters is reported as a function of their diameter, this law appears as distinct lines according to age and decreasing with the diameter. At increasing age the straight line moves upwards following an increase in craterization. The dating is done by comparing the data with a sample diagram.

Obviously in a planet with atmosphere and erosive agents, for example on Earth, we must take into account the destruction of the smallest craters over time (giving a more a curved shape to the graphic in figure 4.27). The opposite effect is observed instead on the Moon where there is an increase of the small craters due to the production of secondary impacts from material expelled by large impacts that falls around the primary crater (in figure 4.27, this effect is shown as a deviation in the upper part of the straight line).

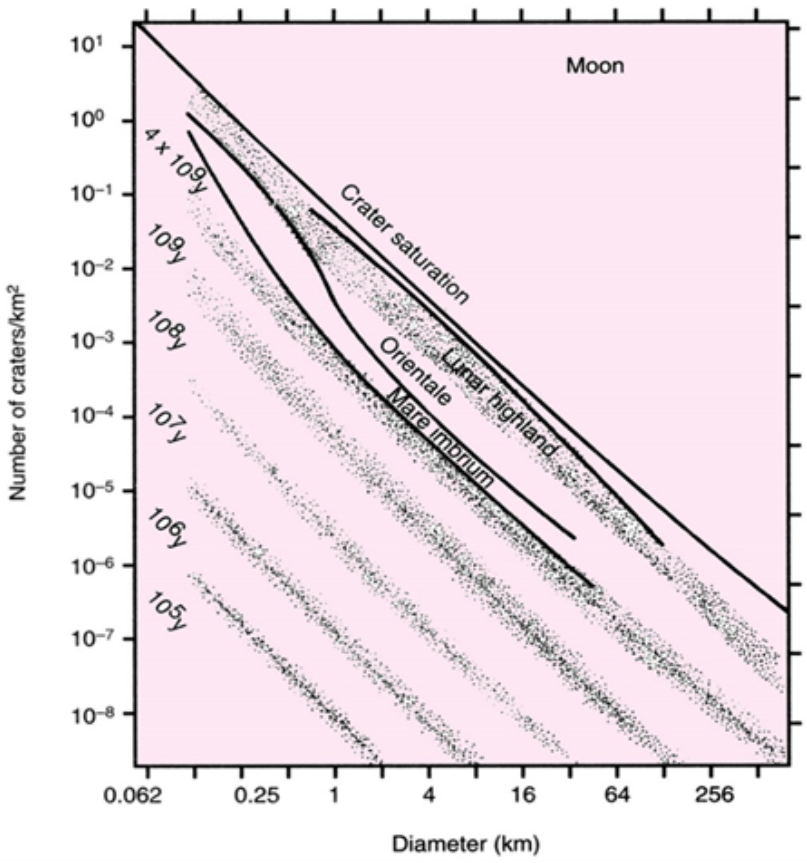

Moon
Crater saturation
Lunar highland
Orientale
Mare imbrium
4 x 10⁹y
10⁹y
10⁸y
10⁷y
10⁶y
10⁵y
Number of craters/km²
Diameter (km)
0.062
0.25
1
4
16
64
256

Figure 4.27

Chapter 5

Exoplanet

An exoplanet or extrasolar planet is a planet outside the Solar System. The first confirmation of an exoplanet orbiting a main-sequence star was made in 1995, when a giant planet was found in a four-day orbit around the nearby star 51 Pegasi.

5.1 Extra solar planets: research techniques statistics

There are different ways to discover an exoplanet.

- Direct imaging
- Astrometric perturbation
- Radial velocity
- Photometric eclipses: the transit method
- Radio detection
- Pulsation
- Radar analysis

There are also some minor ways. However we will treat only the first four method. The others are less used and less common.

Every technique allows to know specific parameters about the planetary system or about the planet but every method has also some biases, a king of selection of what kind of planets we can detect, due to the intrinsic nature of a specific method.

5.1.1 Direct imaging

The discovery of extrasolar planets through direct images is the most obvious way but also a very difficult one due to the difference in brightness between the star and the planet, so the issue is not about the separation from the star and the planet.

Suppose to observe a planet like Jupiter, which is faraway 5 AU from the star, around a solar type star at a distance of 5 pc from the observer, like in figure 5.1.

By definition of the astronomical unit, at 1 pc you see 1 AU under an angle of 1 arcsecond. Therefore at a distance of 5 pc, a separation of 5 AU is seen as an angle of 1 $arcsecond$. Nowadays, with modern telescopes, we can reach a separation also of 0.5 $arcsecond$ so, as said before, the separation is not the problem. The real issue is the difference in brightness between star and planet. A star like the Sun at a distance of 5 pc has a visible magnitude of $V_\odot = 3.4$ while Jupiter $V_J = 24.6$. Therefore the difference in luminosity between a planet like Jupiter and a star like the Sun is 21.3 magnitudes,

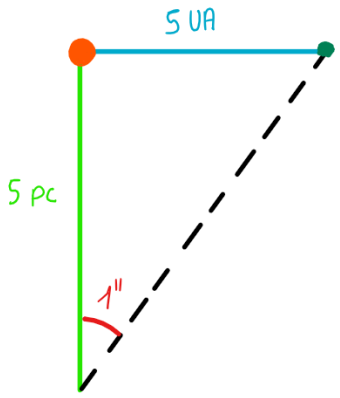


Figure 5.1: Direct imaging

about 10^{-8} ratio because every 5 magnitude, it increases of a factor 100. This means that, considering a brightness profile, we have to consider values faraway in the tail.

Instead, considering a planet like the Earth in this planetary system, its visual magnitude at this distance is $V_\oplus = 26.4$ which means another factor 100 in terms of luminosity. Therefore it is more difficult to discover planets of Earth-size.

However actually there is no technique for the direct detection of a planet with these characteristics. Direct detection has recently been obtained on some planets where the separation is wider (e.g. 2M1207b has a separation of 54 AU) and the star/planet luminosity ratio is much lower (about 100).

The problem can be partially resolved using systems in IR range, where young planets emit a lot of radiation due to their internal heat, or radio range where the stars like the Sun are much fainter than the planet.

Obtainable parameters From direct images we can obtain:

- brightness of the planet;
- eventually the distance of the planet from the star, if you know the distance of the star from the Sun. However we don't know exactly where the planet is along its orbit so the distance we measure is $\leq a$, with a the real axis of planet's orbit;
- eventually we can get also the period but it is very hard because there is a strong bias on large a because to observe the planet, it must be very distance from the star.

Therefore you don't have many outputs from direct imaging. It is suitable eventually for spectroscopy but nothing more.

Bias We detect only planet with a large distance from its star, at 30/40 AU, like Pluto. The period is about hundred years so it is almost impossible to follow the orbit and this is the reason why a is not very well determinate.

5.1.2 Astrometric perturbation

This technique is based on the astrometric measurements of the motion of the star around the common star-planet center of gravity. In figure 5.2, is represented a simple framework of the system.

For classical mechanics, the equation that describes equilibrium around the barycenter is:

$$M_s a_s = M_p a_p \tag{5.1}$$

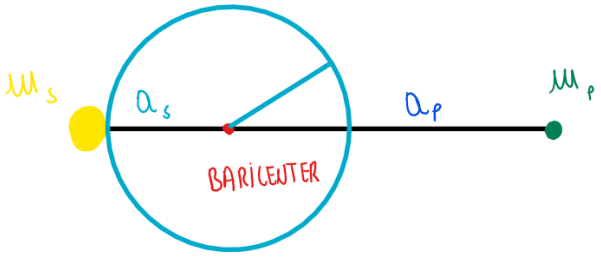


Figure 5.2: Astrometric perturbation

where M_p and a_p are mass and distance from the center of the planet and M_s, as those relative to the star.

- The mass of the star M_s is dominating (a very large number compared to the mass of planet) and it can also be easily deduced by other methods, like stellar models.
- Observing the motion of the star around the barycenter position, it can be deduce the radius a_s, that it is usually a very small number.
- Observing the star moving around the barycenter, following its path it is possible to get the period P

Therefore, once known M_s and the period P, from the third law of Kepler, we also have a_p:

$$P^2 = \frac{4\pi^2}{G(M_s + M_p)} a_p^3 \tag{5.2}$$

where M_p is eventually negligible.

So, in equation 5.1, the only unknown is M_p, one of the most important parameter.

Obtainable parameters In summary, from the astrometric method we have the following parameters:

- distance from the center of gravity of the star a_s;
- period of rotation P around the center of gravity (of the star and therefore also of the planet);
- distance of the planet from the center of gravity a_p (which is practically the distance of the planet from the star);
- mass of the planet M_p.

What is important to understand is that we can get the mass of the planet because the motion of the star is driven by a gravitational effect.

Bias The discoveries are based on amplitude we can measure. The amplitude of the astrometric orbit depends on two factors, on the distance of the planet from the barycenter (which means from the star) and n the mass of the planet:

$$a_s = \frac{a_p M_p}{M_s} \tag{5.3}$$

Therefore the selection favors the discovery of planets far from the star (wide separation) and of great mass, and/or orbiting planets around stars of small mass (the radius of the star or planet are not relevant). Note, however, that greater distances on the planet from the star imply longer rotation periods, hence, as discussed above, prolonged observations over time.

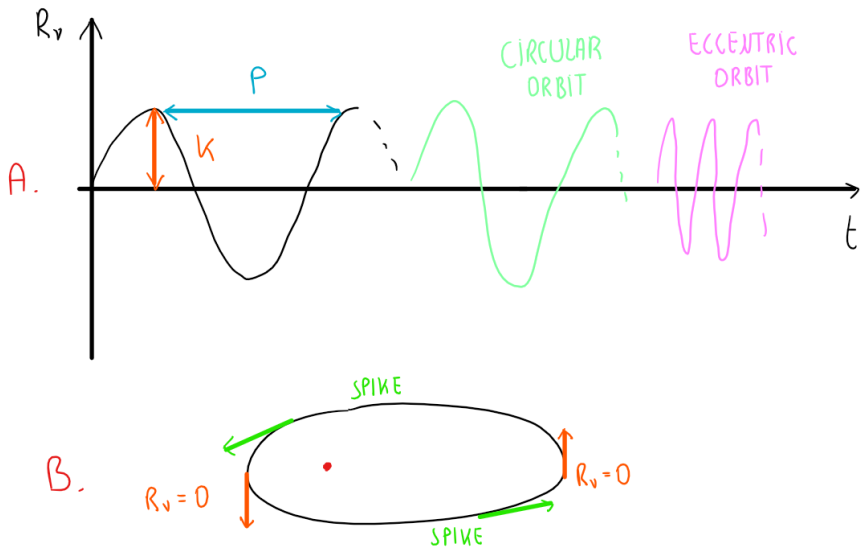


Figure 5.3: Radial velocity

This is the oldest method that has been used however they never discover a planet in this way because, as said before, it is necessary a very high accuracy. In particular it depends on a_s that is a projection on the sky as angular size so this method can be used only fro nearby stars.

For example, for a planet like Jupiter at 5 Kpc, a_p is about only 1 $milliarcosecond$ (1 mas) which is 1000 $\mu arcsecond$ while from Earth we can only get some mas. In the case of Gaia, the accuracy is about 20 $\mu arcsecond$ and this will allow the discovery of new planets. So it is a promising technique for the future.

To be more specific, astrometry and radial velocity have something in common. In the case of radial velocity, we measure the motion along the line of sight while in the case of astrometry we measure the motion tangentially to the plane but there is also a big difference. Even thought the physics is the same and then we will get the mass of the planet, in the case of radial velocity the amplitude of the variation is not dependent on the distance so the detection can be done as far as we want. Instead, in the case of astrometry, the separation rapidly decreases with the distance, in a linear way. More details about radial velocity are given in the next subsection.

5.1.3 Radial velocity

Almost 1000 exoplanets known to date have been discovered, or studied, through the study of relative star motion around the common center of gravity, along the line of sight (radial velocity), from the Doppler effect. In particular, measurements of radial velocity of the star as function of time give a typical diagram such as figure 5.3, part A.

It is known from classical astronomy that the radial velocity, along the line of sight of the observer, of a mass star M_s, due to a companion of mass $M_p \sin i$, with orbital period P, inclination of orbit i and eccentricity of orbit e, is given by:

$$K = \left(\frac{2\pi G}{P}\right)^{1/3} \frac{M_p \sin i}{(M_s + M_p)^{2/3}} \frac{1}{(1-e^2)^{1/2}} \tag{5.4}$$

- K is the half amplitude of the graph (V_r, t) so it is known from observations of radial velocity as function of time.
- Period P is known from the graphic as distance in time between two peaks. Indeed, as visible in

5.3 part B, radial velocity is much higher when planet approaches the star then there is radial velocity equal to zero and then another spike up to the opposite position where the radial velocity is zero again and so on.

- The eccentricity is derived from the analysis of the radial velocity curve. As visible in figure 5.3 part A, from the shape of the graphic, more compressed or not, it is possible to derive an estimate of eccentricity of the orbit.
- Mass of the star is known from isochrones and/or the spectrum of the star. Usually they have masses similar to the Sun, eventually 4 or 5 $M_\odot$, however for these last one we can't measure radial velocity because they are too massive to be moved around the barycenter by gravitational influence of the planet.
- The term $(M_s + M_p)^{2/3}$ is known due to the fact that M_s is usually much more bigger than the planet one so M_p is negligible.

Therefore, knowing K from observations, $M_p \sin i$ can be obtained. In the case of Jupiter the value of K is 12.6 m/s, while the Earth would give only 0.09 m/s. The best spectrographs today reach a few meters per second. It is believed that the planets can be revealed up to a value of K around one m/s, below which the motions of the stellar photosphere can hide the effects of motion around the common center of gravity.

The inclination i is the angle between the direction of observations and the perpendicular to the orbit. Of course, when $i = 0°$, this means $\sin i = 0$ so the equation goes to zero. It is obvious, because when the orbit is perpendicular to the line of sight, we have any radial velocity component. On the contrary, when the plane of orbit approaches the direction of the observer, $i = 90°$ which means $\sin i = 1$, then the amplitude is maximum and you have more chances to detect a planet: higher is K, higher is the probability to detect a planet and higher is the accuracy you get on measurement of K.

In particular, since $\sin i$ goes from 0 to 1, the mass M_p we get is lower than the real one due to this degeneracy. To remove this degeneracy, we have to know the inclination i; to do this there are some tricks.

- If we can have simultaneously the detection of radial velocity as function of time and also a transit, this means that the inclination $i = 90°$ so $\sin i = 1$ (best condition obtainable).
- Let's think about Solar System. In our planetary system planet's orbits are almost perpendicular to the spin axis of the Sun (there is a tilt of only 5°. So the issue is: it is necessary to find the spin axis orientation relatively to our line of sight of the star and then we can assume that orbits are perpendicular to the spin axis. How to get the spin axis? From the spectrum we can study the Doppler broadening due to the rotation which is $V_{rot} \sin i$. We don't have V_{rot} but it is easy to get. If the rotational period and the radius of the star are known then we get the real tangential velocity V_{rot}. Therefore, knowing the velocity, i is known. In particular from standard models we can get the radius while observing stellar spots we get the rotational period.

Obtainable parameters The most important parameter we can get by this method is, as seen before, the mass of the planet M_p. From radial velocity graphic we can also obtain the period P, and putting P and M_p inside the third Kepler law, we can get also the axis a.

Bias

- The K decreases with the increase of the period P, that is, from the third law of Kepler, linked to the orbital radius. It is deduced that the discovery of planets close to the star is favored.
- The K increases linearly with the mass of the planet, but depends on the projection of the orbit along the line of sight.
- The K also increases with the eccentricity of the orbit e.

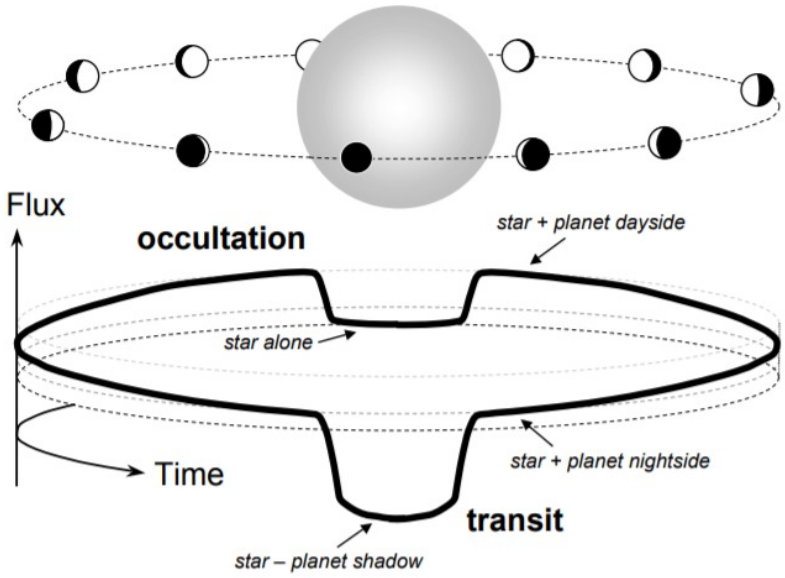


Figure 5.4

- Finally the K decreases as the mass of the star increases.

In conclusion the method favors the discovery of planets of great mass and high eccentricity of the orbit, orbiting close to the star with short periods (e.g. the hot Jupiters), possibly around small mass stars.

The case of simultaneous observability of radial velocity curves and transits (photometric eclipses) is important, because in this case we have both the mass and the radius of the planet and therefore its density.

Finally, note that planets with "face on" orbits, that is, with $V \sin i = 0$ do not give observable radial velocities and therefore escape this type of investigation. In this case they can be observed only with the astrometric technique or by direct observations.

It should be noted that the accuracy of the radial velocities (hence of the K) depends on the stability of the spectrograph, as well as on the signal/noise ratio, not necessarily on the size or quality of the telescope. However, the main problem on accuracy is the degeneracy due to $\sin i$.

Of course, there can be some false positives: effects of binary stars can be mixed up with a planetary system, while for multiples systems it is necessary go thought models.

5.1.4 Photometric eclipses: the transit method

An alternative method of detection is based on the observation of the transit of planets in front of the star (figure 5.4). The variation of brightness of the star is proportional to the ratios of the projected areas, i.e. to the squares of the radii:

$$\Delta I = \frac{R_p^2}{R_s^2} \tag{5.5}$$

If the radius of the star is known, the radius of the planet is immediately obtained from the depth of the measured transit translated into a difference in intensity ΔI. From the repetition of the transits we also obtain the period P and, if the mass of the star is known for other ways, from the third law of Kepler, we have the distance of the planet from the star.

There can be also an occultation, with a minor depth, when the planet passes on the other side. In this case the total intensity is only given by the intensity of the star, and not the sum of star and

planet intensities.

In the event that the transit could be observed due to the transit of a planet like Jupiter in front of a solar-type star, there would be a variation of ΔI of the star's intensity of 1%. In the case of Earth transit, instead, we would have about 300 times less, corresponding to a depth of ~ 70 *ppm* which means $70 \cdot 10^{-6}$. However this last measurement can't be done on the ground, due to several issues.

- The limitation is due to the statistics of the photons (poissonian noise) and to the atmospheric scintillation. This las one is the main issue. It depends on wavelength, pupil size and distance of turbulent layers of the atmosphere from the pupil. The latter is given by (Dravins et al., 1998):

$$\sigma = 0.09 D^{2/3} (\sec Z)^{1.75} \exp -h/h_0 (2T)^{1/2} \tag{5.6}$$

  where D is the diameter of the telescope in centimeters, h the height of the observer, h_0 the atmospheric scale height (the standard is about 8000 m), T the integration time in seconds, $\sec Z$ is the airmass while Z is the zenith distance. Eventually the scintillation can be smoothed down by increasing observing time however, exposing for very long time we can't observe the transit. So we have to expose for limited time, about some minutes (indeed transits usually last for some hours). We can re-write the equation above in a simple way:

$$\sigma = \frac{\Delta I}{I} = \frac{0.06 D}{\sqrt{T}} \tag{5.7}$$

  There are also some techniques to reduce scintillation using some sparse pupils connected each other however on the ground it remains the main problem.

Therefore in order to detect a planet like the Earth rotating around a star like the Sun, we have to go into space to make observations.

Moreover to observe a transit, it is necessary to observe a specific geometry condition. Indeed we can observe a transit only if the planetary system is on the line of sight. In particular there is a tilt condition on the inclination of the planet orbit. So preferentially we will discover planet with very low tilting angle. However, supposing to have a wide tilting, we can miss the transit but if the distance from the star is small, we can still see the transit. So preferentially we will discover planets close to the star.

The probability of observing the transits is, however, very low because it requires that the tangent of the angle of inclination between the planet's orbit and the direction of observation is less than the ratio between the radius of the star and the radius of the planet's orbit around the star.

$$\tan \alpha \leq \frac{R_s}{a_P} \tag{5.8}$$

It is clear that this can occur only for very small angles.

Of course, in this case the mass of the planet is irrelevant due to the fact that photometry doesn't include any gravitational effect.

Obtainable parameters The method allows to get the period P, obtainable measuring the time interval between two transits and the radius of the planet, from equation 5.5. Then, using the third Kepler law, we can get also the orbital axis a.

Of course, knowing the radius from the transit and the mass from radial velocity, we can get the density, a fundamental parameter to investigate the nature of the planet.

Bias

- As seen before, the probability of observing eclipses is very small because it requires a stringent inclination angle condition of the planet's orbital plane along the line of sight ($tan\alpha < R_s/a_p$). Furthermore it requires continuous and prolonged observations to catch the time of transit which is of very short duration compared to the orbital period. In the case of the Earth in front of the Sun, we have about $t = 2R_s/v_t = 1.400.000/40 = 35.000s$, that is almost ten hours, compared to a whole year. This is a probability of almost 1 over 900 ($\frac{10}{24} \cdot 365$). So planets transiting near the star are favored because of their shorter periods compared to the transit times.
- The radius of the star acts in two opposite directions. Larger radius gives greater (linearly) the probability of transit, but shallower depth of the eclipse, with quadratic law. So for relatively small planets where the signal to noise ratio is important, the radius of the star must be as small as possible. If we then consider that in the main sequence the radius is a function of mass, taking into account the third law of Kepler, we deduce that small mass stars and therefore small radii are favored.
- The continuity of the observations and the duration of the same are very important in order not to miss the "rare" transit events.
- The instrumentation plays an important role because the signal-to-noise ratio increases with the telescope diameter both because it improves the photon statistics and because, from observations from the ground, the effect of atmospheric scintillation, which contributes to noise, decreases. The duration of the exposure time can compensate for the telescope diameter, but it cannot be too long not to compromise the sampling during the eclipse.

Summarising, this method introduce a selection, as visible in equation 5.5, for smaller stars and bigger planets with small tilt in orbit.

Of course there can be some false positives: suppose to observe a bright star with a periodicity (maybe it is in blend with a faint star on the background so it is a binary system). To understand if it is a double system of stars or a planetary system we need confirmation from radial velocity method.

Chapter 6

Supernovae remnants

Since the first investigations into the radio emissions of our galaxy has been found that the radiation comes from an intense component concentrated on galactic plane, partly widespread, partly made up of numerous discrete sources with an angular diameter less than one degree. Some of these have a thermal spectrum and have been identified as HII regions, while others, with non-thermal spectra, have been interpreted as the remnants of galactic supernovae.

The association between supernova remnants and radio sources was initially suggested by Bolton, Stanley and Slee in 1949 for the Tau A source coinciding with the Crab Nebula, the supernova remnant observed by the Chinese and Japanese in 1054. Indeed they found it was expanding at very high velocity, it was rich in heavier elements than H and it had a peculiar shape and internal structure. Later works confirmed the close association between supernova remains and non-thermal sources showing that even the strong Cas A source was located in a filament rich nebula with typical properties of supernova remnants. Further evidence came from the identification of radio-sources with the supernovae regions of Tycho (1572) and Kepler (1604). From the examination of ancient astronomical documents Minkowski concluded that two other supernovae have been observed, in addition to the other three cited, in the last 3000 years: the supernovae of 185 and 1006. It can be assumed that the number of galactic supernovae actually observed in historical times is slightly less than ten. They are reported in the following timetable.

Modern observations on supernovae come from the study of extragalactic objects. So far, thousands supernovae have been cataloged with spectroscopic and/or photometric data in the modern era. These are traditionally classified (old, simplified classification) into two main categories based mainly on spectral characteristics. While in the type I supernovae spectra (SN) the hydrogen lines do not appear, in the type II SNs the lines of the Balmer series are identified, in particular the $H\alpha$.

Date	Name	Mag	Time of visibility
185	Centaurus	V=-8	20 months
393	Scorpius	V=-1	8 months
1006	Lupus	V=-8, -10	years (Ia)
1054	Taurus (Crab)	V=-6	22 months (II)
1161	Cassiopeia	V=0	6 months (II)
1572	Cassiopeia (Tycho)	V=-4	18 months (Ia)
1604	Ophiucus (Kepler)	V=-2.5	12 months
1680 (?)	Cassiopeia	V=0 (?)	? (II)

Table 6.1: Historical supernova remnants in the last 2000 years

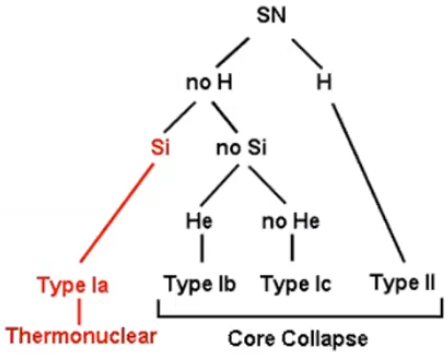


Figure 6.1: SN classification.

Type I

Origin: explosion SNI comes from a complex mechanism inside a binary system, usually composed by an evolved giant star (about 3 or 4 $M_\odot$) and a dwarf white. The process implies transfer of mass from the giant star to the white dwarf until the white dwarf reaches a critical mass and it explodes in a thermonuclear explosion with generation of heavier elements up to the mass of iron and also more. The ejection is rich in all elements, including Fe, while it travels at a speed of 10000 km/s.

Features This type of supernovae has a light curve with a period of about 50 days around the maximum, followed by an exponential decrease with a time scale of around $60 - 80$ days. The color index is around $B - V = 0.5 - 0.7$.

The optical spectrum is characterized by wide emission bands with radial velocities around $10.000 - 20.000\ km/s$. Type I supernovae are observed in all types of galaxies, including elliptical ones. For this reason they are associated with stars of intermediate mass of population II. However, these supernovae are also observed in the discs of spiral galaxies. It is believed that the supernovae of Tycho and Kepler belong to this category. Recent studies suggest a further subdivision of this class into SNIa and SNIb depending on the presence or absence of SiII line at 6150 $\mathring{A}$ in absorption. The SNIb are systematically weaker than the SNIa and therefore closer to the type II SNs.

Type II

Origin: implosion The SNII came from intermediate stars with masses of the order of $\sim 8 - 9M_\odot$, which generate all the heavier elements than H in thermonuclear reaction still iron. This SN accours thank to a collapse of the iron core when it reaches a mass about Chandrasaker limit, forming a neutron star with $\sim 1.4M_\odot$. Usually, core size of massive stars goes, more or less, from 1 $R_\odot$ (700.000 km) to 10000 km, leaving, after the collapse, a neutron star of about 8 km of radius. Therefore there is a huge release of gravitational energy. In particular the collapse is translated into thermal energy and kinetic energy ejecting the outer layers. As consequence, iron is captured inside the neutron star and transformed into neutrons so the ejection is made up by lighter elements. Moreover, the implosion triggers a shock-wave with velocity of the order of 5000 km/s.

For their origin they are also called **core-collapse supernovae**.

Features These supernovae have different light curves from one another. The classification is based mainly on the characteristics of the spectrum that shows, at most, a strong continuous blue and weak structures identified as H Balmer lines. The expansion speed is around 5000 km/s. Type II supernovae are characteristic of spiral galaxies and are associated with the population I. Further subdivisions have also been suggested for this class (SNIIn, P, L).

Statistics based on extragalactic supernovae indicate that type I supernovae are about twice those

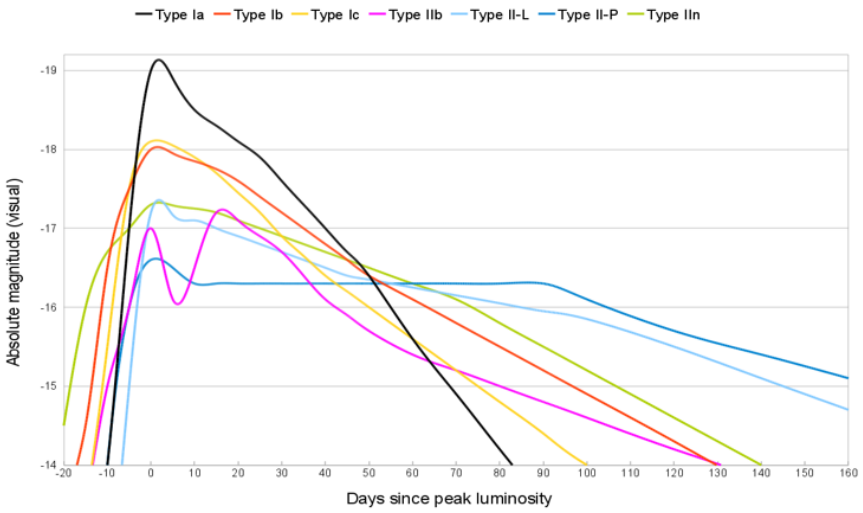


Figure 6.2: Spectrum of two types of supernovae

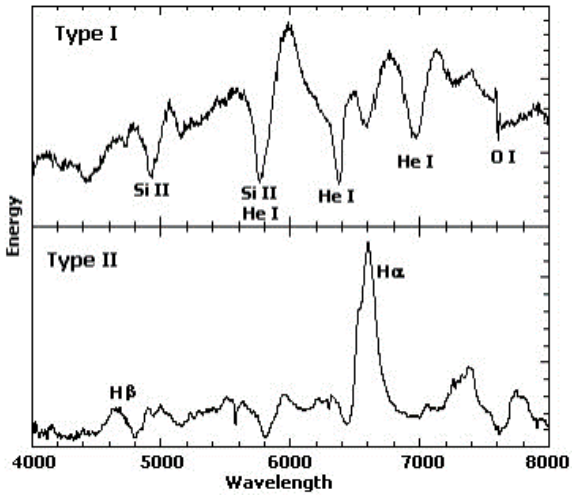


Figure 6.3: Spectrum of two types of supernovae.

of type II. However, we have to take into account the selection effects that tend to favor the type I frequency, which has a higher magnitude at maximum. In spiral galaxies the samples, corrected for selection effect, give the type II supernovae at least comparable to those of type I. The frequency depends on the brightness and the type of galaxy. For galaxies like ours, or M31, we have to expect from 1 to 3 supernovae every 100 years. However, these may not necessarily always be observable.

SN magnitude and spectrum

Figure 6.2 shows the absolute magnitude in visual band for different SN types as function of days since peak luminosity.

As visible in graphic 6.2, the core-collapse SN (SNII) are below $M_v = -18$ while SNIa, given by double star explosion, are just above $M_v = -19$ so there is a difference of more than 1 magnitude. These are statistical indicators. Figure 6.3 reports the spectrum of the types of supernovae. In particular it is visible that core-collapse SN shows H lines while SNI have very broad features and lines such as $SiII$ and He but not hydrogen.

Historical SN

As visible inside the table, there are only few SN historically known. Their frequency is very low; 10 SN have been detected in the last 2000, several in the last 500 years so in average 1 SN every century.

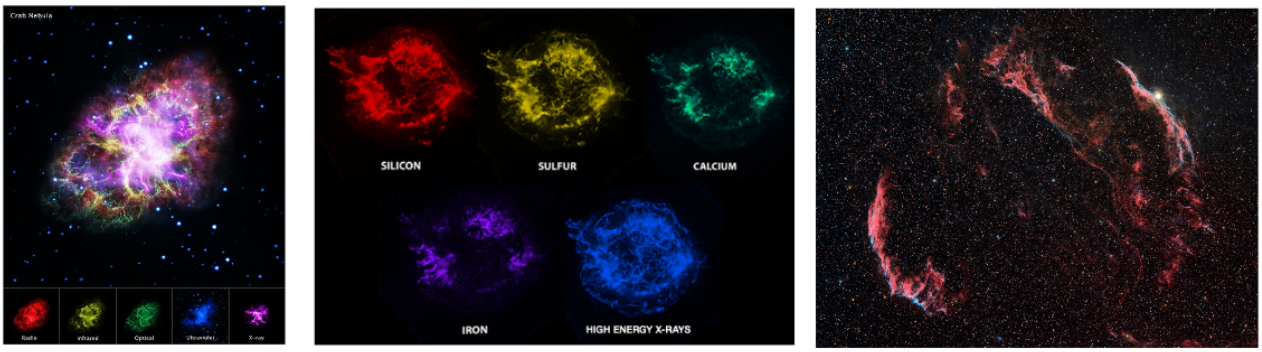


Figure 6.4: SNR examples.

Of course, they had been observed inside our galaxy and in particular, especially on the anti-galactic center direction and this is a contradiction because on the other side there are more stars. Of course, without instruments, historically only a small volume of the galaxy has been explored. Calculating all the SN we lost in the past and extrapolating in all the galactic disk, we can estimate a frequency of 100 SN in the last thousands years, which means 1 SN every 10 years.

6.1 SN remnants

A supernova remnant (SNR) is the structure resulting from supernovae events. The supernova remnant is bounded by an expanding shock wave, and consists of ejected material expanding. The interstellar material is sweeped up and shocked along the way.

Catalogues contain 215 SNR in our galaxy but the vast majority SNR we know are extra-galactic, usually observable only in radio due to the extinction by dust.

Figure 6.4 shows three exemplifying and important cases of supernovae remnants:

1. Crab Nebula in Taurus
2. Cassiopeia A
3. Cygnus Loop

6.1.1 Crab Nebula

The Crab Nebula, classified as optical as M1 (from the Messier catalog), occupied a fundamental position in the study of supernova remnants. Identified with the supernova of 1054 by Duyvendak, Mayall and Oort in 1942, is the brightest remnant of supernova, in radio, after the Cassiopeia A and one of the youngest SNR, with Cassiopeia A. Shklowsky in 1954 proposed that the continuum was produced by synchrotron radiation, hypothesis confirmed later from the polarization of the signal. The Crab Nebula is also one of the most intense known X sources and, along with the rest of the Vela supernova, it is the only remnant of supernova optical associated with certainty to a pulsar (another fifteen possible ones identifications has recently been proposed). The nebula is dominated by a filament structure and a sphere volume relatively filled with uniformly distributed material at high temperature and ejected at velocity of some thousands km/s. On the sky it has an apparent size of 4 $arcmin$ and thanks to its nature, M1 is also called "plerione" (dal greco, pleres, che significa pieno).

Distance limits have been obtained from the proper motion of the filaments using the so-called **nebular method**. If one knows the proper motions together with the radial velocity (measured by doppler shift of spectrum taken at the center), the distance geometric can be expressed by:

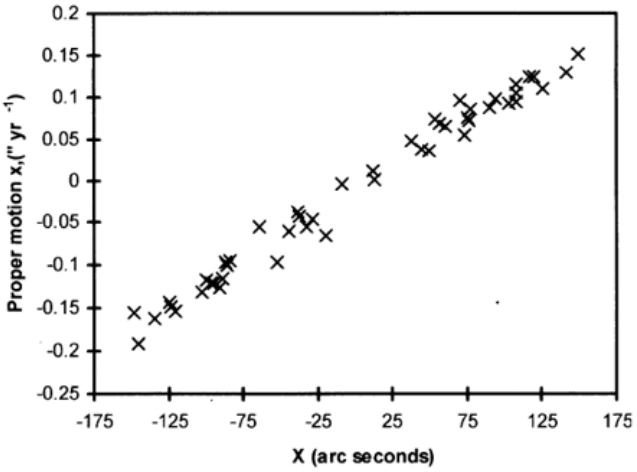


Fig. 1.—The x-component of proper motions vs. x-coordinate showing radial motion of the filaments.

Figure 6.5: Proper motion of M1.

$$d = \frac{v_r}{4.74\mu} \tag{6.1}$$

where d is the distance in Kpc, v_r is the radial velocity in km/s, μ is the proper motion in arcsecond per year, $"/yr$. 4.74 is a coefficient resulting from combining the different units of measurement. This is a solid method because it is a pure geometric one, obtained without assumption.

Given that, in the Crab Nebula, the radial velocity in the center it is 1450 km/s and the proper motions reach a maximum along the major axis of $0".22/yr$, the distance obtained is around $d = 1.4$ Kpc. This is probably a lower limit because the maximum speeds observed radials have a selection effect towards lower values. In general is assumed $d = 2$ Kpc. However today Gaia measurements give a distance $d = 3.37$ Kpc so there is something to more to understand.

Assuming the distance calculated by Gaia and calculating the apparent magnitude from historical documents (about $m_v = -4.4$), taking into account the reddening effect, from distance module we obtain $M_v = -18.2$. This means that M1 is one of the brightest core-collapse SN we know in our statistic.

The proper motions of the filaments provide two other useful elements, the time of expansion and the absolute position of the origin of the expansion. Extrapolating the supposedly uniform expansion movement (isotropic expansion) back in time yes obtains the beginning of the expansion. The first work on this was realized by Nugent in 1998. He scanned and compared the resent images of M1, measuring the proper motion. First he obtained the graphic in figure 6.5 which represents the proper motion as function of X-axis in arcsecond.

As visible in figure 6.5, at the center if the Crab Nebula the proper motion is zero. Going outside the center, from one side or the other up to the edge, proper motion increases up to the maximum value.

Then he took all vectors of proper motion and he went back to the origin of the expansion and the he found the barycenter point of expansion, as visible in figure 6.6.

After this huge work he found two important results:

- From analysis of proper motion vectors he found the starting expansion at 1130 ± 16 years but we know with certain from the documents that M1 started in 1054; a mistake of almost 80 years that gives a shorter age of about 20%. Of course there is not a mistake on historic document so the assumption is wrong: the expansion is not uniform. This means that the expansion has

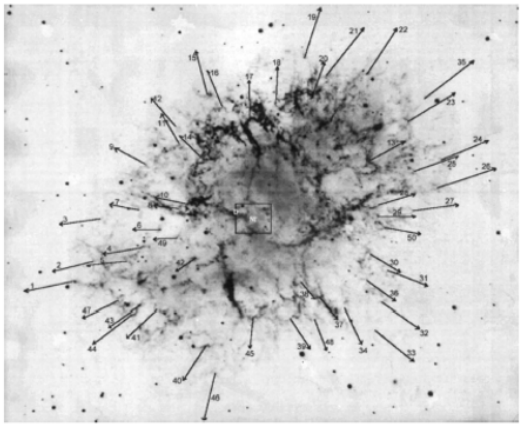


Fig. 4.—Arrows indicate the motions of the filaments in 250 yr at current expansion rates. The box at the center corresponds to Fig. 3. The expansion centers of Duncan, Trimble, and Nugent are shown by D, T, and N. This photograph was taken in 1960 with a narrow-band red filter showing $H\alpha$ emission.

Figure 6.6: Proper motion vectors of M1.

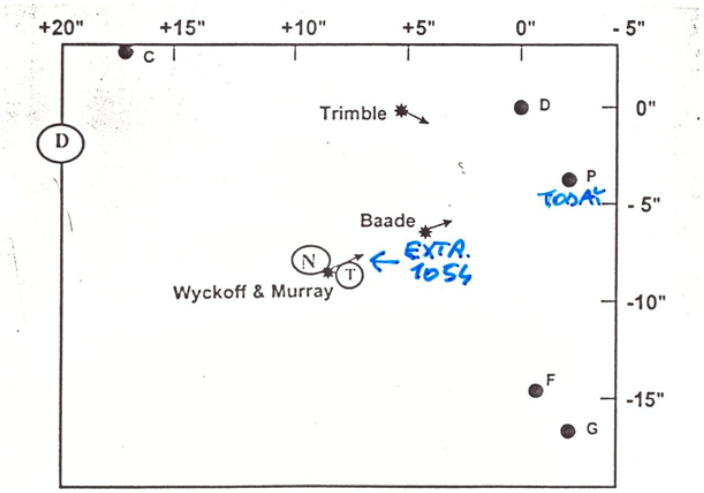


Fig. 3.—P is the current position of the pulsar. The expansion centers of Duncan, Trimble, and Nugent are shown by D, T, and N (see text). These are circled in proportion to the size of their average uncertainty. The symbol * shows the A.D. 1054 positions of the pulsar from proper-motion studies in the direction of motion. These absolute proper motions are from Wyckoff & Mur-

Figure 6.7: Pulsar position in M1.

accelerated and now it is faster than in the past. However acceleration it is not foreseen by the current models of the supernova remnants in the phase of free or adiabatic expansion, unless introducing a relevant contribution of pressure operated by cosmic rays emitted by the pulsar. So there is something to be studied in details.

- Nugent found also what is visible in figure 6.7. This graphic is a very expanded view of Crab Nebula center. The pulsar, what remains from the explosion, is on point P of the graphic but the barycenter position of proper motion is on point N so they don't coincide. To explains this, the idea was that the pulsar has a very high proper motion and then it moved away from the origin. Indeed, measuring proper motion of the pulsar and going back to 1054, we found that the pulsar was in point N (Wyckoff and Murray).

Barlow et al., in 2013 confirmed teh detection of $^{36}ArH^{+}$ in the Crab Nebula, a very well defined line, clearly visible in the radio spectrum. This tell us that there is a lot of ^{36}Ar produced by massive stars.

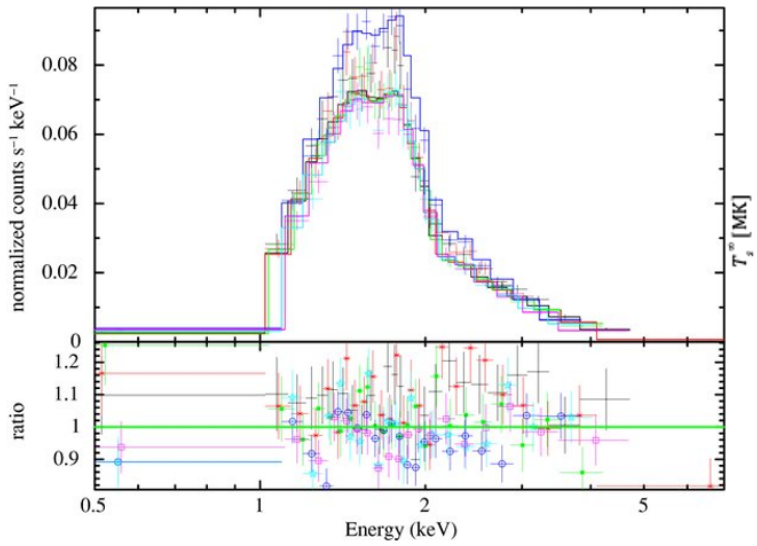


Figure 6.8: Cassiopeia A.

6.1.2 Cassiopeia A

Cassiopeia A (Cas A) is a supernova remnant in the constellation Cassiopeia and the brightest extra-solar radio source in the sky at frequencies above 1 GHz. The supernova occurred approximately 3.4 Kpc away within the Milky Way; given the width of the Orion Arm it is placed in the next-nearest arm outwards, the Perseus Arm, about 30 degrees from the Galactic anticenter. The expanding cloud of material left over from the supernova now appears approximately 3 pc across from Earth's perspective. In wavelengths of visible light, it has been seen with amateur telescopes down to 234 mm with filters. It is expanding at high velocity but it is empty inside; it is just a thin shell. SNR like Cassiopeia A are the most common type of supernovae remnants.

The youngest neutron star in the galaxy is the one in Cassiopeia A In X-ray it is a black body emission with a peack at wavelength about 1000 times shorter than visible corresponding to 2000 elettronvolt (figure 6.8) giving a surface temperature of about 2 million degrees.

In 2005 an infrared echo of the Cassiopeia A explosion was observed on nearby gas clouds using Spitzer Space Telescope. The infrared echo was also seen by IRAS and studied with the Infrared Spectrograph. Previously it was suspected that a flare in 1950 from a central pulsar could be responsible for the infrared echo. With the new data it was concluded that this is unlikely the case and that the infrared echo was caused by thermal emission by dust, which was heated by the radiative output of the supernova during the shock breakout. The infrared echo is accompanied by a scattered light echo. The recorded spectrum of the optical light echo proved the supernova was of Type IIb, meaning it resulted from the internal collapse and violent explosion of a massive star, most probably a red supergiant (about $15-20\ M_{\odot}$) with a helium core which had lost almost all of its hydrogen envelope. This was the first observation of the light echo of a supernova whose explosion had not been directly observed which opens up the possibility of studying and reconstructing past astronomical events.

Moreover in the last 10 years, recent measurements confirm that there is a very hot residual material with temperature decreasing from 1.7 to 1.6 million degrees, as visible in figure 6.9.

6.1.3 Cygnus Loop

Instead Cygnus Loop is a wide SNR with an apparent size on the sky of about some degrees. In this case you can just see the compression wave, which is in fact a supersonic shock wave, moving across the space. This is one of the oldest SNR that we know, with an age of about $10^5 - 10^6$ years, and it is the result of a massive explosion: we think the progenitor had 13/14 $M_{\odot}$. However what is visible is not the residual of explosion so it is not the ejected material (the material is so diluted in space

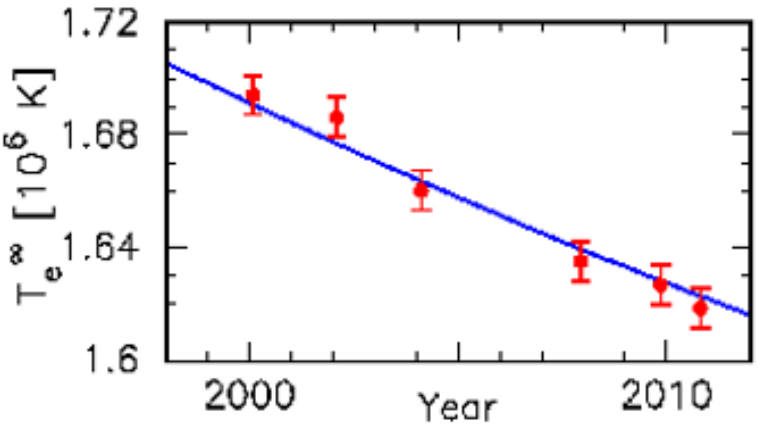


Figure 6.9: Cassiopeia A decreasing temperature.

that we can't see it anymore) but we see just the mechanical compression of interstellar medium left by the "piston" which is the ejected material.

6.1.4 Statistical results on supernova remnants

Of the 240 possible cataloged supernovae remnants, about 60% of the objects has been studied in detail. Milne concludes that almost all of them have a clear shell structure or at least a brilliant outer part. The remaining ones have a more uniform morphology like the Crab Nebula (Plerione type).

An important diagram for studying the evolution of supernovae is the surface brightness as a function of the diameter of the source, both in logarithmic scale. If the expansion mechanisms were identical and the energy involved comparable for all supernova remnants, the meaning of the diagram would be purely evolutionary. We have already seen that an expanding shell decreases in brightness exponentially over time. We therefore expect that the diagram can be interpolated with a power law.

$$\Sigma = const \cdot D^{-3.7} \tag{6.2}$$

where D is the diameter and the slope is 3.7. This important graphic is shown in figure 6.10.

From fig 6.10 we can show that the points are in according with the evolution: the diameter increasing and the brightness decreasing, becoming cooler with time. However it is not obvious that there is a small dispersion from the straight line, in spite of their double origins, double stars or core-collapse SN. This means they somehow lost their memory of their origin quite soon after the generation of SN remnants. This implies that then the process is driven by other factors then the initial velocity ejection and initial energy. Pay attention: R in the graphic is the real radius (not the apparent one) so $\log 2R$ is in parsec. Of course to know the real diameter, not the apparent one, it is necessary to know the distance.

Alternative method for SNR distance Another empirical result is the following: ones we have this relation for well known SN remnants, we can observe any kind of SN remnants, enter the surface brightness we see and then we can get the real radius. Then, comparing the real radius to the apparent one, we can get the geometric distance, independent on reddening and position.

Therefore, while the nebular method can be applied only to nearby stars visible in optical range, this last method is more powerful: it can be applied also for SNR detected in radio.

However the problem is that there is almost a 1 order of magnitude in brightness of dispersion. This means that there is another parameter to take into account. We know only that this dispersion is NOT caused by the different types of SN but there is something more to be studied.

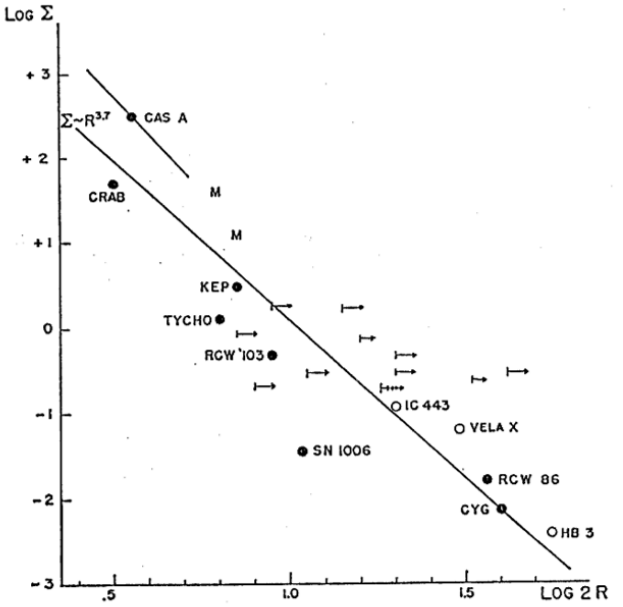


FIGURE 1: The Σ–*R* diagram. ●–well determined distances; ○–distance from interaction with H II region; ⟼–H_2CO, OH, or C_3H_2O absorption distance; ⟼–same but poor; M–Magellanic Cloud objects. The line through Cas A represents its present evolutionary track.

Figure 6.10: Surface brightness vs real diameter of supernova remnants from Woltjer (1972)

Brightness-diameter relation Clark and Caswell in 1979 made an interesting work about the relation between brightness and linear diameter of SN remnants. Figure 6.11 show the result for two different frequencies: 408 MHz and 5000 MHz. The slope is more or less the same as well as the dispersion and the line is well defined until the point corresponding to a diameter D of $40 - 50$ pc. Therefore, from this graphic, we understand that wider SNR have different evolution: the line is broken so there is something more in their evolution.

To be more specific, Clark and Caswell , from a selected sample of SNR, find a index of -3.0, almost constant in the range of frequencies between 408 and 5000 MHz.

From the calculations of Shklowsky and Kesteven for free expansion we get that the superface brightness should decrease with an index of -6 for a uniform sphere and -4.5 for a shell of constant thickness, both higher of the value -3.7 found empirically. It can be deduced that most of the supernova remnants expand in a phase in which the expansion is not free but conditioned by the interstellar medium. The diagram also shows a considerable dispersion around the line of interpolation, in many cases higher than measurement errors. This indicates that galactic supernovae produce expansion velocities in a broad interval. The phenomenon of evolution could be much more complex than as assumed with the hypothesis of a single phase of free expansion, as will be discussed in the next chapter.

The statistical analysis of Clark and Cashwell gives an inclusive spectral index (representing the energy of material inside SNR) of the continuum in radio between -0.1 and -0.8, with an average around -0.45. We know that the spectral index of the synchrotron radiation depends on the energy distribution of the electrons α. Assuming the syncrotron as dominant emission mechanism, from -0.45 we get $\alpha = -1.9$, much lower than the index observed of -2.5 for cosmic rays. Since it is assumed that the diffuse cosmic rays result from supernova explosions, a mechanism must therefore be found to explain the difference in the energy distributions.

Another important result of the Clark and Caswell statistics is that there is no correlation of the spectral index (index of the slope of the spectrum) with the diameter (and therefore with the age) of supernova remnants (figure 6.12), suggesting a rapid expansion compared to the internal energy decay of relativistic electrons. In other words, the kinematic evolution is faster than the loss of energy of SNR.

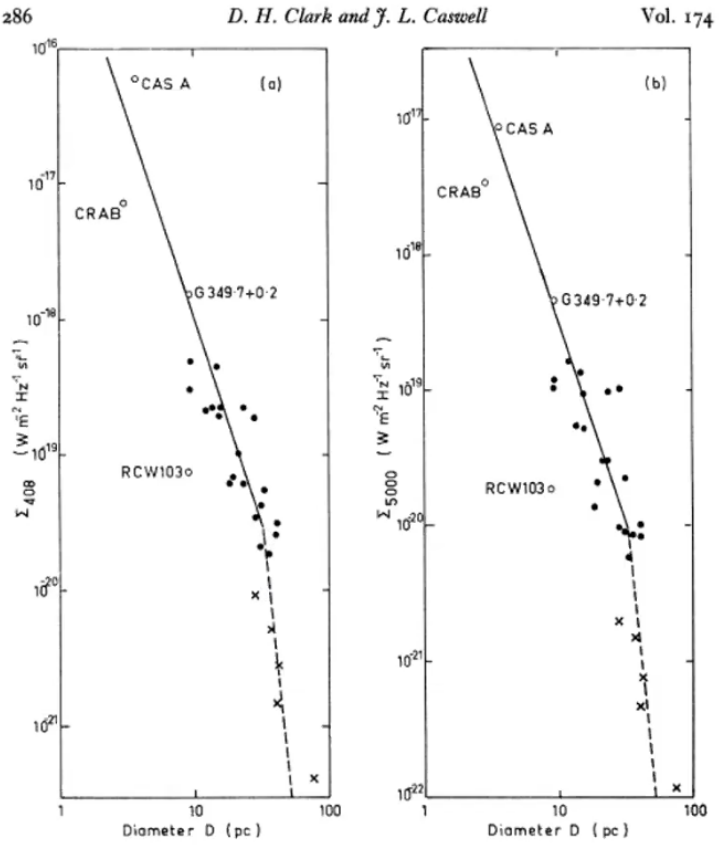


Fig. 1. *Surface brightness vs linear diameter: (a) at 408 MHz, (b) at 5000 MHz. The class 1 calibrators used to determine the solid line (adopted relationship) are shown as filled circles; the open circles, crosses and the broken line are explained in the text.*

Figure 6.11: Clark, Caswell, 1979: Brightness/linear diameter SNR relation

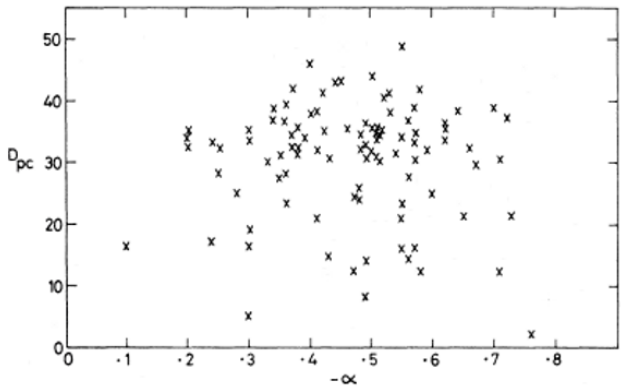


Figure 6.12: Spectral index vs. Diameter for supernova remnants from Clerrk and Caswell (1976)

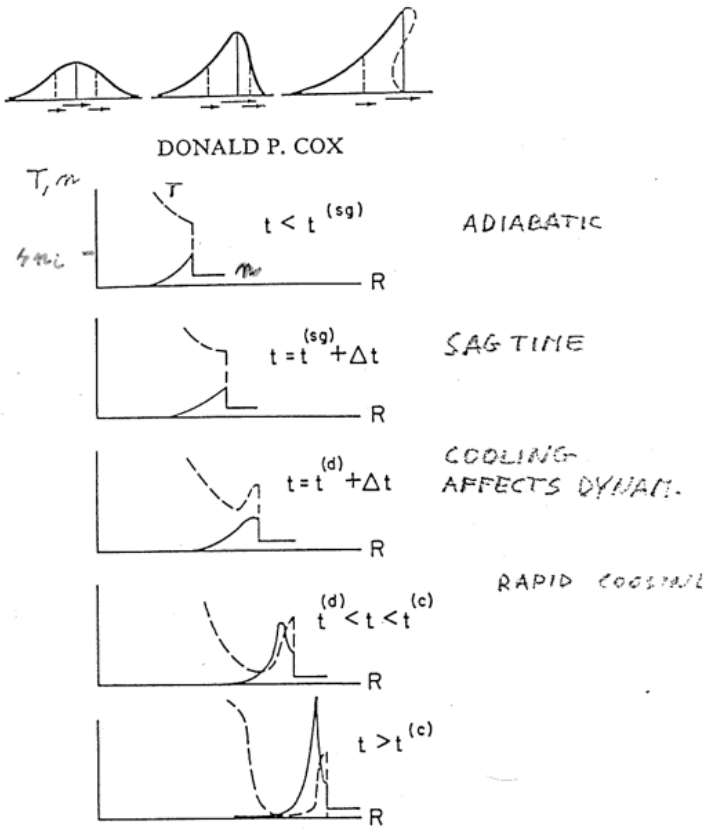


FIG. 1.—Qualitative diagrams illustrating changes in the run of T/T_s (*dashed curve*) and n/n_0 (*solid curve*) with R for several epochs in the history of a remnant. In the uppermost diagram, the swept-up interstellar mass is assumed to greatly exceed the mass originally ejected. In the last diagram, the shell thickness is greatly exaggerated.

Figure 6.13: Supersonic shocks, evolution of temperature t and density n in adiabatic and isothermal phase (adapted from Cox, 1972).

6.2 Evolution of the supernovae remnants

Current evolutionary models foresee, as originally described by Woltjer and from Ilovaisky and Lequex in 1972, three distinct phases of evolution:

1. short-term free expansion in 100-300 years;
2. adiabatic (or Sedov) expansion with shell development;
3. isothermal expansion accompanied by cooling by radiation;
4. total mixing with the interstellar medium.

Most supernova remnants would be found in the second phase or between the second and third one. Only the youngest, such as Cassiopeia A, may have just come out of the first phase, while at the other extreme we have the Cygnus Loop which should be in isothermal phase.

The classical treatment of the evolution of supernova remnants, in the different phases, is due to Cox and Woltjer (1972), while diagnostics with the calculation of observable quantities (line intensity, continuum emission) is discussed by Dopita and Chevalier (1977). In a more recent paper Kafatos and collaborators (1980) have instead calculated the effects interactions of supernova remnants with an interstellar medium with different densities and temperatures. Reynolds (1988) developed a non-classical treatment, already introduced by Chevalier (1977), which includes the effects of turbulent propagation.

A summary of the results of the classical treatment, according to the model of Woltjer, of the 3 expansion phases is illustrated in the next subsections and in figure 6.13.

6.2.1 Free expansion

The free expansion phase assumes that the density of the interstellar medium is negligible compared to that of the rest of supernova. SNR gas acts as a piston that compresses the interstellar gas and the speed of the ejection does not undergo an appreciable deceleration.

The expansion occurs between low density gases with mean free path of hundreds of parsecs. This implies that there is no direct collision between particles. The interaction between SNR particles and the interstellar medium occurs through magnetic coupling.

In the typical conditions of supernova explosion the perturbation of the medium is quite strong, with velocity much higher than the speed of sound, causing the formation of shock waves which propagation depends on gas properties, such as density.

It can easily be verified that the wave due to an explosion of a supernova is highly supersonic. In fact the speed of the typical sound of the interstellar material goes from 10 to 100 km/s, while that of gas ejected from a supernova remnant moves at about $10,000$ km/s. At this high velocity, the ejected material, more or less $1 - 0.5$ $M_{\odot}$, is diluted in the space in a very short time that it gets the same density of space.

Unlike sonic waves, the approximation of $dP/P \ll 1$ (normally used for sonic waves) is not applicable and the wave deforms with the peak which tends to anticipate the entire perturbation, given that the speed of propagation of the perturbation is greater in the center, at higher density (see the upper part of figure 6.13). Therefore it releases a lot of energy in the gas and creates a discontinuity of limited spatial extension where the material of the surrounding medium undergoes a sudden compression and thermalization which means interstellar gas is suddenly shocked and reaches very high pressure and density. The gas is not removed but it is shocked. By the way, since the gas is ionized, there is some UV and X-ray emissions from the front of the shock wave.

The classic solutions of the shock wave (currently called shock), in this situation, can easily be obtained by imposing the condition of continuity and the equivalent of the laws of motion, in the condition generally verified in the initial phase, of adiabatic transformations. The post-shock density, pressure and temperature, as a function of density of the medium and the speed of expansion (Sedov solutions), are derived:

$$\rho = 4\rho_0 \tag{6.3}$$

$$P = \frac{3}{4}\rho v^2 \tag{6.4}$$

$$T = \frac{3}{16} m \frac{v^2}{K} \tag{6.5}$$

where:

- P is the post-shock pressure;
- ρ the post-shock density:
- ρ_0 the density of the medium;
- T the post-shock temperature;
- v velocity of shock wave;
- m the average molecular mass.

6.2.2 Adiabatic expansion

These conditions remain valid also in the second phase, when the SNR density, as a result of expansion, has become comparable to that of the medium (or, in other words, the amount of swept matter from the shock wave it becomes comparable to that of the supernova itself). To be more specific, the density reaches a peak, 4 times higher than interstellar one, and no more. This limit is due to the fact that when a gas is compressed in adiabatic transformations, a counter pressure is created because it is hitting up therefore the gas resists to further compression. Density increases while, behind the shock temperature increases blocking the new increasing density.

At this point the speed starts to significantly decrease. With decreasing velocity, also pressure and temperature decrease. This condition is reached when the radius $R(t)$, function of time because it is expanding, satisfies the relation:

$$\frac{4}{3}\pi R(t)^3 \rho_0 = M_0 \tag{6.6}$$

where M_0 is the ejected mass and ρ_0 is the density of the interstellar medium in atoms per cm^3.

However, energy conservation conditions continue to apply (we can also demonstrate that kinetic and thermal energy are conserved separately):

$$\frac{1}{2}M(t)v^2 = E_0 \tag{6.7}$$

where E_0 is the initial energy, constant in this phase, and is comparable to the (kinetic) energy released by the explosion, since the irradiated energy is at least an order of magnitude lower. In this equation the initial mass M_0 is neglected, being the component $M(t)$ rapidly growing with time.

However, the total mass is variable and increases at the expense of the interstellar medium. The expansion speed therefore decreases as a function of the radius (or of the time). It is clear that, under the same conditions of the explosion, the transit to this phase is shorter, greater is the density of the interstellar medium. At this stage the only energy losses are due to free free transitions and mostly free-bound emission lines are starting. Equation 6.7 in this phase can also be written as:

$$\frac{1}{2}\left(\frac{4}{3}\pi R(t)^3 \rho_0 + M_0\right) v(t)^2 = E_0 \tag{6.8}$$

where $\frac{4}{3}\pi R(t)^3 \rho_0$ is the mass of interstellar medium while M_0 is the mass of the original ejected material. Of course, as $R(t)$ increases with time, as a consequence of increasing mass, the velocity v decreases and therefore shock waves slow down. The integration of 6.8 gives the radius as a function of time:

$$R(t) = const \cdot \frac{E_0}{\rho_0}^{1/5} t^{2/5} \tag{6.9}$$

Equation 6.9 together with the equations that regulate density and temperature in the wave fronts, describe the conditions of supernova remnants in the phase of adiabatic expansion. It is easy to show, for a speed of $1,000\ km/s$ that the temperature inside the shock front reaches about $10^8\ K$. The density instead increases up to the maximum value of $4\rho_0$.

Equation 6.9 is important to derive the characteristic parameters of the SNR. In general the only observable data are the radius R (once known the distance) and thus two unknowns remain, the E_0/ρ_0 ratio and the age t. The study of historically known supernova SNRs as Ticho, Kepler and the Crab Nebula, allow to obtain directly E_0/ρ_0 (then the calibration) being known the time t.

The values obtained from Clark and Caswell (1976) are around at $10^{51} - 10^{52}\ ergcm^{-3}$. Assuming a density of interstellar gas around to a hydrogen atom per cubic centimeter, we obtain that the energy release from the explosion of the supernova remnant in kinetic form is around $10^{51} - 10^{52}\ erg$. This is about two orders of magnitude lower than the total graviational collapse energy released (mainly as neutrinos energy) by the supernova. Once this parameter is set, the relationship radius–time is therefore defined. In a histogram of distribution of the number of lower SNRs at a certain radius, against the radius we should therefore expect a straight line with slope 2.5 on a logarithmic scale. Moreover from the observed radius, the age can be easily obtained.

Energy losses, are small ($dE \ll E_0$), due mainly to free free emission expressed by:

$$-\frac{dE}{dt} = cost \cdot n_e^2 t^{1/2} \cdot volume \propto t^{3/5} E_0^{4/5} \tag{6.10}$$

As the evolution progresses, the gas cools down, the shock wave slows down and the loss of free free slowly increases as a result of the increase in total mass which compensates for the decrease in temperature. It is estimated that the free free losses in $6 \cdot 10^4$ years (which is the typical age of the observed SNRs) are less than 1% of the total energy, therefore the hypothesis that the process is adiabatic in this time scale is valid.

When the temperature drops to a few million degrees, the radiation from free-bound and bound-bound transitions, due to elements heavier than hydrogen (C, O, N), gets dominant. In this situation energy losses become important, with an inversely proportional power output with the temperature (radiative phase). This indicates that losses grow rapidly over time:

$$-\frac{dE}{dt} = cost \cdot \rho_0^{1/5} E_0^{9/5} t^{12/5} \tag{6.11}$$

The high power index of time, compared to the free free losses, demonstrates what has been said. We can define a time t', at which half of the initial energy is irradiated and beyond that the transformation are no more in adiabatic conditions:

$$t' = 5 E_0^{4/17} \rho_0^{-9/17} \tag{6.12}$$

6.2.3 Isothermal expansion

Then SNR enters the isothermal phase: the energy transferred to circumstellar material (CSM) is nearly immediately radiated leaving temperature T unchanged. Therefore this phase is characterized by constant temperature in a very thin outer layer. At the same time, just behind this thin layer, density is rapidly increasing to a value which is much more than a factor 4, due to the cooling. Indeed gas can not oppose anymore with thermal pressure and the density increases. Moreover in this phase the momentum (mass per velocity) is constant.

6.2.4 Last phase

Then there is a total mixing with the interstellar medium at a sonic speed.

Values For a density of the interstellar medium of 1 $atom/cm^3$ we obtain:

- $R = 40\ pc$;
- $t' = 10^5\ years$;
- $v(t') = 100\ km/s$.

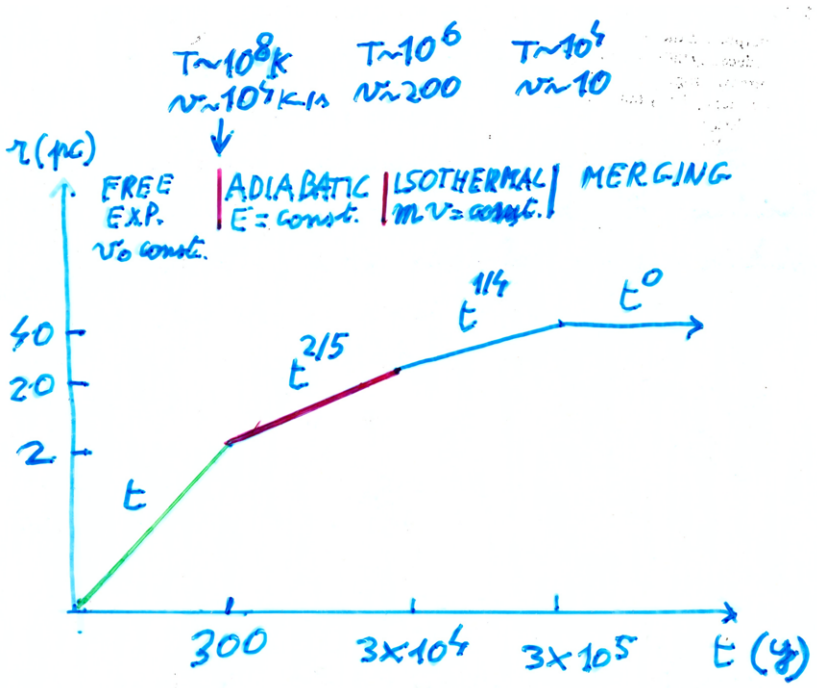


Figure 6.14: Evolution of the size with time (from Ortolani)

These values are not very different from those of the largest SNRs observed such as Cygnus Loop or IC 443. The calculation of the total mass within a sphere of radius $R = 40\ pc$ gives a mass of the order of one 100 $M_\odot$. A first consequence of these results is that most SNRs are in the adiabatic phase (they have $r < 40\ pc$) and have an age of less than 100,000 years. The dependence on the initial conditions (E_0 and ρ_0) is not very strong. We see also that the evolution proceeds faster at higher densities of the interstellar medium.

A schematic representation of all these phases is visible in figure 6.14 and in figure 6.15.

In figure 6.16, the first table is another scheme of SNR evolution while table 1B and 1C report a particular effect: hot cavity of SNR. Suppose to have an explosion of SN in a star cluster. The heat realised to interstellar matter makes the gas expanding and decreasing the density. In these conditions, a second massive exploding star expands in a hot cavity with very different expansion time due to minor interstellar density (much longer by factor 10). This means that we may not see the second explosion of SN until it is so fast to reach the edge of the previous one.

6.3 SNR frequency

Frequency of SN remnants, which means $1/t$ where t is the time, is given by:

$$f(< R) = \frac{N(< R)}{t(< R)} \tag{6.13}$$

where R is a specific radius, to make a selection in size and time, $N(< R)$ is the number of events of SN and t is time. Assuming a diameter of 40 pc, $f = \frac{200}{3 \cdot 10^4} \rightarrow \frac{1}{75}\ yr^{-1}$ which means 1 event every 75 years. In this case 200 is the number event give by catalogues while $3 \cdot 10^4$ years is the age of adiabatic phase.

Figure 6.17 shows the number of events as function of a specific diameter. It is possible to observe that before 40 pc, the line is almost constant and we are near to the completeness without distinction on SN of type I or II.

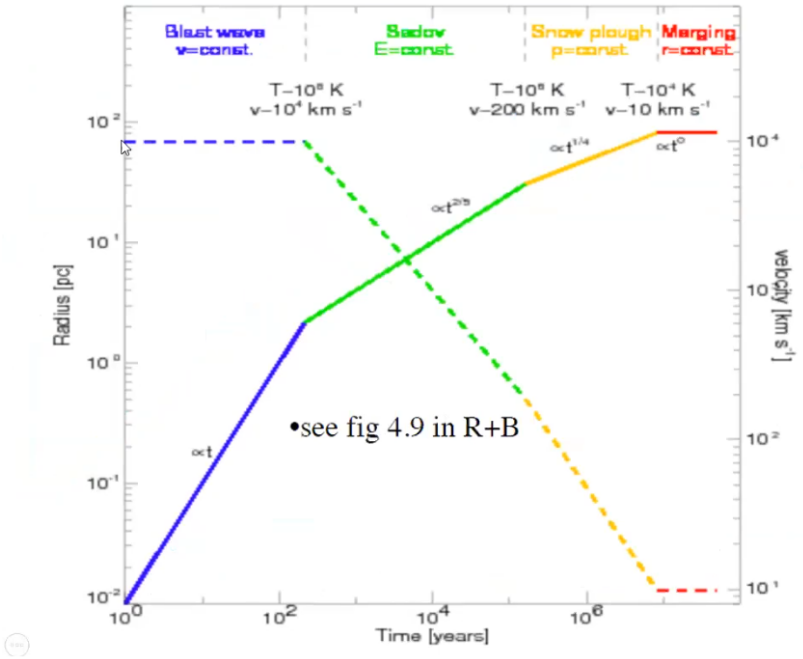


Figure 6.15: Evolution of the size with time (from Janfei Jang)

TABLE 4. Schematic evolution of an SNR[a]

	t (yr)	R (pc)	V (km/sec)	M ($\odot$)
Phase I–II	90	0.9	10,000	0.2
II–III	22,000	11	200	180
III–IV	750,000	30	10	3600

[a] $\epsilon_0 = 10^{50}$ ergs, $V_0 = 10{,}000$ km/sec, $M_0 = 0.1\ M\odot$, and $n_{in} = 1$ cm^{-3}. The age t, radius R, velocity V, and mass M are given at the transition points between the four evolutionary phases.

TABLE 1B

THE HOT CAVITY SNR ($n \sim 10^{-2}$ cm^{-3}, $T \sim 5 \times 10^5$ K)

t(yr)	V_S(km s^{-1})	T_S(K)	R_S(pc)	Remarks
0.....	5000	3.5×10^8	0	Explosion occurs
3.7×10^3...	5000	3.5×10^8	19	Free expansion ends
1.4×10^5...	300	1.3×10^6	105	SN shock encounters moving bubble shell, and it quickly gets decelerated to the bubble velocity (21 km s^{-1})

TABLE 1C

THE CLASSICAL SNR ($n \sim 1$ cm^{-3}, $T \sim 10^2$–10^4 K)

t(yr)	V_S(km s^{-1})	T_S(K)	R_S(pc)	Remarks
0.....	5000	3.5×10^8	0	Explosion occurs
6.2×10^2...	5000	3.5×10^8	3.1	Free expansion ends
2.9×10^4...	265	10^6	19.6	Adiabatic phase ends
2.6×10^5...	5	2.5×10^3	73.6	Shell stalls

Figure 6.16: Interstellar medium effects on the SNR evolution (classical Woltjer, 1972, hot cavity, classical).

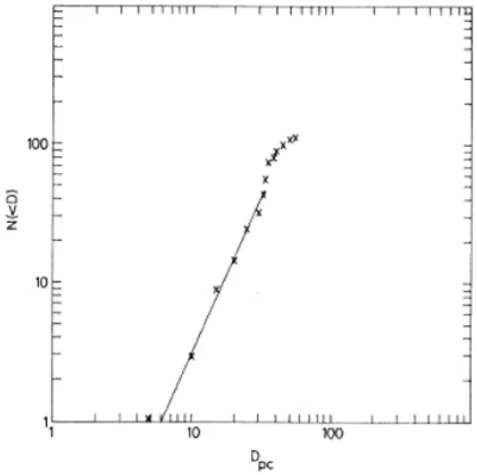


Fig. 4. *N–D relationship for the whole of the Galaxy.*

Figure 6.17: D-N relation

Instead figure 6.18 shows SN frequency in the Milk Way from different tracers (Rozwadowska et al., 2020). The red X-axis to left corresponds to SN events per year, so the frequency while the X-axis to right represents SN events per year and per pc^2.

First of all, historical SN, as said before, corrected by bias, give 1 event every 10 years. Pulsars, residual of core collapse, are 1 every 40 years. SN in external galaxies are distributed in a wide range because it depends on the distance of the galaxy. Since they are measured using the Hubble expansion rate, their frequency depends on the expansion of the universe. Adopting a standard value of $H = 65$ km/s, the frequency is 1 event every 70 years. Instead the frequency of SNR is 1 every 65 years. Finally Wolf Rayet stars, the most massive ones, have a very low frequency: 1 every 200 years. In general, we can see that there is a wide dispersion of estimate, from 1/10 to 1/200.

Exercise Figure 6.19 is an interesting exercise considering galactic disk volume vs. supernovae remnants total filling volume. Suppose to have a frequency of 1 SN every 10 years with radius 70 pc and $t = 10^5$ yr. How much is the total volume if we put together all these spheres of SNR? Do they fill all the galaxy?

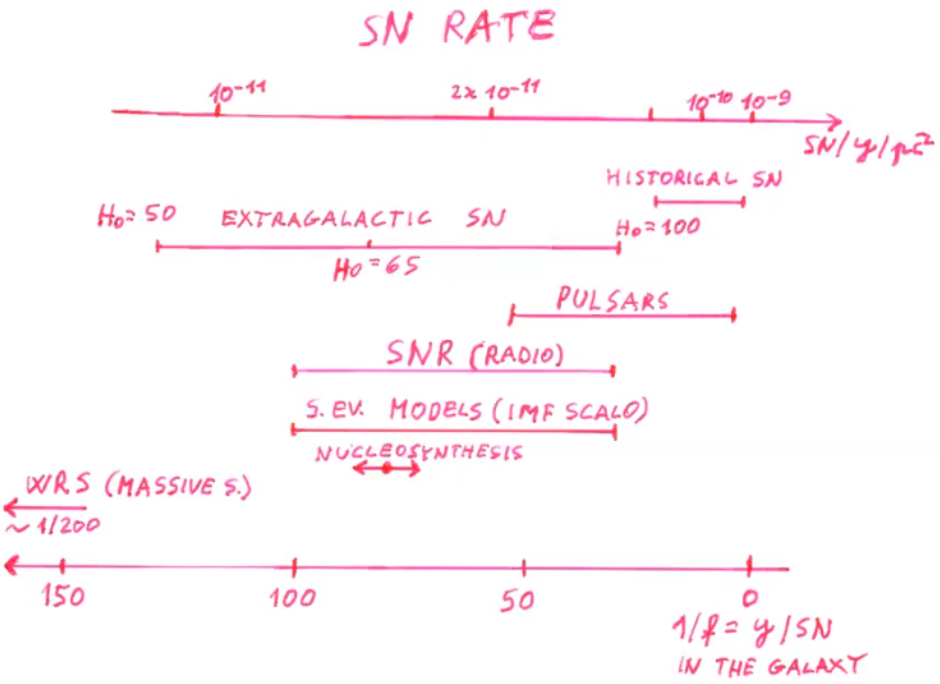


Figure 6.18: SN Rate

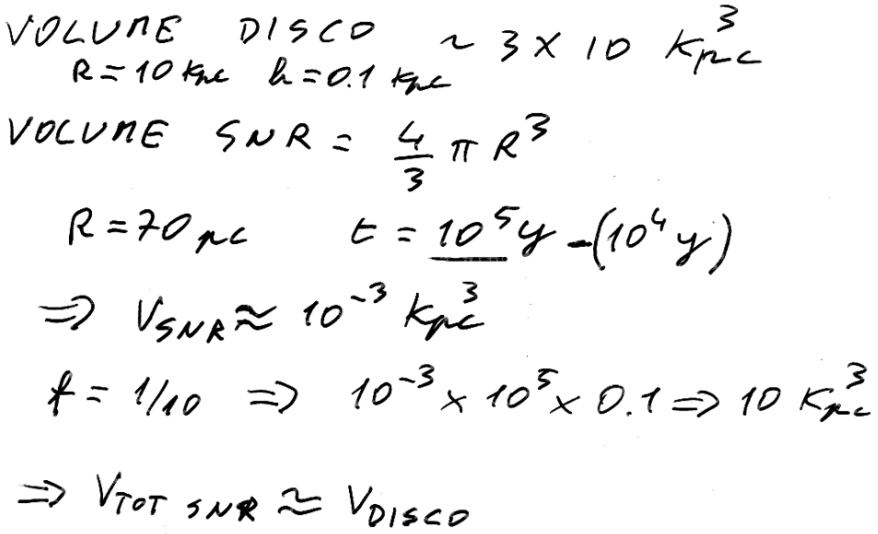


Figure 6.19: Exercise.

Chapter 7

Maser

7.1 Molecular lines and maser emission in the galaxy

In general, molecules can have 3 types of quantized transitions: the first is referred to the electronic transitions, the second one to the rotational transitions and the third to the vibrational transitions. A molecule therefore has a total energy E_t that can be written as a sum of three factors, electronic energy, vibrational energy and rotational energy:

$$E_t = E_e + E_v + E_r \tag{7.1}$$

Electronic transitions with energy E_e are transitions of electrons between different quantum levels. Usually they are not visible in IR or radio because they are very energetic so they are usually visible only in optical and eventually X-ray windows. This is not, however, a rigid rule, since, for example, a transition between electronic states produces one of the most common lower frequency radio lines: HI line at 21 cm. It is a forbidden line due to its tiny separation between levels: it is caused by a spin-invesrion transition at very low energy. In general the energy bond between ions is relatively weak compared to that of first ionization, i.e. the energy required to separate two ions that form a molecule is normally smaller than that required to extract an electron from a neutral atom [1]. Only some alkaline atoms have ionization energies smaller than the highest molecular binding energies.

E_v is instead the energy necessary for vibrational transitions inside molecules. The vibrational levels are therefore of low energy with the highest energy level corresponding to the molecular binding energy. Transitions between vibrational states produce lines the infrared regions up to about $10 - 20$ micron. On the other hand the vibrational transitions are not observed normally in the optical domain for spectra of astrophysical importance. The relative vibrations of two atoms can be seen as harmonic oscillations. These are quantized.

In addition to vibration, the molecules have another degree of freedom which is the possibility of rotating around the molecular center of gravity with energy E_r. Also this energetic state is quantized and the energy differences between the rotational states are in general smaller than those of the vibrational states. Both in case of vibrational and rotational transitions, the levels are narrow enough to observe transitions in the IR and radio wavelength.

Therefore in general it is valid:

$$E_e > E_v > E_r \tag{7.2}$$

[1] In the case of the diatomic molecule of hydrogen H_2, for example, the bond energy is 4.48 eV, against an ionization energy of the atom of H is 13.6 eV

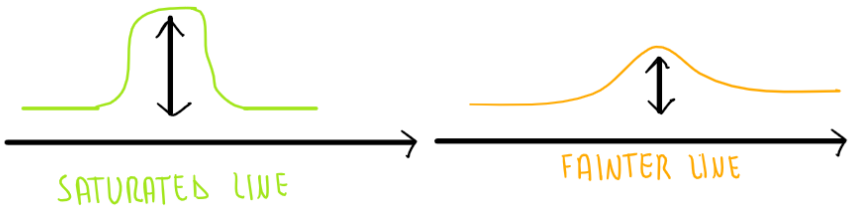


Figure 7.1: Types of lines.

As concerning E_v and E_r, we need a condition about electronic dipole emission. If the molecule is symmetric (barycenter of mass coincident with barycenter of charge), there is no dipole emission. So in the case pf the most common molecule in interstellar matter, which is H_2 perfectly symmetric, we can only observe transitions in UV and not radio or IR. However it is very important to have light in radio in order to observe, thought the galactic disk, sources that are not visible in optical due to absorption by dust.

Moreover, the vibrational states are further divided into rotational levels with smaller separations. Transitions between these states lead to lines especially in the radio field. As a first approximation most of the molecules can be compared to rigid rotator, obtaining as a result a series of equally spaced lines. Transitions between sublevels in which individual rotational lines are divided are also important in the radio region. Important examples are the Λ splitting and the separation of hyperfine structure.

7.2 The emission of CO

Among the most important interstellar molecular lines certainly the rotational lines of the CO at 2.6 mm occupy a dominant position. CO is among the most stable molecules ($E_e = 11.1\ eV$) and derives from the combination of the two most abundant elements, after hydrogen and helium. It can resist to cosmic rays, X-rays anda radiation in interstellar matter for some thousands years so about 10^3 years. It is a very small time but we can see it and other elements with short time-life because these molecular lines are continuously produced in interstellar matter by different processes.

Although its abundance is about 10^4 times lower than that of the H_2 molecule, the CO has been detected everywhere in the galactic disk, with high concentrations in regions of small diameter and small dispersion of radial velocities, identifiable with compact molecular regions of low temperature (about 10 K). The CO emission maps in the radial-galactic velocity diagrams are very similar to those obtained from the 21 cm of the hydrogen, but at high resolution show a complex structure of compact regions. The CO lines are currently used both to obtain the molecular abundances in our galaxy, in particular abundance of H_2 which is not visible, and for the temperatures and density of the Bok globules (small part of dark nebulae). We can get the temperature when the line is saturated (left one on figure 7.1): when optical depth is very high, brightness temperature is more or less the black body temperature so the physical one. Instead, when the line is fainter and optical thin (right one in figure 7.1), you can measure the column density because you observe thought the cloud and not just the surface as when it is opaque. Then, form size, we can get the total mass of the cloud.

There is also something peculiar: from unsaturated lines we can measure the width. The broadening of a line normally is dominated by thermal-Doppler shift so by particles thermal movement described in Maxwellian distribution of velocities. However the broadening given by thermal movement is not enough to explain the width we measure. This means that there are other phenomena in the cloud that increase the kinetic energy (for example collapse, turbulence and so on).

However CO is something more complex because it has many lines around 2.6 mm due to the present of isotopes ^{12}C and ^{13}C. The ratio between these two isotopes is about 40 on average for interstellar medium while on Earth it is about 86 so something had changed locally during Earth formation. This ratio is very important because it give information about origin of carbon but how to get the ratio?

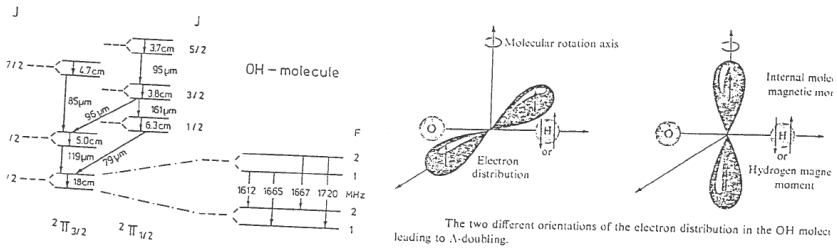


Figure 7.2: Different orientation of OH molecule. The transition between these two orientation generates the lines at 18cm and the transition is called the Λ-doubling.

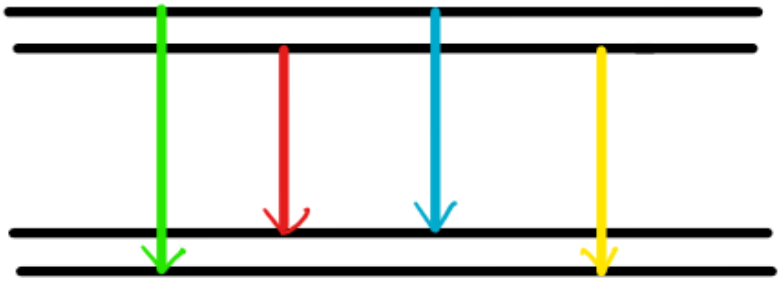


Figure 7.3: Hyperfine structure.

The relative abundances of the lines originating from ^{12}C ^{16}O against ^{13}C ^{16}O allow, in theory, to derive the isotopic interstellar abundances ^{12}C / ^{13}C. However, the measurement is made difficult by the small difference in energy and by the frequent saturation of the ^{12}C ^{16}O, from which the temperature of the clouds is more easily obtained. So the solution is just to measure another line which is abundant and never saturate: ^{12}C ^{18}O. Therefore usually they do the ratio $\frac{^{12}C^{18}O}{^{13}C^{18}O}$.

For the isotopic ratios it is preferred today to use instead the H ^{12}C ^{16}O / H ^{13}C ^{16}O ratio since these lines are almost always optically thin.

7.3 The emission of OH lines

OH is a quite common molecule (with abundance about 10^{-5} compared to H) an it is an asymmetric molecule. As visible in figure 7.2, there are two different orientation for rotation, both orthogonal to the joining line between two atoms. The momentum is different depending on rotational axis we consider. A transition between these two orientation generates the lines at 18 cm: this is called Λ doubling.

The 18 cm lines of hydroxyl (OH) were observed, for the first time, in absorption, in the direction of Cas A in 1963. They were the first molecular radio lines revealed in radioastronomy.

These lines come from hyperfine structure transitions of the two levels corresponding to the Λ splitting of the lowest electronic, vibrational and rotational state of the OH molecule. In this state two different orientations of the electron distribution are possible: one along the axis of rotation of the molecule and one perpendicular, in the plane of rotation. The transition between these two states, due to the splitting, gives rise to the line at 18 cm (figure 7.2). In turn, each of these two levels has a hyperfine structure and is separated into two components each. There are therefore four transitions between these levels around 18 cm, as visible in figure 7.3.

The line at 18 cm has often been observed in emission. The lines are relatively narrow, corresponding to Doppler widening of 100 K. We can calculate the transition probabilities of each of this line and we can check from observational data if they correspond to theoretical expectation. They do not: the

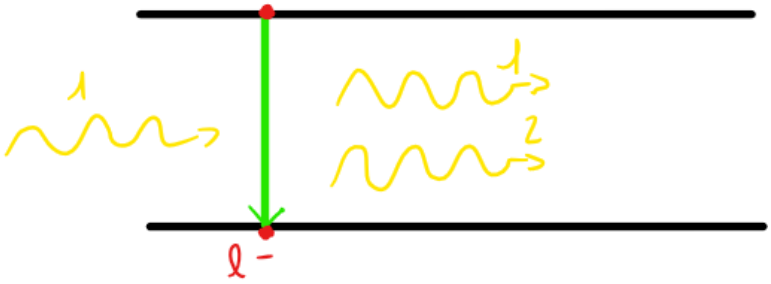


Figure 7.4: Stimulated emission.

relatively intensities are very different from what we expect. Moreover, when they have been detected for the first time, it has been noticed that these lines are very bright, very thin and also with a high degree of polarization, both linear and circular. So nothing to do with other common lines. The intensity is very high as well as surface brightness (around 1 $L_\odot$ in only one line) because the angular size of sources is very small, around the milliarcsecond, and the corresponding brightness temperatures therefore reach very high values, around $10^{10} - 10^{15}$ K, difficult to reconcile with a thermal Doppler width of no more than 100 K.

It should be noted that, unlike the hyperfine structure transition of the 21 cm hydrogen atom, these transitions are permitted. Their probability of transition is about 4 orders of magnitude higher. This justifies the observed intensity of these lines despite the effect of a factor 5 of less abundance than OH compared to H.

All this indicates that the emission of OH does not occur in conditions of thermodynamic equilibrium. This is why the 18 cm lines of hydroxyl are now explained with the maser mechanism.

7.4 Maser emission

The maser emission mechanism is based on the stimulated emission in environment where there is a metastable state of excitation called "population inversion". This term indicates that the number of atoms or molecules present in a high energy level is greater than that of the atoms or molecules that are at a lower energy level.

Normally the emission works in this way: we have an electron and when a photon arrives, the electron goes to the upper level. Then this electron can go down to ground state by spontaneous decay and spontaneous emission or by the stimulated emission. In particular, if then arrives a photon (while the electron is on the upper state) has the exactly energy corresponding to the difference between the two levels, it causes the stimulated emission. In this specific case, the electron goes down emitting two photons: the first one the original one, the second one is coming from the transition (figure 7.4).

In conditions of local thermodynamic equilibrium this is not generally possible as the collisions tend to redistribute the populations according to the Boltzmann equation (statistical equation), that is with a population that decreases exponentially with increasing energy levels:

$$\frac{n_1}{n_0} = \frac{g_1}{g_0} e^{\frac{h\nu}{KT}} \tag{7.3}$$

where n_1 and n_0 are the number of atoms/molecules in the upper state and lower, respectively, g_1, g_0 the corresponding statistical weights of the levels, K the Boltzmann constant and T the kinetic temperature of particles. $h\nu$ is the energy of the photon corresponding to the two levels, the ground state and the upper state while in general $g_1/g_0 = 1$.

It is clear from 7.3 that, for unitary ratios of the statistical weights of levels, there is no temperature value T for which the ratio n_1/n_0 can become unitary. For values of T close to zero the ratio n_1/n_0

also tends to zero, while at increasing values of T, the ratio increases, while remaining always less than 1 and tends asymptotically to this value for T increasing to infinity. From the formal point of view one can have a population inversion (break the thermal equilibrium), so have more electrons in the upper level then the lower lever, only by introducing in the equation 7.3 **negative temperature values**.

From a physical point of view, deviations from this equilibrium condition are possible only if the density is lower than a critical value, which depends on the radiation field and on the ratio atoms/molecules present, to prevent collisions from "thermalizing" the distribution with greater speed of the opposite process that disturbs the balance. In general, cosmic masers operate at densities of the order of about $10^5 - 10^{10}$ particles per cubic centimeter, which correspond to $N(OH) = 10 - 10^6$ since $N(OH)/N(H_2) = 10^{-4}$ (the density of the Earth's atmosphere, by reference, at sea level is $2 \cdot 10^{19}$ $particles/cm^3$).

7.5 Three-level maser mechanism

How to overpopulate the upper level? Of course, not by collision or radiation because in this way we pumping the upper level but at the same time it is pumping down particles by stimulated emission. So there are other explanations of overpopulation.

Since the edge is continuously destroyed and ri-created, we may think that OH molecules are chemically created at upper level.

However, what we think is more efficient, is the following mechanism: the **three-level model**. The molecules are somehow pushed to the excited state (3) by a pumping mechanism. From this upper level (3), by spontaneous transition and emission, mostly in the infrared, they pass to an intermediate level (2) which represents the upper state of the maser transition. If the probability of transition between the two upper levels is much greater than that between the upper (3) and lower level (1), then the depopulation of the lower level (1) can be very strong to the advantage of overpopulation of the upper level (2), which is an intermediate metastable level with long life time. This means that electrons stay here for relatively long time (some seconds or a fraction of seconds) before going to level (1). The transition between the upper maser level (2) and the lower level (1) can occur spontaneously or by stimulated emission by photons.

It is well known that the stimulated emission by photons requires inducing photons with an energy identical to the energy difference between the two levels. From a single exciting (triggering) photon two identical photons are produced. They have same wavelength, same phase, same direction and polarization.

In the case of population inversion the probability of stimulated emission far exceeds that of absorption. As the two identical photons, emitted by the transition between the metastable level (2) and level (1), propagate through the emitting medium, the mechanism is amplified by stimulated, or induced emission. Moreover, since the emitted photons have the same characteristics as the photons that induce the transition, the stimulated radiation is coherent. This process in the laboratory is called laser, but maser in the case considered of emission in the interstellar medium in the radio field. In the figure 7.5 is shown the mechanism.

A maser in a machine that absorbed the radiation from a source and remits energy in a peculiar emitted lines. The maser mechanism is maintained until the energy pumping remains active. There are three main types of pumping.

1. Photon pumping, in which the upper states are populated by photon absorption.
2. Collisional pumping in which collisions populate the upper state. This type of pumping is often used in artificial lasers where collisions are due to electrons accelerated by electric fields.
3. Chemical pumping. In this case the molecules are preferentially created in excited states, including the upper ones responsible for the maser emission.

In general, however, none of the three mechanisms alone is capable of explain the observed transitions, even if in some cases it is clear that an intense IR flow excites the higher states of the maser.

A characteristic of maser radiation is that of being strongly polarized. The stimulated emission in fact releases photons that have the same polarization direction as the incident photon. This means that all the photons that derive, directly or indirectly, from the same progenitor, are all polarized in the same direction and therefore the radiation is 100% polarized.

The production of photons by stimulated emission, under certain conditions, can become higher than the re-population of the upper level by pumping. In this case the efficiency of the maser is strictly linked to the efficiency of the pumping mechanism and the amplification of the signal will be strongly limited. In this case we say that the maser is "saturated". The saturation effect can induce rapid apparent variations in the angular dimension of the sources because the amplification, and then the depopulation, is higher along the maximum length segment.

The excitation mechanism of the maser (i.e. the source of the photons which they trigger the stimulated emissions) may be due to the continuous source that also supplies the infrared photons of pumping (in this case the maser is produced exclusively in the direction source-cloud) or by spontaneous transitions. In the latter case the maser does not have a preferential direction (see the stellar maser).

The maser activity requires some fundamental conditions:

- a high optical depth of OH lines in order to have interaction between photons and atoms ($\tau_{OH} > 1$) but it should be low in IR otherwise the photons cannot go inside the cloud deep enough to overpopulate all the cloud ($\tau_{IR} \sim 1$)
- the gas density must be much higher than that of the interstellar spaces (at least $10^5\ atom/cm^{-3}$ while interstellar medium as a density of 1 $atom/cm^{-3}$) to have amplification. Indeed, if density is too low, particles don't interact;
- a high brightness source must be present ($L > 10L_{\odot}$) to provide pumping energy.

Amplification Therefore maser is a continuum mechanism with very low efficiency converting flux coming typically from a star into energy of a peculiar and very amplified line, such as OH. Indeed this line is very intensive but pay attention: the peak intensity doesn't correspond to a physical brightness temperature, which instead is around $10^{10} - 10^{12}\ K$, out of thermal equilibrium. It is just a peculiar amplification.

Moreover amplification is very peculiar for two reasons:

- it is an exponential amplification;
- it violates the transfer equation. We know that transfer equation is $\frac{I}{I_0} = e^{-\tau}$ where I is the output radiation, I_0 the original one and $\tau = kl$ the opacity coefficient (k is the absorption coefficient and l the length of the cloud). Since $\tau > 0$, I is always less than I_0. However, in the case of maser, it is the contrary: greater is the optical depth in the line, more intense is the output radiation. Therefore for masers we have a negative temperature T and negative absorption k.

7.6 Stellar and interstellar masers

There are two main types of maser sources in our galaxy: **interstellar masers** and **stellar masers**.

7.6.1 Interstellar masers

Interstellar masers are associated with regions of recent star formation and pumping energy can be provided by young stars type O-B, while the maser is triggered by compact condensations of interstellar material located near these stars.

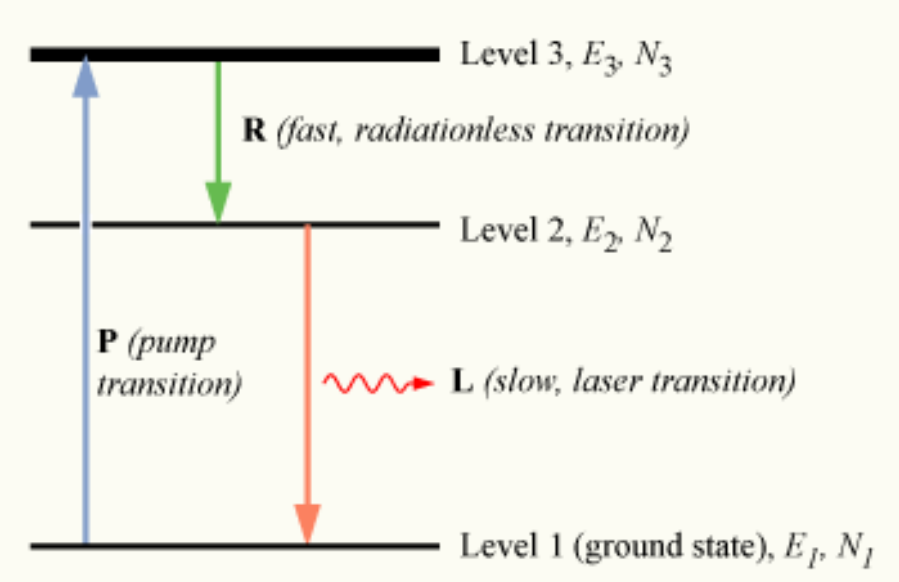


Figure 7.5: Consider a group of N atoms, each atom is able to exist in any of three energy states, levels 1, 2 and 3, with energies E_1, E_2, and E_3, and populations N_1, N_2, and N_3, respectively. We assume that $E_1 < E_2 < E_3$. Initially, the system of atoms is at thermal equilibrium, and the majority of the atoms will be in the ground state, i.e., $N_1 \sim N, N_2 \sim N_3 \sim 0$. If we now subject the atoms to light of a frequency $\nu_{13} = \frac{1}{h}(E_3 - E_1)$, the process of optical absorption will excite electrons from the ground state to level 3. This process is called pumping, and does not necessarily always directly involve light absorption; other methods of exciting are electrical discharge or chemical reactions. The level 3 is sometimes referred to as the pump level or pump band, and the energy transition $E_1 \to E_3$ as the pump transition, which is shown as the arrow marked P in the diagram on the right. Thanks to pumping an appreciable number of atoms will transition to level 3, such that $N_3 > 0$. To obtain the population inversion it is necessary that these excited atoms quickly decay to level 2. The energy released in this transition may be emitted as a photon (spontaneous emission), however in practice the $3 \to 2$ transition (labeled R in the diagram) is usually radiationless, with the energy being transferred to vibrational motion (heat) of the host material surrounding the atoms, without the generation of a photon. An electron in level 2 may decay by spontaneous emission to the ground state, releasing a photon of frequency ν_{12}, which is shown as the transition L, called the laser (or maser) transition in the diagram. If the lifetime of this transition is much longer than the lifetime of the radiationless $3 \to 2$ transition, the population of the E_3 will be essentially zero ($N_3 \sim 0$) and a population of excited state atoms will accumulate in level 2 ($N_2 > 0$). If over half the N atoms can be accumulated in this state, this will exceed the population of the ground state N_1. A population inversion ($N_2 > N_1$) has thus been achieved between level 1 and 2, and optical amplification at the frequency ν_{21} can be obtained.

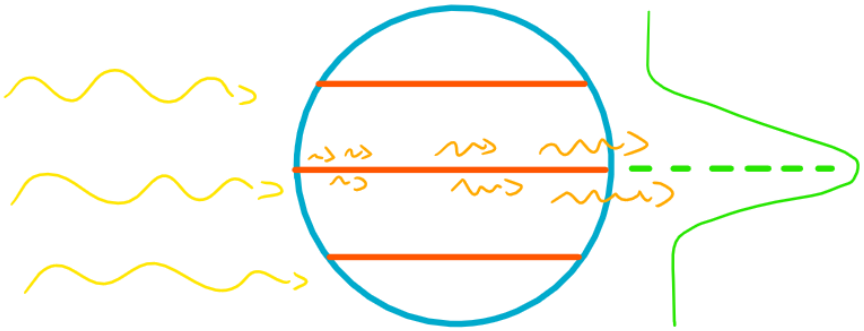


Figure 7.6: Interstellar masers.

Suppose to have radiation coming from an outside source like a young star, passing thought a sphere of interstellar matter (figure 7.6).

These photons are provided by radiation outside the cloud and therefore have a very precise direction. In this case, since the amplification depends on $\tau = kl$, at maximum length l (the direction source-cloud-observer), there is also the maximum amplification. All other segments are shorter so the amplification is less. The result is that the output intensity has a profile strongly peaked in the direction thought the center of the sphere and it looks like a very narrow source. In other words: this is an extreme case of limb darkening and it gives the appearance of a smaller angular size of the source due to this kind of amplification. This is true even the cloud is irregular.

Moreover, it happens that at the begging of the sphere, in the region much close to the star, there are few photons while approaching the other side, in the direction of observer, there are more photons due to the strong amplification. It happens that in this region you may have the "saturation": there are so many photons that the pumping is not efficient enough in keeping the inversion population because there are so many depopulating photons that it is not enough. As consequence, the line reduces the intensity, not at the edges, but at the center. As result, the size of this object changes very quickly. It is not an expansion of the cloud; it is am effect of variability of the saturation in the set of the maser. It is an apparent super-luminous effect due to the fact that winds remain the same while the peak change as well as the FWHM of the line.

The typical star in this system, as said before, is a young star having $L > 10L_{\odot}$ while the cloud must contain OH in rather narrow interval of density, not to high to avoid collisions, not to low in order to have quite high optical depth.

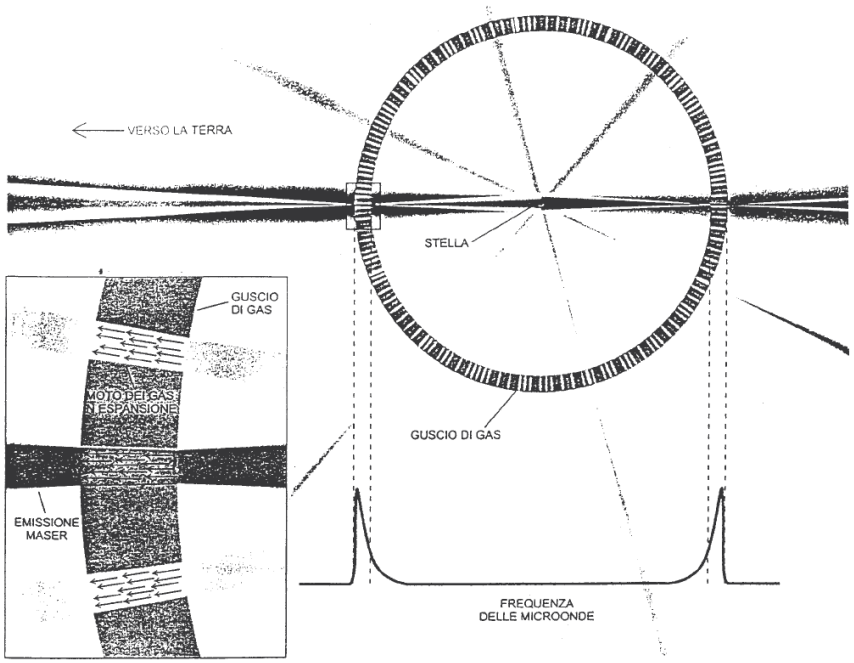
Summarizing, due to this strong amplification thought the center of interstellar cloud, the profile is very peaked so this type of masers are very good for accurate astrometry, still in radio (which means still to the GC). Then, combining with measurements of proper motion, we can get the distance up to $8 - 10$ Mpc, so external object, with a method that is purly geometric.

7.6.2 Stellar masers

In stellar masers, instead, the emission takes place in a different way: the energy source is the star itself, usually a giant or super-giant star surrounded by big shell of expanding material at low velocity $(10 - 15\ km/s)$. The size of the shell is much more bigger than the size of the star but it has a delimited thickness.

The interstellar and stellar masers differ substantially in the trigger mechanism, that is in the origin of the photons responsible for the stimulated emission.

In interstellar masers these photons are provided by radiation outside the cloud and therefore have a very precise direction, while the stellar ones are usually triggered by spontaneous transitions within the circumstellar shell and can therefore radiate in all directions.



La struttura di un guscio contenente maser a ossidrile intorno a una gigante rossa è mostrata schematicamente in sezione. Dato che l'amplificazione si verifica solo all'interno di una regione dove il gas si muove più o meno alla stessa velocità (*nel riquadro*), quasi tutta l'emissione maser si sviluppa lungo linee radiali; dalla Terra si vede perciò solo quella proveniente da due piccole zone del guscio. L'emissione maser di aree diverse del guscio si distingue a causa dello spostamento Doppler; il caratteristico doppio picco è un segno distintivo delle giganti rosse, rilevabile anche se la stella è invisibile ai telescopi ottici.

Figure 7.7: Geometry of a stellar maser.

In stellar masers the radiation that reaches the observer comes only from two narrow caps located in the star-observer direction (figure 7.7). This because we said that the photons must have the right energy to stimulated the emission, the exactly energy. We know that due to Doppler effect, the energy of a photon change due the composition of the motion between the photon incoming and the motion of the atom targeted in the same direction. At the edges of the shell, from our point of view, the composition change in any interaction because the atoms or molecules of the shell move radially as the shell expands. So the energy along our line of sight changes and the stimulated emission is blocked. Therefore we can observe maser emission only along the line of sight connected to the center of the shell.

We can also see the lines coming from the behind side because the photons for the stimulated emission are coming from the spontaneous decay and not from the star. So, the spontaneous emission generates other masers over all the directions. The photons from the behind side can reach us also because the star is much smaller than the maser size.

In general, the spectrum shows a characteristic splitting of the lines due to shell expansion. The two components correspond to the cap located between the star and the observer (component shifted to a shorter wavelength) or to the cap located on the side opposite the star, for the component shifted to greater wavelengths (figure 7.8). The average of the two wavelengths gives the motion of the star and of the envelope as a whole.

In stellar masers the star is a long-term asymptotic branch variable, with a luminosity of the order of $10^4 L_\odot$ and temperatures around 2000 K. The splitting of the lines, of $10 - 50$ km/s, is easily appreciable. The diameter of the emitting region is up to about $D = 10^16 cm$ (the astronomical unit is about $1.5 \cdot 10^{13}$ cm). There are also supergiants with infrared excess, such as NML Cygni or VY Cma, of $10^5 - 10^6 L_\odot$. The star is often a LPS (long period star), a variable star, with a period of $100 - 200$ days therefore the pumping is variable according to the variable emission of the star.

Distance measure with stellar maser We know that the diameter of the shell around the star has a size of several AU. Therefore, even if the radiation from star allows pumping increase in both direction at the same time, fro the observer the amplification from the cup below the star, on the

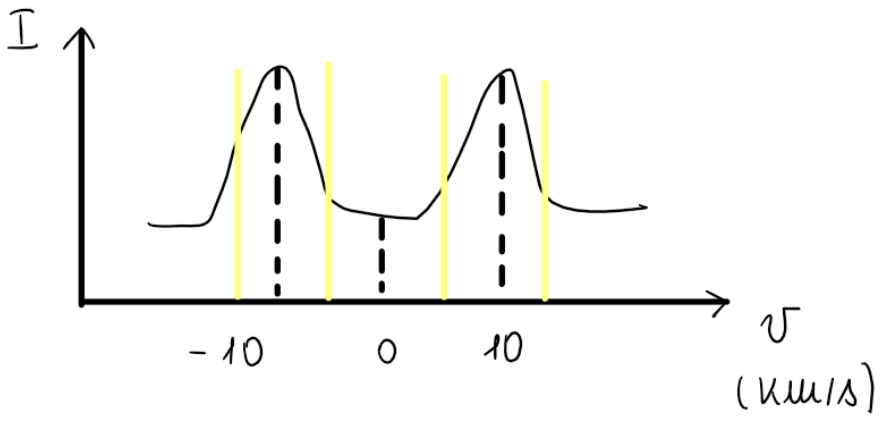


Figure 7.8: Intensity versus velocity of stellar masers.

opposite side, takes more time to arrive due to the diameter of the cloud. As result, we measure a non-synchronous variation. The difference in time is equivalent to the time necessary to photons to pass all the shell. Therefore, knowing the speed of light, we can get the true diameter of the sphere. Then, measuring by interferometry the apparent size, we can get the real distance using a purely geometric method.

This technique is very useful: in this way we can measure distance stars thought the disk because measurements are in radio so there is no absorption. So this method can be used to to map the kinematics of galactic disk as function of the distance from the GC. This is fundamental to study black matter.

Other chemical elements in stellar masers The H_2O emissions at 1.4 cm derive from rotational transitions. The H_2O maser emission is often associated, but not always, with that OH. Unlike the latter, the H_2O emission source is generally smaller than the OH and the velocity structure is different. The intensity variations are more evident. The radiation is often polarized and the sources are compact and variable in a time scale of a few weeks. Brightness temperatures are very high. The very low Doppler effect and temperatures are reminiscent ofOH emissions and suggest that these lines are also produced by stimulated emission. Finally we recall the importance of the maser emission between rotational states of SiO at 7 mm, and of methanol (CH_3OH) at 1.2 cm, discovered in the direction of a region of active star formation in Orion. Many other molecules show population inversion, but do not have sufficient optical depth to produce an effective maser mechanism.

7.7 Usefulness of masers

The small angular size of the masers makes them suitable also for high precision astrometric measurements. Exploiting the expansion motions of the various sources of a cloud, or the statistical parallaxes of a cluster, or even the proper motion due to the galactic rotation, it is possible to obtain geometric distance measurements on a galactic and extragalactic scale.

The OH maser allowed accurate measurements of the Sun-galactic center distance. Simultaneous measurements of radial velocity and proper motions in the Sagittarius complex have given a distance of $7.1 \pm 1.5 Kpc$, shorter than that derived from the majority of optical astronomy methods (about 8 Kpc). The reason of this discrepancy is not clear. The measurements made in the Orion complex gave a distance of 480 Kpc, is, instead, in excellent agreement with the optical data and thus demonstrating the reliability of the method.

The prospects for direct measurement of extragalactic distances are exciting. Using the rotation speeds in galaxies and proper motions, geometric distances were measured up to 10 megaparsecs, thus including the galaxies of the local group.