Extra-Solar Planets

Lecture, D-PHYS, ETH Zurich, Spring Semester 2014

Friday: 8:45 - 10:30, HIT J52, Hönggerberg

Exercises: Friday: 10:45 - 11:30

Dates: 21.02.; 28.02.; 07.03.; 14.03.; 21.03.; 28.03.;
04.04.; 11.04.; 02.05.; 09.05.; 16.05.; 23.05.; 30.05.

Website: www.astro.ethz.ch/education/courses/Extrasolar_Planets

Lecturer: Prof. Dr. H.M. Schmid,
Office, HIT J22.2, Tel: 044-63 27386; e-mail: schmid@astro.phys.ethz.ch

Teaching Assistant and Co-Lecturer:

Dr. Kamen Todorov,
HIT J23.2, Tel: 044-63 30661; e-mail: todorovk@phys.ethz.ch

ETH Zurich, Institut für Astronomie,
Wolfgang Pauli Str. 27
ETH-Hönggerberg, 8093 Zurich

Contents

Chapter 1

Introduction

1.1 Planets and the Universe

The fact that we exist indicates that we must live in a Universe with worlds that can harbor life. This is the basic statement of the antropic principle. We live on a terrestrial planet in a planetary system and it seems very likely that this situation is a very favorable one for harboring life. The properties of planetary systems and planets and the search for signatures of life is the astronomical aspect of our quest on the origin and place of life in the Universe. There is also an important biological aspect addressing the conditions required that life can emerge in a system.

This lecture concentrates on the physical properties of planetary systems and the processes which are important for the formation and evolution of planets. Another strong focus is set on observational data which provide the basic empirical information for our models and theories of planetary systems.

The place of the planet Earth in the Universe can roughly be described as follows. Earth resides since 4.6 billion years in the solar system. Our sun is just a kind of normal star among 100 billions of stars in the Milky Way, which is itself a quite normal spiral galaxy among billions of galaxies in the observable Universe. The galaxies were essentially born by the assembly of baryonic matter in the evolving potential wells of dark matter concentration in an expanding Universe. This process started about 14 billion years ago with the big bang.

Terrestrial planets consist of heavy elements. These were "cooked" from the light elements H and He originating from the big bang by nuclear processes in previous generations of intermediate and high mass stars. Stars form through the collapse of dense, cool interstellar clouds. Then they evolve due to nuclear reactions until they expel a lot of their mass at the end of their evolution in stellar winds or stellar explosions (supernovae). This matter, enriched in heavy elements by the nuclear processes, goes back to the interstellar gas in the Milky Way disk and may form there again a new generation of stars.

Planets form in circumstellar disks around new-born stars. For this reasons the planet properties depend a lot on the parent star and on the environment in which they were born. The following diagram (Fig. 1.1[1]) gives a very general overview on the place of planets in the Universe, especially within the stellar evolution cycle taking place in galaxies.

[1] to be completed by the reader according to the discussion in the lecture

.

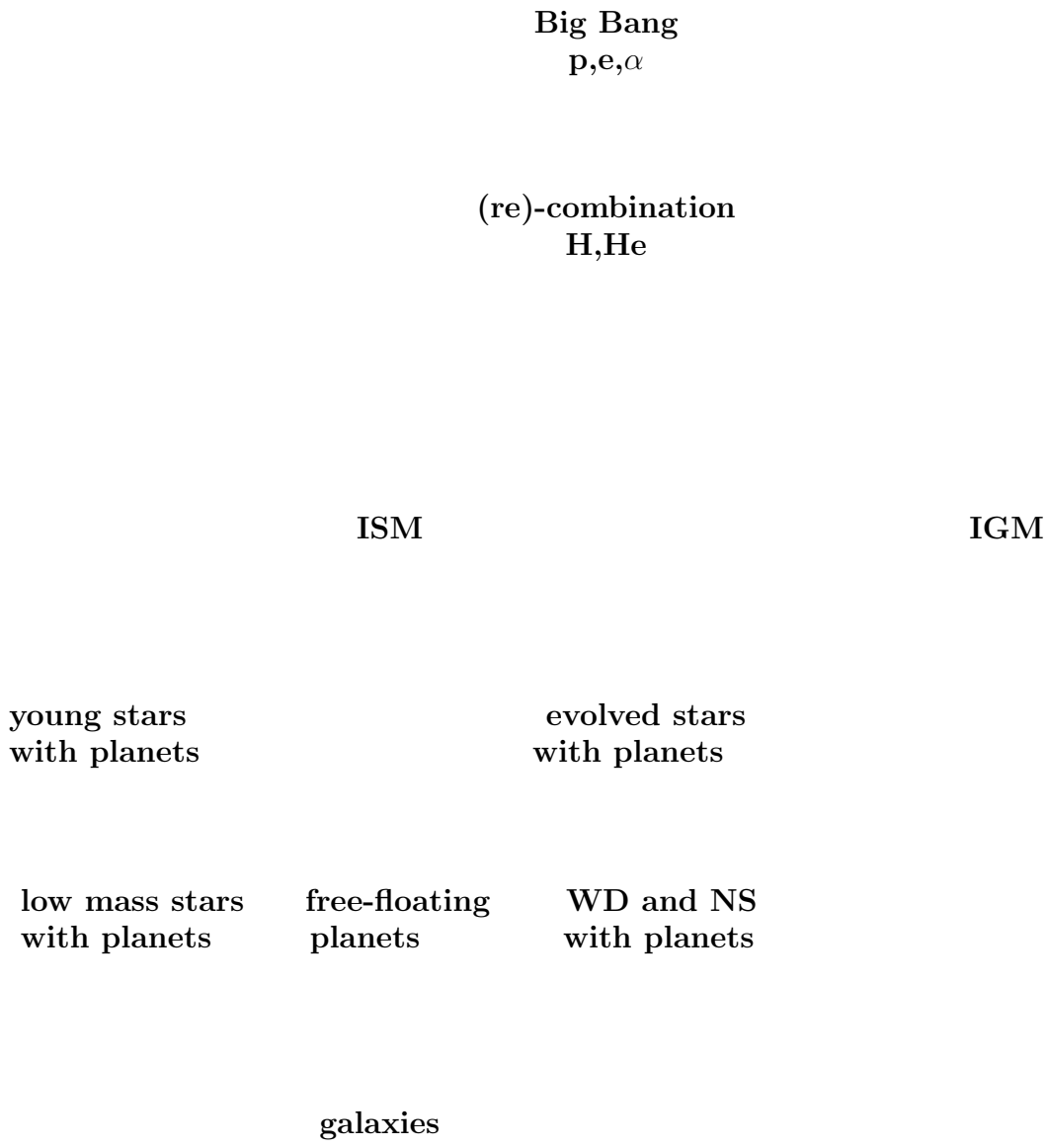

Figure 1.1: The position of the planets within the stellar evolution cycle in galaxies and in relation to the evolution of the baryonic matter in the Universe. ISM stands for interstellar matter, IGM for intergalactic matter, WD for white dwarfs, and NS for neutron stars.

1.2 The solar system

The solar system has been studied in much detail with many different types of investigations ranging from geological studies, meteorite analysis, ground-based observations of solar system bodies, visits by Astronauts on the moon, and visits to other objects with sophisticated robotic systems. We will never be able to study extra-solar system in such detail. The solar system is therefore a most important source of information for extra-solar planetary systems. In particular many physical models and properties which apply to the solar system are also valid for extra-solar systems. On the other hand the models and theories for extra-solar system must also apply in some way to the solar system. For this purpose the solar system provides a number of well studied examples of planets and other objects which serve as laboratories and test cases for the interpretation of not well understood properties of poorly observed extra-solar systems.

In this lecture we use the solar system as example for basic physical principles and for the characterization of typical objects in planetary system. The discussion concentrates on topics which are most relevant for extra-solar planet research. We neglect many aspects, like surface morphologies, small scale structures, or planetary magnetism which are important topics in solar system research but completely unconstrained for extra-solar planetary systems because no observational data are available today and will not be available in the near future.

Objects in the solar system

The sun dominates the solar system with its gravitational potential (> 99 % of the total mass) and its energy generation by nuclear reactions. Beside the sun, this system contains objects of very different size, composition and physical properties:

- sun: central star of spectral type G2V, age 4.5 Gyr, and quite normal metallicity when compared to other stars,
- 8 planets: 4 terrestrial planets, Mercury, Venus, Earth and Mars, and 4 gas giants, Jupiter, Saturn, Uranus and Neptune (see Slide 1.1),
- several dwarf planets: Eris, Pluto, Makemake, Haumea, Ceres, and perhaps some more,
- moons and satellites of planets and dwarf planets,
- small bodies of the solar system including asteroids, trans-neptunion objects (TNOs), comets, meteorites, and dust.

The objects have a characteristic distances to the sun as indicated in Figure 1.2.

Figure 1.2: Distance of solar system objects to the sun.

Planets. There exists a well established definition for the planets in the solar system. Planets are in orbits around the sun. They are large enough for their self-gravity to overcome rigid body forces. Thus, planets are round except for some rotational flattening. Further, they have cleared the neighborhood of their orbit.

Dwarf planets differ from planets because they have not cleared the neighborhood of their orbit. Thus, they are large and round, but they are not a moon of a planet. Ceres is the only dwarf planet in the asteroid belt because it is the only asteroid which is large enough to be round. Pluto is the most famous dwarf planet in the Kuiper belt, but there are others of similar size out there (Eris, Makemake, Haumea) and perhaps even more which have not been detected yet.

Moons are in orbit around planets and dwarf planets. They can be as large as dwarf planets. Titan and Ganymede are even larger than the planet Mercury, but they have a smaller mass. The satellites around planets span a large range down to the size of large rocks ($R \approx 100$ m).

Asteroids are rocky bodies, located mainly in the (main) asteroid belt between Mars and Jupiter. The largest asteroid is the dwarf planet Ceres, then comes Pallas, Vesta (Slide 1.8) and about 200 more with $R > 100$ km. There are much more small asteroids (Slide 1.9) down to the size of a few centimeters, which are called meteorites if they fall onto Earth. More than 100 000 asteroids are known and it is expected that there are of the order 1 million asteroids with a diameter of 1 km or larger. Their radius distribution $N(R)$, the number of asteroids per unit size interval, can be described by a power law

$$N(R) = N_0(R_0) \left(\frac{R}{R_0} \right)^{\alpha}$$

The exponent is about $\alpha = -3.5$ with a normalization value of $N_0(10\,\mathrm{km})/dR \approx 1000/\mathrm{km}$. This implies that most of the mass is in the largest bodies and most of the surface area is in the smallest bodies. The total mass in the asteroid belt is estimated to be about 0.1 % of the mass of the Earth $\mathrm{M_E}$. There are many different subgroups and types of asteroids which are classified according to their orbits and composition.

Transneptunian objects (TNOs) are all objects in the solar system that orbit the Sun with an average distance larger than Neptune. Pluto is the first detected TNO. In the meantime more than thousand TNOs were found with radii in the range 20-1000 km. TNOs are classified in different groups, like Pluto and the plutinos which are in a 2:3 resonance orbit with Neptune, Kuiper belt objects located at a distance of 30 - 55 AU and scattered disk objects further out. It is expected that there are more than 100 000 Kuiper belt objects with diameters larger than 100 km. The total mass of the Kuiper belt is of the order 0.1 $\mathrm{M_E}$ or about 100 times the mass of the asteroid belt.

Comets are icy objects which start to evaporate when they come close ($d < 3$ AU) to the sun (Slide 1.11). The comet forms then a coma (extremely tenuous atmosphere) and a tail due to the solar wind and the solar radiation pressure. Typically a bright comet has a nucleus with a diameter of about 1 to 10 kilometers. There are about 500 short period $p < 200$ yr known. Many comets where observed only once because they originate from the very distant solar system.

Meteorites are rocks which have fallen onto Earth. They are mostly debris pieces from collisions between asteroids but there are also meteorites which are associated to debris from impacts of asteroids on the Moon or Mars. Stony meteorites originate from the

mantel or crust while iron meteorites are from the core of differentiated parent bodies (Slide 1.10). A third group are the primitive meteorites or chondrites which are condensed directly from the solar nebula and not much changed in composition since then. Meteorites are important because they can be analyzed in laboratories. For example, the chondrites provide very good abundances for many elements and isotopes in the solar system.

Zodiacal dust are interplanetary particles in the ecliptic plane with a size between a fraction of a μm to a few mm. These dust particles originate, like the larger meteorites, from collisions but also from evaporating comets. This dust produces the zodiacal light, a forward scattering effect near the sun, and the "gegenschein" a backward scattering effect in the anti-solar direction (Slide 1.12). In addition the dust is heated by the solar radiation and produces a strong emission in the mid-IR spectral region. Meteors ("Sternschnuppen") are large zodiacal dust particles (mm - cm in size) which penetrate into the Earth atmosphere, are heated up by friction and evaporate before they reach the ground.

1.3 Short history of the detections of extra-solar planets

The development of the research field on extra-solar planets can be well illustrated with the evolution of the known planets in the mass - separation diagram (see Slides 1.13 to 1.7). The following table gives a chronology of the most important detections.

Table 1.1: Chronology of important detections in extra-solar planet research.

year	important detection or event
1992	Two planets around the pulsar PSR 1257+12 are discovered (A. Wolszczan and D. Frail) based on the timing measurements of the periodic variation of the pulsar's radio pulses. It is not clear whether these planets were formed by the supernova (SN) explosion or whether these are the remaining cores of gas giants which lost there atmosphere/envelope due to the SN event.
1995	Detection of 51 Peg b, the first extra-solar planet around a sun like star by M. Mayor and D. Queloz (Geneva Obs.) by measuring the periodic radial velocity variations of the steller reflex motion. For many astronomers it was a big surprise that there could exist Jupiter-like planets in a 4 day orbit. The surface temperature of this planet must be $T > 1000$ K because of the strong irradiation. These objects are now called "hot Jupiters".
1999	Detection of the first transiting planet HD 209458 b. The transit confirmed that the close-in planets have the size of gas giants. The combination of RV-measurements and transit depth provides an estimate of the bulk density of the planet.
2000	About 40 planets are detected with the RV-method (see Slide 1.14).
2003	The HARPS high precision RV-spectrograph starts operation at the ESO 3.6m Telescope. This instrument is more accurate than previous instruments and capable to search for Neptune-like planets, super-earths and even terrestrial planets in tight orbits.
2004	The OGLE-team announced a first convincing case of a micro-lensing event by a star-planet system. The mass ratio between the low mass star $M < \mathrm{M}_\odot$ and the planet is about 250 : 1. This yields a mass of about $\approx \mathrm{M}_J$ for the planet. The system is at the distance of a few kpc in the direction of the lensed background star located in a dense stellar field near the center of the Milky Way.
2005	There are about 160 RV-planets known, about 12 transiting planets, and 3 planets were seen in micro-lensing events. Most planets known are giant planets with a mass like Saturn or larger (see Slide 1.16).
2007	The SPITZER satellite determines the phase curve of the transiting hot Jupiter HD 189733 b which allows to determine the temperature difference between the illuminated hot and the non-illuminated cold side of the planet. The data show further that the hottest point is offset from the sub-stellar point indicated strong atmospheric circulation.

2008	Direct detection of 3 planets around the A-star HR 8799 and 1 planet around the β Pic with high contrast imaging. These are young systems with quite luminous, still contracting planets with temperatures of about 1000 K.
2010	15 year after the detection of 51 Peg b there are about 360 RV-planets, about 110 transiting planets, 10 micro-lensing planets, 5 pulsar planets, and 5 directly imaged planets known. The detection limits for RV-planets and transiting planets are pushed significantly towards terrestrial planets (Slide 1.4).
2011	First results from the Kepler-satellite are presented. This instrument surveys about 150'000 stars for planetary transits. Kepler finds within a few months more than 1000 planet candidates with this method. Many of the candidate planets have radii of 2 R_E or even smaller. There are also many systems with multiple planets detected. The system Kepler-11 is the most extreme case with 6 transiting planets.

From the compilation in Table 1.1 it becomes obvious that extra-solar planets are studied with several different detection methods which are listed in Table 1.2. The table does not list characterization techniques, like transit spectroscopy, which will in general follow the initial detection. The main techniques for planet detection are currently the radial velocity (RV) searches for reflex motion and transit searches. It is expected that in about 5 years there will be many, of the order hundred, successful detections of planets with astrometry and IR high contrast imaging.

Table 1.2: Planet detection methods and order of magnitude numbers for successful detection until 2013 of gas giants and terrestrial planets.

method	detected planets		comment
	giants	terr.	
reflex motion			
radial velocity (RV)	> 100	> 10	established
pulsar timing	< 10	< 10	only for special systems
transit timing	> 10	≈ 10	KEPLER multiplanet systems
binary eclipse timing	< 10	< 10	only for special binaries
astrometry	< 10	0	great potential with GAIA
transits			
transit light curves	> 100	> 100	≈ 1000 KEPLER candidates
microlensing			
light curves	≈ 10	< 10	measurements not repeatable
astrometric effects	0	0	future technique
direct imaging			
infrared (IR)	≈ 10	0	great potential
optical	0	0	difficult
other methods			
radio waves from planets	0	0	expectations unclear

1.4 What is a planet?

The definition of a planet in the solar system is well established. However for extra-solar objects there exists often only very limited information and there are different kinds of definitions for substellar objects. In the following we give some definitions.

Star: This is an object which has an extended main-sequence phase during which the radiated energy originates from stable nuclear hydrogen burning. This phase lasts $>$ Gyr for low mass stars.

Substellar object: The mass of a substellar object is too low to burn hydrogen in equilibrium because the temperature and pressure in the core do not reach high enough values during the gravitational contraction of a low mass gas sphere. The mass limit between stars and substellar objects is about $0.08\,\mathrm{M}_\odot$.

Brown dwarf: This is a substellar object which cannot burn hydrogen in the core but still has some significant phase of nuclear energy generation due to deuterium- or lithium burning. This includes objects in the mass range $0.015 - 0.08\,\mathrm{M}_\odot$ $(15 - 80\,\mathrm{M}_\mathrm{J})$.

Planets: Planets are substellar, spherical object which were formed in a circumstellar disk.

This definition of planets is based on the formation process. According to this definition a substellar object which formed like a star out of a gas cloud, is not a planet, even if its mass is similar to Jupiter. On the other side there could exist deuterium-burning brown dwarfs which were born in a disk and are called planets. Also free-floating objects, which were formed in a circumstellar disk but where then ejected out of their circumstellar orbit by some dynamical interaction, are also classified as planets.

An alternative definition for planets is based on the explicit mass limit, e.g. $< 15\mathrm{M}_\mathrm{J}$, for object without nuclear burning. This definition allows a simple classification according to a mass determination. But, this does not consider fundamental differences in the formation process which have most likely also a very strong impact on the physical properties of the object.

1.5 Contents and literature

Content of this lecture: Important topics covered by this lectures are:

- properties of planets of the solar system,
- detection techniques,
- interpretation of the observed extra-solar planet properties,
- difference between low mass stars, brown dwarfs, and giant planets,
- theory for the formation and evolution of extra-solar planets,
- search for life.

Textbooks on the solar system:

- *Planetary Sciences.* I. de Pater, J.J. Lissauer. 2001, Cambridge University Press
  Comprehensive treatment of many aspects of the physics in the solar system.

– *The Solar System.* T. Encrenaz, J.-P. Bibring, M. Blanc. Second edition 1995, Springer
A good overview on the solar system.

– *The Earth as a Distant Planet.* M. Vázquez, E. Pallé, P. Montañés Rodriguez. 2009, Springer
A textbook with a very special focus on Earth including a unique compilation of Earth observations and data.

Textbooks on extra-solar planets:

– *Exoplanet atmospheres.* S. Seager. 2010. Princeton University Press.
Very careful and useful description of the physical processes for extra-solar planetary atmospheres.

Review articles or collection of review articles on extra-solar planets: The review articles provide usually more detailed and more actual information on specific topics with the drawback that they are often more rapidly outdated than textbooks.

– *Extra-solar planets.* P. Cassen, T. Guillot, A. Quirrenbach, Saas-Fee Advanced Course 31, 2006, Springer.

– *Exoplanets.* S. Seager (eds.) and 38 authors. 2010. University of Arizona Press.
A very good starting point for many topics in extra-solar planet research.

Textbook on special topics related to this field

– *Stellar Structure and Evolution.* R. Kippenhahn, A. Weigert. 1990. Springer.
The standard textbook on stellar structure.

– *New Light on Dark Stars.* I.N. Reid, S.L. Hawley. 2005. Springer.
A careful and comprehensive treatment of all aspects of low mass stars and brown dwarfs.

On-line sources:

– **http://www.exoplanets.eu**
On-line catalog of known extra-solar planets. This source includes tools for the selection and statistical analysis of exoplanets.

– **http://adsabs.harvard.edu/abstract_service.html**
NASA astrophysics database system. Essentially all scientific articles on extra-solar planets are available through this source. Many are freely available. Essentially all articles are available from an ETH account.

– **http://www.ency-astro.com/eaa/eaa/eaa/index.html**
'Encyclopedia of Astronomy and Astrophysics'

Chapter 2

Reflex motion: masses and orbits

The mass is a most important parameter for the characterization of an astronomical object. This is also the case for planets. For the planets in the solar system one finds with decreasing mass: giant planets Jupiter and Saturn, the "ice" giants Neptune and Uranus, the terrestrial planets Earth and Venus with substantial atmosphere, then Mars with only a thin atmosphere and finally the bare, rocky planet Mercury without atmosphere. This anti-correlation between planet mass and extent of the "gaseous" envelope may be also common for extra-solar planets.

The orbit is another key property for a planet. The orbital distance to the star defines largely its surface temperature. The orbital eccentricity, inclination, and the mutual orbit geometries in multi-planet systems provide essential information about the dynamic properties and evolution of a system.

All this information is available without "seeing" the planet, but by measuring the reflex motion of the bright parent stars induced by the planets. The determination of masses and orbits based on the reflex motion of unseen companions was introduced more than 100 years ago for binary stars. But because planets are much less massive than stars, the required measurement accuracy for extra-solar planetary systems is extremely demanding and could only be realized with modern technologies. Once, detections of the reflex motion were possible, the investigation of extra-solar planetary systems became immediately a major new field in modern astronomy, and a substantial fraction of all observational information on extra-solar planets available today is based on this technique. Mass and orbit determinations via the stellar reflex motion will remain a backbone for extra-solar planet research in the near future. Therefore we discuss these methods in much detail.

2.1 Kepler's laws and the stellar reflex motion

2.1.1 Kepler's laws

Kepler formulated three generals laws for the orbits of the solar system planets based mainly on the astrometric data from Tycho Brahe:

1. The planets move on elliptical orbits with the sun located in one of the focal points,
2. the line connecting the sun and a planet sweeps the same area in equal time intervals (Flächensatz),

3. The squares of the orbital periods of the planets are proportional to the cubes of their semi-major axis.

Kepler's laws follow also from the 2-body mechanics based on Newton's theory of gravitation for the limiting case $M_\odot \gg m_P$. They are therefore valid for all 2-body problems where one body dominates in mass, for example for moons around planets. We summarize here important implications of Kepler's laws. The full derivations of the formula given in this section can be found e.g. in standard textbooks on mechanics or comprehensive physics textbooks.

Figure 2.1: Geometry of an elliptic orbit.

First Kepler law (elliptical orbits):

– The eccentricity ϵ characterizes the elliptic orbit $0 \leq \epsilon < 1$. The major axis a and minor axis b of the ellipse are related by

$$b = a\sqrt{1-\epsilon^2}\,. \tag{2.1}$$

– The location of the sun in the focal point F is offset from the ellipse center by the distance $a \cdot \epsilon$ on the major axis. The perihel $r_{\min}$ and aphel $r_{\max}$ distances are

$$r_{\min} = a\,(1-\epsilon) \quad \text{and} \quad r_{\max} = a\,(1+\epsilon)\,.$$

– For circular orbits there is $\epsilon = 0$ and $a = b = r_{\min} = r_{\max}$.
– For $\epsilon \to 1$ the orbit approaches a parabola ($\epsilon = 1$) which is not a closed orbit.

Second Kepler law (Flächensatz):

– The second Kepler law is equivalent to the conservation of angular momentum $\vec{L} = m_P \vec{v} \times \vec{r}$

$$\Delta A = \frac{1}{2}\,|\vec{x} \times \vec{v}|\,\Delta t = \frac{1}{2m_P}\,|\vec{L}|\,\Delta t\,,$$

where ΔA is the swept area per time interval Δt.

– The second Kepler law can also be written as

$$\frac{\Delta A}{\pi a b} = \frac{t - T_0}{P}\,, \tag{2.2}$$

where T_0 is the time of perihel passage and P the orbital period. From this relationship one can derive numerically (not analytically) the "central" orbital angle E from the mean orbital anomaly $M = E(t) - \epsilon \sin E$ and then the orbital phase angle as function of time $\phi(t)$ from

$$\tan\frac{\phi(t)}{2} = \sqrt{\frac{1+\epsilon}{1-\epsilon}}\tan\frac{E(t)}{2} \quad \text{where} \quad E(t) - \epsilon\sin E(t) = \frac{2\pi}{P}(t - T_0)\,.$$

The radial distance is defined by the orbital phase angle and the eccentricity

$$r(\phi) = a\frac{1-\epsilon^2}{1+\epsilon\cos\phi}\,, \tag{2.3}$$

– The orbital velocity has a maximum at perihel and a minimum at aphel

$$v_{\max} = \frac{2\pi a}{P}\sqrt{\frac{1+\epsilon}{1-\epsilon}} \quad \text{and} \quad v_{\min} = \frac{2\pi a}{P}\sqrt{\frac{1-\epsilon}{1+\epsilon}}\,,$$

which becomes of course a constant orbital velocity $v = 2\pi a/P$ for a circle ($\epsilon = 0$).

Third Kepler law:

– The third Kepler law states that $a^3/P^2 = \text{const}$ for the solar system planets. The constant is directly related to the mass of the sun and therefore the third Kepler law is a most important tool for the determination of the the mass M of an astronomical object by measuring the separation and the orbital period of a small companion $m \ll M$:

$$\frac{a^3}{P^2} = \frac{G}{4\pi^2}M\,, \tag{2.4}$$

where G is the gravitational constant.

– For circular orbits the velocity $v_{\rm circ}$ is related to a and P by $v_{\rm circ} = 2\pi a/P$ so that the above relation can be written as

$$v_{\rm circ}^3 \cdot P = 2\pi G\,M\,.$$

indicating that the mass of the central object can also be determined by measuring the orbital velocity $v_{\rm circ}$ and the period P.

– The third Kepler gives also the orbital velocity of small objects around a central mass M as function of distance. For circular orbits $P = 2\pi a/v_{\rm circ}$ there is

$$v_{\rm circ} = \sqrt{\frac{G\,M}{a}} \propto \sqrt{\frac{1}{a}}\,. \tag{2.5}$$

– The third Kepler law follows also from the force equilibrium $-F_G = F_Z$ for a circular orbit, where the gravitational force is $F_G = -GMm/r^2$ and the centrifugal force $F_Z = m\omega^2 r$ (where $r = a$ and the angular velocity $\omega = 2\pi/P$).

2.1.2 Generalization for a two body problem

Kepler's laws follow from Newton's two body problem with M_1 and M_2, for the special case where $M_1 = M \gg m = M_2 \to 0$. Kepler's law can be generalized for the two body problem with $M_2 > 0$:

- The exact formula for Kepler's third law for a 2-body system is

$$\frac{(a_1 + a_2)^3}{P^2} = \frac{G}{4\pi}(M_1 + M_2)\,. \tag{2.6}$$

- The individual orbits of the two masses M_1 and M_2 are ellipses with eccentricity ϵ and semi-major axis a_1 and a_2 respectively, with the center of mass at the focal points. The total semi-major axis a of the orbit of M_2 with respect to M_1 is

$$a = a_1 + a_2\,.$$

- The semi-major axis of the two ellipses behave like

$$M_1 a_1 = M_2 a_2\,, \tag{2.7}$$

  according to the lever rule (Hebelgesetz). The radial distances and orbital velocities behave accordingly

$$M_1 r_1(\phi) = M_2 r_2(\phi) \quad \text{and} \quad M_1 v_1(\phi) = M_2 v_2(\phi)\,.$$

2.1.3 Orbital elements

Seven quantities are necessary for the definition of the elliptic orbit of an object in a two body system. If the orbit of one object M_1 or M_2 is given relative to the center of mass then the ellipse has the semi-major axis a_1 or a_2 respectively. If the orbit of one object is given relative to the position of the other object then the semi-major axis is $a = a_1 + a_2$. One object needs to be identified for the definition of the position angle for the orientation of the orbit, because the other object has an orbit which is offset by 180° in orbital phase with respect to the first object.

Four parameter describe the elliptical orbit:

- P: Orbital period,
- ϵ: eccentricity of the ellipse,
- T_0: time of periastron (or perihel) passage,
- a_1, a_2, or a: semi-major axis of the ellipse of component 1, component 2, or of one component relative to the other component.

Three quantities describe the orientation of the elliptical orbit on the sky:

- Ω: position angle of the line of nodes for the orbital plane and a reference plane with respect to a reference point. In the case of a solar system object this is the ecliptic plane and the "March" equinox. For a stellar system this is the sky plane and the position angle with respect to N (measured from N over E).
- i: inclination of the orbital plane with respect to the reference plane (ecliptic or sky plane)
- ω: Angle between the line of node and the periastron (or perihel) passage. For a stellar system this can be defined e.g. by the angle from the line of node passage with positive radial velocity (red shift).

Figure 2.2: Illustration for the definition of the orientation of the elliptical orbit.

2.1.4 Orbital parameters for the solar system planets.

Table 2.1 gives orbital parameters for the solar system planets.

Table 2.1: Masses and orbital parameters for solar system planets. P is the orbital period, a the semi-major axis, v the mean orbital velocity, ϵ the eccentricity, and i the inclination with respect to the ecliptic plane.

planet	M [M_J]	P [yr]	a [AU]	v [km/s]	ϵ	i
Mercury	0.00017	0.241	0.387	47.9	0.206	$7°00''$
Venus	0.0026	0.670	0.723	35.0	0.007	$3°24''$
Earth	0.0031	1.00	1.00	29.8	0.017	$0°00''$
Mars	0.00034	1.88	1.52	24.1	0.093	$1°51''$
Jupiter	1.0	11.9	5.20	13.0	0.048	$1°19''$
Saturn	0.299	29.5	9.55	9.64	0.056	$2°30''$
Uranus	0.046	84.0	19.2	6.80	0.046	$0°46''$
Neptune	0.054	165.	30.1	5.43	0.009	$1°47''$

$M_J = 1.90 \cdot 10^{27}$ kg, 1 year = 365.25 days, 1 AU = $1.50 \cdot 10^8$ km

2.1.5 Expected reflex motion due to planets

The reflex motion of a star due to orbiting planets follow from Newton's mechanics. The star will circle around the center of mass of the planetary system and mirror the motion of the planets. This motion can be measured with three different techniques:

- **Radial velocity**: periodic radial velocity variation of the star with respect to the line of sight, which can be measured with high precision spectroscopy,
- **Astrometric motion**: periodic motion of the star around the center of mass in the sky plane which can be measured with high precision astrometric techniques,
- **Timing variation**: arrival time variations of a well defined periodic signal due to light path variations or changes in the system configurations.

The expected motion of the sun as seen from a distance of 10 pc is shown in Slide 2.1 and Table 2.2 gives the values induced by Jupiter and Earth. The solar motion is dominated by the reflex motion due to Jupiter which has a period of 12 years. The typical amplitude (radius) of the circular motion is about 0.005 AU $= 7.5 \cdot 10^5$ km ($0.001''$ at 10 pc is 0.01 AU). This orbit is comparable to the radius of the sun.

Table 2.2: Typical values for observables for the reflex motion for Jupiter-like and Earth-like planets (mass and orbit) around a solar mass star.

observable	Jupiter	Earth	meas. limit (2013)
radial velocity amplitude	12.8 $\mathrm{m\,s^{-1}}$	0.1 $\mathrm{m\,s^{-1}}$	≈ 1 $\mathrm{m\,s^{-1}}$
astrometric amplitude $d = 10$ pc	500 μas	0.3 μas	≈ 1 mas
astrometric amplitude $d = 100$ pc	50 μas	0.03 μas	≈ 1 mas
timing residuals	2.5 s	1.5 ms	depends on signal

The time dependence of the measured reflex motion provides orbital parameters like period, eccentricity etc., while the amplitude of the observed effect is proportional to the planet mass. If the mass of the star is known then one can estimate the mass of the planet. The different methods provide not exactly the same information and they favor the detection of different kinds of systems as summarized in Table 2.3.

Table 2.3: Resulting planet parameters from measurements of the reflex motion.

method	mass	orbit	favored systems
radial velocity	$M_p \sin i$	$P, T_0, \epsilon, a \sin i, \omega$	shorter period planets, around bright, quiet stars with many absorption lines
astrometry	M_p	$P, T_0, \epsilon, a, i, \omega, \Omega$	longer period (> 1 yr) planets around nearby stars
timing residual	$M_p \sin i$	$P, T_0, \epsilon, a \sin i, \omega$	measurable only for stars with well defined periodic signal

2.2 Radial velocity method

2.2.1 The radial velocity signal

With the radial velocity method one can only measure the velocity component of the star parallel to the line of sight $v_r = dz/dt$. We consider here the effect of one planet only. In this case the temporal dependence of the z-component of the orbit is

$$z(t) = r_S(t)\sin(\omega + \phi(t))\sin i\,, \tag{2.8}$$

where, $\sin i$ is the inclination of the orbital plane with respect to the line of sight, and ω the orientation of the periastron phase away from the line of sight.

The resulting radial velocity signal induced by a planet follows by differentiation of Equation 2.8, the ellipse equation, and Kepler's 2$^{\rm nd}$ law as given below:

$$v_r(t) = \frac{dz(t)}{dt} = \frac{2\pi}{P}\frac{a_S \sin i}{\sqrt{1-\epsilon^2}}\Big(\epsilon\cos\omega + \cos(\omega + \phi(t)\Big) \tag{2.9}$$

For the observed radial velocity there is in addition a (usually) constant radial velocity of the center of mass v_0

$$v_r(t) = v_0 + K\Big(\epsilon\cos\omega + \cos(\omega + \phi(t)\Big) \quad \text{with} \quad K = \frac{2\pi}{P}\frac{a_S \sin i}{\sqrt{1-\epsilon^2}}\,. \tag{2.10}$$

K is called the RV semi-amplitude of the stellar reflex motion which is relevant for the determination of the planetary mass. The term in the brackets defines the shape of the periodic radial velocity curve. The only time-dependent quantity is $\phi(t)$ as described in Section 2.2.

The following equation provide the derivation of Equation 2.9 from Equation 2.8 using the ellipse equation and Kepler's 2$^{\rm nd}$ law (see Section 2.2). Derivation yields:

$$\frac{dz}{dt} = \sin i\,\frac{dr}{dt}\,\sin(\omega+\phi) + r\,\cos(\omega+\phi))\frac{d\phi}{dt}\,.$$

The derivation dr/dt follows from the ellipse equation $r = a(1-\epsilon^2)/(1+\epsilon\cos\phi)$

$$\frac{dr}{dt} = -\frac{a(1-\epsilon^2)}{(1+cos\phi)^2}(-\epsilon\sin\phi)\frac{d\phi}{dt} = \frac{\epsilon\sin\phi}{a(1-\epsilon)^2}\,r^2\frac{d\phi}{dt}\,.$$

One can replace $r^2\,d\phi/dt$ according to the 2$^{\rm nd}$ Kepler law

$$r^2\frac{d\phi}{dt} = \frac{2\pi}{P}\,ab = \frac{2\pi}{P}a^2\sqrt{1-\epsilon^2}$$

and $r\,d\phi/dt$ follows by dividing the Kepler law with r (ellipse equation)

$$\frac{dr}{dt} = \frac{2\pi}{P}a\,\frac{\epsilon\sin\phi}{\sqrt{1-\epsilon^2}} \quad \text{and} \quad r\frac{d\phi}{dt} = \frac{2\pi}{P}a\,\frac{1+\epsilon\cos\phi}{\sqrt{1-\epsilon^2}}\,.$$

Inserting these two relations in dz/dt and using as a trick the following trigonometric relation

$$\cos\omega = \cos\big((\omega+\phi)-\phi\big) = \cos\phi\cos(\omega+\phi) + \sin\phi\sin(\omega+\phi)$$

yields then Equation 2.9 for the radial velocity variation.

Mass for the planet. From the radial-velocity parameters K, P and ϵ follows a_S, the semi-major axis for the orbit of the star around the center of mass

$$a_S \sin i = \frac{P}{2\pi}\sqrt{1-\epsilon}\, K\,. \tag{2.11}$$

Using Kepler's third law $a^3/P^2 = G(m_S + m_P)/4\pi^2$, $a = a_S + a_P$ and $m_S a_S = m_P a_P$ yields the so called mass function

$$\frac{(m_P \sin i)^3}{(m_P + m_S)^2} = \frac{P}{2\pi G} K^3 (1-\epsilon^2)^{3/2}\,. \tag{2.12}$$

For planetary systems there is $m_P \ll m_S$ and one can use for the mass of the planet the approximation

$$m_P \sin i \approx \Big(\frac{P}{2\pi G}\Big)^{1/3} K\, m_S^{2/3} \sqrt{1-\epsilon^2}\,. \tag{2.13}$$

Thus one can derive the mass of the planet $m_P \sin i$ from the radial velocity data provided that the mass of the central star m_S is known. Usually m_S is known within an uncertainty of a few percent. The inclination factor $\sin i$ cannot be determined with RV-data alone.

Impact of the "sin i"-factor. With the radial velocity method alone one can get the planet mass m_P only together with the unknown $\sin i$ factor for the orbit inclination. The measured value $m_P \sin i$ is therefore a lower limit value, also called minimum mass for a planet. Thus one needs to consider that a given system could be seen almost pole-on, with an orbit inclination close to $i \approx 0°$ or $\sin i \to 0$. In such cases the planet mass is strongly underestimated.

If one assumes a random orientation of the orbital planes then the distribution of orbital inclinations is $n(i) \propto \sin i$. The probability to have a system within the inclination range [0,i_x] is

$$\int_0^{i_x} \sin i\, di = 1 - \cos i_x$$

Thus, only 13.4 % of all systems have $i < 30°$ or $\sin i < 0.5$ and therefore a mass which is more than twice the measured minimum mass. 71 % of all systems have $i \geq 45°$ and $\sin i > 0.71$ and 50 % $i \geq 60°$ with $\sin i > 0.86$ and the real mass of the planets is underestimated less than about 30 % or 15 % respectively.

Convenient formulae. The formulae derived above can be expressed in convenient units. The planet mass is:

$$m_P[\mathrm{M}_J]\, \sin i \approx 3.5 \cdot 10^{-2} K[\mathrm{m\,s^{-1}}]\, (P[\mathrm{yr}])^{1/3}\, (m_S[\mathrm{M}_\odot])^{2/3}\,.$$

The semi-major axis a of the orbit of the planet relative to the star follows from the 3$^{\mathrm{rd}}$ Kepler law. With the approximation $m_P \ll m_S$ and one can write

$$a[\mathrm{AU}] \approx (P[\mathrm{yr}])^{2/3} (m_S[\mathrm{M}_\odot])^{1/3}$$

Similarly one can express the radial velocity semi-amplitude for given system parameters:

$$K = \frac{28.4\,\mathrm{m\,s^{-1}}}{\sqrt{1-\epsilon^2}}\, m_p[\mathrm{M}_J]\, \sin i \Big(\frac{1}{m_S[\mathrm{M}_\odot]}\Big)^{2/3} \Big(\frac{1}{P[\mathrm{yr}]}\Big)^{1/3}$$

2.2.2 Examples for the radial velocity signal induced by planets

Slide 2-2 to 2-4 illustrate a few examples for measured RV-curves for extra-solar planetary systems:

51 Peg b: Slide 2-2 shows the historical detection measurements for 51 Peg b, the first extra-solar planets around a normal star, from Mayor and Queloz. The orbital period is about 4.2 days and the RV-semi-amplitude 59 m/s. The scatter of the data points with respect to the fit is 13 m/s. The measurements which were taken during many different orbits have been folded into a phase curve.

HD 4113 system: Slide 2-3 shows the RV-measurements and the phase curve for a longer period (529 days) planet with a very high eccentricity of $\epsilon = 0.90$. In addition there is a long term trend which points to a brown dwarf or low mass star with an orbit > 20 years. A good sampling and long term monitoring are essential to uncover the real nature of such systems. Clearly visible are the yearly observing seasons for this object. Only 4 orbital periods (≈ 6 years) after the initial detection of RV variations the very short periastron passage could be sampled with observation. The definition of the orbit of the outer companion must be completed by the next generation of astronomers.

HD 40307 system: State of the art measurements of the three super-Earth system HD 40307 are shown on Slide 2-4. A measuring precision of ≈ 1 m/s is required to achieve the detection of planets which induce RV-variation with a semi-amplitude of $K = 2 - 5$ m/s. Multi-planet systems like this one require many data points to reveal the very complicated RV-pattern. From a few data points one would interpret the data just as noise induced by the instrument or the central star. A careful analysis is required to extract finally the correct solution.

2.3 Measurement of radial velocities

2.3.1 Science requirements.

A very high measuring precision is required for the detection of the planet induced reflex motion of stars. The Doppler effect gives the relation between radial velocity v_r and the relative wavelength shift $\Delta\lambda/\lambda$, which is:

$$\frac{v_r}{c} = \frac{\Delta\lambda}{\lambda} \tag{2.14}$$

If one aims to achieve an accuracy of $\Delta v = \pm 3$ m/s. then this corresponds to a measuring precision of

$$\frac{\pm\Delta\lambda}{\lambda} = \frac{\pm\Delta v_r}{c} = \frac{\pm 3 \text{ m/s}}{300'000 \text{ km/s}} = \pm 1 \cdot 10^{-8} \,. \tag{2.15}$$

The next generation of instruments aims at an accuracy which is another order of magnitude higher. What this very demanding science requirement really means is illustrated in Slide 2-5. The slide shows a section of the solar absorption line spectrum and picks a single line which is then compared to the shift expected by the radial velocity variation induced by a planet. The effect is much smaller than the line widths. It is obvious that

the technical requirements for building such a measuring instrument are very demanding and require sophisticated techniques.

In the following sections we introduce first the basic principles for astronomical spectroscopy before addressing the details of the modern RV measuring techniques.

2.3.2 Telescopes

Basic principles of a telescope We first consider the astronomical refractor to explain the basic layout of telescopes. The astronomical refractor consists of:

- a converging (convex) objective lens or aperture lens,
- an intermediate focus with eventually a field stop,
- a collimating lens,
- a parallel beam section with an intermediate pupil,
- a converging lens (camera lens).

Figure 2.3: Basic principles of a telescope.

The following quantities are used to describe a telescope:

- f_1: focal length of objective lens L_1,
- f_2: focal length of the collimating lens L_2 (or the eyepiece),
- f_3: focal length of the camera lens L_3,
- y_1: or D, the diameter of the entrance pupil or aperture,
- y_2: diameter of the intermediate pupil,
- α: semi-angle of the field of view in the entrance pupil,
- θ: semi-angle of the field of view in the exit pupil.

Some basic principles are:

- An image of the object is formed in the focal planes. The image plate scale is defined by the effective focal lengths f given by the F-ratio of the image forming lens L_1 or L_3 and the diameter of the entrance pupil D:

$$f = F_x \cdot D\,. \tag{2.16}$$

– In a pupil plane the light from a distant object forms parallel rays. The information is in the angular direction of the rays. In the case of the astronomical refractor the collimating eyepiece refracts the light into a more or less parallel beam suited for observation with the eye. The magnifying power m (angular magnification) of the telescope is:

$$m = \frac{f_1}{f_2} = \frac{\tan\theta}{\tan\alpha} = \frac{y_1}{y_2}\,. \tag{2.17}$$

– The collimated beam section between L_2 and L_3 is the typical location for diffraction gratings or prisms for spectroscopy.
– The field stop in the first focal plane can be used for field selection (e.g. a spectrograph aperture) or for a coronagraphic mask.
– Well designed pupil and field stops can be very helpful in reducing the stray light and the background level of an instrument.

2.3.3 Spectrographs

Equations for grating spectrographs. A diffraction grating consists of a large number of very fine, equally spaced parallel and periodic slits separated by a. The wavelets from each slit are strongly enhanced in a few directions θ_n in which all the wavelets are in phase so that constructive interference occurs. θ_1 is the first order with a path difference of λ etc. For wavelength λ the angular displacement θ_m is

$$\sin\theta_m = \frac{m\cdot\lambda}{a}\,, \tag{2.18}$$

where a is the periodic separation between the grating lines and m an integer number for the interference order. Because θ_m depends on λ the light is dispersed and spectra are formed. One should be aware that an overlap of the different orders ($m-1$, m, $m+1$, etc.) can occur. The same effect is obtained for a reflection grating which has fine periodic rulings.

Figure 2.4: Principle of a spectrograph grating.

The angular dispersion $d\theta/d\lambda$ of the grating follows from differentiation (determine first $d\lambda/d\theta$):

$$\frac{d\theta}{d\lambda} = \frac{m}{a\cos\theta}\,.$$

The angular width W_θ of a monochromatic interference peak is broad for few grating lines and it becomes narrower as the number of illuminated grating lines N increases like

$$W_\theta = \frac{\lambda}{N a \cos\theta}\,.$$

This width can also be expressed in a wavelength width $\Delta\lambda$

$$\Delta\lambda = W_\theta \frac{d\lambda}{d\theta} = \frac{\lambda}{N a \cos\theta} \cdot \frac{a\,\cos\theta}{m} = \frac{\lambda}{N\,m}\,,$$

or a resolving power R for a more general characterization of the diffraction limited resolution of the grating

$$R = \frac{\lambda}{\Delta\lambda} = N\,m\,. \tag{2.19}$$

This formula indicates that the resolving power depends only on the number of illuminated grating lines N and the dispersion order m.

Gratings, in particular reflective gratings, are often inclined with respect to the incoming beam by an angle i which is the angle of the grating normal to the incoming beam. The angle θ is then defined by the interference order m and the zero order. In this case Equation (2.18) includes the term $\sin i$ for the grating inclination:

$$\sin\theta = \frac{m \cdot \lambda}{a} - \sin i\,. \tag{2.20}$$

This is called the grating equation. The resolving power is larger for inclined gratings because more lines are illuminated for a given beam diameter

$$R = \frac{\lambda}{\Delta\lambda} = \frac{N\,m}{\cos i}\,. \tag{2.21}$$

The grating equation describes also how one can change the central wavelength and the wavelength range for a given deflection angle θ by changing the tilt angle i. The following list gives the dependence of spectrum parameters on grating properties:

- the resolving power R depends only on the number of illuminated lines and the diffraction order
- R increases if the number of lines per mm of the grating are enhanced for a given beam (=pupil) diameter
- R can be enhanced for a given grating by a larger illuminating beam (larger pupil) or by tilting the grating so that the illuminated lines increase like $N_i = N_{i=0}/\cos i$
- R is substantially larger for higher diffraction orders $R \propto m$ (there is the restriction that overlap of grating orders occur)
- the wavelength region for a given deflection angle can be selected by changing the inclination i of the grating.

2.3.4 Different types of spectrograph gratings

Simple gratings. Typical grating have about $100 - 1000$ rulings/mm. This yields for the first order diffraction spectrum and a collimated beam diameter (pupil diameter) of 1 cm a grating resolving power of $R = 1000 - 10000$. The first order spectrum can be contaminated by the second order spectrum with $\lambda_{m=2} = \lambda_{m=1}/2$ or higher order spectra $\lambda_{m\geq 3}$. For ground based optical spectroscopy this happens for $\lambda_{m=1} > 660$ nm, when second order light from above the UV-cutoff $\lambda_{m=2} > 330$ nm sets in. The second order can be suppressed with a short wavelength cutoff filter. For example a BG430 filter cuts all light short-wards of about 430 nm, allowing first order spectroscopy from 430 nm to 860 nm without second order contamination.
The same grating can also be used in second order with twice the resolution of the first order. In this case one has to select for a given wavelength range the correct pass-band filter to avoid the contamination by other orders.

Blazed gratings. Simple gratings are not very efficient since the light is distributed to several grating orders. The efficiency of reflecting gratings can be improved by an optimized inclination of the reflecting surfaces so that they reflect the light preferentially in the direction of the aimed interference order. Thus the grating efficiency is optimized for one particular wavelength or diffraction angle θ_b, the so-called blaze angle.

Figure 2.5: Illustration of a blazed grating and an echelle grating.

Echelle gratings. An extreme case of the blazed grating is the echelle grating. It is strongly inclined with respect to incoming beam and more importantly it is optimized (blazed) for high order diffraction directions, say $m = 10 - 100$. With this type of grating the resolving power can be strongly enhanced even if the grating is quite coarse. For example a beam of 2 cm diameter illuminating an echelle grating with 20 lines / mm, inclined by $i = 60^\circ$ ($1/\cos i = 2$) will see effectively 800 grating lines, which produce for $m \approx 50$ a spectral resolving power of $R = 40'000$. Of course for such a grating the free spectral range, without overlap by neighboring pixels, is only small and of the order $\Delta\lambda \approx \lambda/m$. Narrow band filters are required to select one particular order.
A more elegant solution is a **cross dispersion** with a second low order grating or a prism perpendicular to the dispersion of the echelle grating. In this way the individual orders

are displaced with respect to each other and many orders of the echelle grating can be placed on a rectangular imaging detector.

2.3.5 Technical requirements for a RV-spectrograph

The scientific requirement for measuring radial velocities with a precision of 3 m/s imposes very demanding technical requirements for a spectrograph. The following type of measurement must be possible:

- a spectral resolving power $R = \lambda/\Delta\lambda > 50'000$ is required for resolving the individual spectral lines and measure small wavelength shifts,
- a broad spectral coverage is required to cover many thousand lines for higher efficiency or for achieving enough signal for a successful detection,
- a high spectrograph throughput and/or a large telescope which allows also a search for planets around faint low mass stars or more distant stars,
- an efficient measuring and calibration strategy,
- a telescope with enough available observing time for monitoring program.

For achieving these technical requirements there are various problems and instrument effects which must be under control. The following effects introduce shifts which can be 2 or 3 orders of magnitudes larger than the required precision of 3 m/s:

- The variation in the index of refraction of air $n_{\rm air}$ with temperature or pressure introduces strong wavelength shifts. The measured wavelengths is

  $$\lambda_{\rm meas} = n_{\rm air}(P,T)\lambda_{\rm vac}$$

  where $\lambda_{\rm vac}$ is the wavelength in vacuum. A temperature change of 0.1 K or a pressure change of 0.1 mbar introduce a wavelength shift of about 10 m/s! These large amplitude drifts must therefore be corrected with very high accuracy. This is achieved with the simultaneous measurement of a calibration spectrum.
- Mechanical flexures introduced by thermal expansion or variations in mechanical forces can cause substantial shifts of the spectrum on the detector. The typical velocity sampling on the detector is about 1 km/s per pixel, with pixels dimensions of about 10 - 20 μm. Thus a measured wavelength shift of 3 m/s corresponds on the detector to a physical shift of the spectrum of about 50 nm. Any uncalibrated instrumental flexure at this level is very harmful.
- The spectrograph illumination must be very stable. If the illumination is not fully stable due e.g. to telescope guiding errors, then this may introduce variable illumination gradients which may cause large spurious wavelength shift. An absorption cell in front of the spectrograph aperture or a spectrograph illumination with a fiber are two approaches to solve this problem.

These are only the most important disturbing effects which must be considered. Other issues are the definition of accurate wavelengths of the spectral reference (to 9 or 10 digits), the stability of the spectral reference, detector effects like non-perfect charge transfer efficiency and other problems.

2.3.6 HARPS - a high precision radial velocity spectrograph

HARPS is currently the best instrument for the measurement of reflex motion of planetary systems. HARPS is an echelle spectrograph at the ESO 3.6m telescope which was built as dedicated planet search instrument. It can measure radial velocities to a precision of $\Delta\mathrm{RV} = \pm 1$ m/s or even a bit better.

HARPS is optimized for stability and the instrument has no moving components. The light of a star is focussed by the telescope onto an entrance lens of an optical fiber which guides the light to the spectrograph located in a laboratory. Using a fiber has the advantage that it equalizes inhomogeneous illumination from the telescope due to the many internal light reflections. Thanks to this the illumination of the spectrograph is very homogeneous.

The HARPS spectrograph is located in a temperature controlled vacuum vessel (Slide 2.6) to minimize drifts due to temperature and air pressure variations and thermal variations in the instrument. Simultaneously with the target a second fiber is illuminated with a Th-Ar spectral reference light source which provides a rich and accurate reference spectrum for the calibration of each individual measurement. The spectral range covered is 378 nm – 681 nm with a spectral resolving power of about $R = \Delta\lambda/\lambda = 90'000$. The resulting data format of the echelle spectrograph is shown on Slide 2-7.

The main components of HARPS are:

- one fiber head for the target at the Cassegrain focus of the ESO 3.6 m telescope,
- a second fiber input which can be fed by a ThAr emission line lamp,
- an echelle grating, 31.6 gr/mm with a blaze angle of 75° with dimensions of 840 × 214 × 125 mm (Slide 2-6),
- collimator mirror with a diameter of 730 mm ($F = 1560$ mm), which is used in triple pass mode,
- cross disperser grism (a combination of transmission grating and prism) with 257 lines/mm,
- two 2k × 4k CCD detectors which record the echelle grating orders $m = 89 - 161$ in a square image plane.

The HARPS spectrograph achieves a spectral resolution ($\approx$ 3 pixels) of about $\Delta\lambda =$ 0.005 nm or 50 mÅ. This allows us to measure wavelength shifts of the order 10^{-6} nm provided there are a large number, say $> 10'000$, of narrow lines with a width of ≈ 0.02 nm in the spectrum. The measured wavelength shift corresponds to a physical shift of the spectrum of about 10 nm on the detector or $\approx 1/1000$ of a detector pixel.

2.3.7 Stellar limitations to high-precision radial velocity

Measuring the reflex of motion of a star to a precision of $\Delta v \approx \pm 3$ m/s requires that the stellar properties are favorable for such a measurement. The ideal target for radial velocity measurements are bright, G or K main-sequence star (mass range $0.7 - 1.2\mathrm{M}_\odot$) with very low magnetic activity. Of course one would like to investigate planetary systems not only around such stars but for a broad range of systems with different stellar parameters.

Figure 2.6: Stars in the Hertzsprung-Russell diagram and their usefulness for the radial velocity search of planets.

Radial velocity search for different stellar types. We consider different types of stars in the Hertzsprung-Russell diagram in Fig. 2.6 and discuss their suitability for radial velocity searches of planets.

- **G and K main-sequence stars:** These are often the ideal targets for the RV-search of planets. They have a very rich absorption spectrum and they are also sufficiently frequent and bright for a statistically representative sample (> 500 objects). G and K stars younger than 1 Gyr are still rotating quite fast and therefore their atmosphere is not stable enough for very accurate measurements. We will discuss this further below.
- **F main-sequence stars:** Late F-stars can be good targets like G stars but for early F-stars the lines are quite broad and the rotational induced activity is often a problem. Atmospheres of F-stars show quite often stellar oscillations (see Slide 2.8 for an example).
- **A stars:** A stars are problematic for planet searches with the RV method. Hot stars have less absorption lines and their widths are strongly enhanced by pressure broadening and rotation. For example the bright A stars Altair and Fomalhaut show line broadening due to a rotation speed of $v_{\rm rot} \sin i = 240$ km/s and 93 km/s respectively. Therefore it is currently only possible to find for A stars the reflex motion with $K \gtrsim 30$ m/s. Only giant planets with short orbital periods < 1 year can be found around these stars.
- **M type main sequence stars:** Old M-stars are often objects with very stable atmospheres well suited for radial velocity measurements. M stars younger than about 1 Gyr are often too active (like young G and K stars) for accurate RV-searches. An important problem is that M-dwarfs are faint in particular in the visual wavelength range. Larger telescopes and RV-spectrographs optimized for the near-IR wavelengths are required to investigate better M-stars with high precision.

- **G- and K-giants:** These are evolved A- and late B-stars located on the giant branch in the HR-diagram. Because they have a rich spectrum of narrow lines they can be used to assess planets around stars with a masses $\gtrsim 1.5 M_\odot$ (what is difficult for main sequence stars).
- **Other stars:** High-mass stars and luminous red giants are not suited for radial velocity searches of planets. Hot, high mass stars and white dwarfs have only few, broad lines. Red giants show at least strong, disturbing atmospheric oscillations (at the 1 km/s level) if not large amplitude pulsations (> 10 km/s). It seems impossible to measure for such star the potential reflex motion due to a planet.

Magnetic activity of late type stars. As discussed above only late type main sequence stars of spectral type F, G, K and M and subgiants or giants of spectral type G and K are suited for sensitive $K < 30$ m/s radial velocity searches of planets.

For a RV search in the range ΔRV $\approx$ m/s for low mass planets one needs to address the disturbing effects from the stellar atmosphere in more detail. All late type stars have a convective outer envelope and they are born as rapid rotators with rotation periods of a few days. Convective motion and fast rotation induces a dynamo effect and therefore strong magnetic activity which is disturbing the RV measurements. With age the rotation slows down due to the loss of rotational angular momentum via a magnetized stellar wind. For this reason, old stars show much less activity and are therefore better targets for very sensitive radial velocity searches. For comparison, our sun with an age of 4.6 Gyr is a rather old and inactive star with a long rotation period of about one month. It would be suited for the search of planets.

Disturbing effects which are introduced by magnetic activity:

- stellar rotation and spots,
- magnetic suppression of convective surface motion,
- stellar oscillations.

Figure 2.7: Schematic illustration of the disturbing effect of a spot on a rotating star on the structure of a symmetric absorption line.

Stellar spots, like solar spots, are darker and cooler than the surrounding atmospheres. Therefore they produce on a rotating star absorptions lines with an asymmetry on the blue side if the spot is on the approaching side of the rotating star and an asymmetry on

the red side if the spot is on the receding side of the star. Because the spot is cool, it will produce a deeper absorption for low excitation lines (e.g. from neutral atoms an molecules) while the spot absorption are less deep for high excitation lines from e.g. ionized species. Overall, the impact is a systematic displacement of line centers which can significantly affect the RV-measurement of a planet search program. An example for the resulting RV-signal induced by spots is shown for α Cen B in Slide 2.9.

Convection at the surface of late type stars is suppressed slightly by magnetic activity and the presence of stellar spot. Because the upward motion is less strong, we see less gas with strong velocity components towards us and the average RV of the star is red-shifted. This effect is stronger during active phases of the magnetic cycle of stars. For the sun the magnetic activity goes up and down on a cycle with a periodicity of 11 years.

For extra-solar planet search programs these activity cycles can be very disturbing because a very quiet star may turn suddenly into an active star and introduce much enhanced intrinsic RV-velocity variations (see Slide 2.9). For the analysis of RV-data it is important to check always the activity status of a star. A good activity indicator are the Ca II H and K lines, which have for active phases an enhanced central emission components.

Figure 2.8: Schematic illustration of the Ca II emission cores in active late type stars.

Stellar oscillations are induces by the convective gas motion in the envelope. Oscillations are often quite strong for F-stars, because they are located in or close to the pulsation instability strip in the HR-diagram. Typical periodicities of the stellar oscillation are of the order several minutes and the induced RV-shift can be of the order 10 m/s (see Procyon on Slide 2.8) but also much larger.

For more extended giant stars the occurrence of oscillation-like atmospheric instabilities is more frequent. Essentially all F-giants and M-giants are pulsating stars. G and K giants can be quite stable and they are therefore good targets for RV-search of planet around more massive stars than the sun. Typically these giants have masses of ≈ 2 $M_\odot$.

2.3.8 Data reduction and data analysis steps

The RV-search project teams need sophisticated software tools and a good understanding of all instrumental effects for the data reduction and analysis. The HARPS RV-data become public one year after the observations, like all ESO data. Nonetheless, the Geneva team reaches with their own data reduction software, which they developed over the last decades, a significantly higher precision than other researchers using the same data but another software. We address here some basic steps in the data reduction.

Cross-correlation technique. The RV-measurements use a cross-correlation technique of a target spectrum with a reference spectrum of a high quality RV standard star.

For the cross-correlation the data points from the extracted echelle spectrum must be sampled with a constant ΔRV steps, which are constant $\Delta \log \lambda$ steps on a logarithmic wavelength coordinate $\log \lambda$. In this way one can move the entire target spectrum by a shift value δ_i with respect to the reference spectrum and determine the cross-correlation parameter. If the two spectra match then one gets a cross-correlation peak. The exact center of the peak is then obtained with a least square fitting procedure. The result is the offset of the target velocity with respect to the velocity of a reference spectrum.

Barycentric correction. The obtained RV-value must be translated into a barycentric RV, which is the RV with respect to the center of mass of the solar system. The measured value depends on the radial motion of the observer with respect to the line of sight. This barycentric correction is large and includes:

- the Earth motion around the center of gravity of the solar system which is 30 km/s, or 10'000 times larger than the 3 m/s measuring goal,
- the Earth rotation which is of the order 0.3 km/s,
- the Earth motion around the center of gravity of the Earth-Moon system which is of the order 10 m/s.

The corrections for the Earth motion can only be applied with sufficient accuracy if the exact observing time is registered. The radial velocity correction for the Earth motion can change in 15 minutes by more than 1 m/s. Therefore it is not good to observe during cloudy nights because clouds may shift the central time of an exposure (half of the photons collected), if for example the second half of a long exposure is affected by clouds.

Period and RV-orbit search. The first step in the RV-planet analysis is the search for a strictly period RV-variations. The basic principle for a planet search is an approximation of the measured points $v_i(t_i)$ by a fit curve with a sine function:

$$v_{\rm fit}(t) = v_0 + K \sin\Big(\frac{2\pi}{P} t + \delta\Big) \,.$$

There are 4 parameters, system velocity v_0, radial velocity semi-amplitude K, orbital period P, and phase shift δ, which are varied until the best fit is obtained. Usually the best fit is defined by the minimum of the least square parameter σ

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} |v_{\rm fit}(t_i) - v_i(t_i)|^2 \,.$$

One should restrict the period search to reasonable parameter ranges. Good estimates for the starting points are often easy to find for v_0 (= mean RV value) and K ($\approx 0.5 v_{\max} - v_{\min}$), while δ must be varied in the interval $[0, 2\pi]$ for each trial period. The period range to be searched goes from about the shortest possible orbital period (e.g. 0.5 days) to the total time span of the collected data points. Searching in the short period range is only safe if there are also data-points which were taken within a fraction of the searched period. Else the procedure produces good fits with very short, orbital periods which are not well sampled by the data points. A good sampling requires that there are several well distributed data point during at least one orbital period.

A periodogram provides the minimum σ^2-value for each period for the best fitting v_0, K and δ parameters. The highest, isolated peaks should be analyzed more carefully. There could be multiple narrow minima if data points are separated by long time intervals without measurements. Then a periodicity of say 4 days which fits the data taken during one week of one year and another week taken during the following year may have several solutions separated by about 0.04 days. This happens because it is not clear whether the two measurement campaigns are separated by 85, 90 or 95 orbital periods. Additional data point are required to break the degeneracy.

Figure 2.9: Illustration of not well sampled RV measurements.

Similar effects happen because of natural observing gaps or observing periodicities. With Earth-bound observations targets can typically not be observed during some seasons when the sun is at a similar right ascension as the target, or there are monthly cycles because the RV-program obtains only bright time near full moon, which is not demanded by extra-galactic observers. And of course there is the daily cycle because observations are only possible during the night. All these periodicities may cause spurious signals in the period search procedure and aliases of the real period with data sampling patterns.

If a good sine-fit for the data points is found then one should refine the fitting with the full RV-equation (Eq. 2.9) including the eccentricity ϵ and the periastron angle ω.

2.4 Statistical properties of radial velocity planets

2.4.1 On statistical methods

Target samples: Meaningful statistical studies for radial velocity planets are only possible with well defined surveys and well-understood detection thresholds and observational biases. A good approach is to select stars with predefined intrinsic properties, like all single G and K dwarf stars, using one of the following sample criteria:

- a volume limited sample which includes all stars within a fixed distance,
- a magnitude limited sample which includes all stars which are brighter than a certain limit.

In the volume limited case there are relatively less higher mass stars because they have a lower volume density, while in the magnitude limited sample the survey volume is larger for the brighter stars than for the fainter stars. Both approaches have some subtle advantages and disadvantages. However, the important point is that the sample selection is well understood. Table 2.4 give numbers for the frequency of stars for a distant limit of 10 pc and their typical 10 pc brightness.

Table 2.4: Statistics for the stellar systems within the solar neighborhood up to a distance of 10 pc.

object type	numbers	M_V (spec. type)
stars (total)	342	
O and B stars	0	–4.0 (B0 V)
A V stars	4	+0.6 (A0 V)
F V stars	6	+2.7 (F0 V)
G V stars	20	+4.4 (G0 V)
K V stars	44	+5.9 (K0 V)
M V stars	248	+8.8 (M0 V); +12.3 (M5 V)
G and K III giants	0	+0.7 (K0 III)
white dwarf stars	20	
single stars	185	
binary systems	55	
multiple systems (3+)	12	
brown dwarfs	15	
planets	19	
planetary systems	11	

In many surveys there are often some less well defined selection criteria which are due to special target properties, like:

- pulsations,
- fast rotation and high magnetic activity,
- stars in close binary systems with apparent separations of less than a few arcsec,
- early type stars (e.g. A and early F).

Such objects are often excluded from a survey because the planet detection limits are much worse when compared to "normal" stars. From a statistical point of view a clean sample without special cases is easier to understand. This does not preclude a specific study on a sample of special, say e.g. young, fast rotation stars.

Important surveys for the derivation of statistical properties of RV-planets are:

- the volume limited CORALIE and HARPS sample of F, G and K stars with about 2800 stars,
- the magnitude limited Lick, Keck and AAT survey of about 1330 F, G, K and M stars.

The two surveys overlap with many targets in common.

Bias effects. Statistical studies can include various kinds of selection or bias effects which may change the measured distribution of system parameters. Such effects are often unavoidable because of observational constraints, availability of sources, instrumental effects or not considered or unknown sample properties.

For a rough assessment of statistical results the most important bias effects should be known. For the RV-surveys important effects are:

- planets with large mass are easier to detect than planets with small mass for given orbital parameters,
- for a given mass planet with short periods are easier to detect than planet with long periods.

This indicates that in a distribution of measured RV-planets there could always be more, undetected objects with longer periods or smaller mass. On the other hand, if lower mass or longer period planet are more frequent than high mass or short period planets then this is at least qualitatively a robust result. It is very unlikely that there exists a large population of high mass or short period planet which was not picked up by the survey.
In order to achieve accurate and most useful scientific results a careful analysis of the obtained results is required:

- the sample selection should be based on well-understood parameters.
- observations of a predefined sample should be completed,
- if the completion of a sample is not guaranteed, then the targets should be picked according to a scheme which depends not on intrinsic target parameters. In this way an incomplete data set is not affected by a preference of the observers who like to pick the bright targets first and the faint targets later and only if time permits. Often it is good to use a scheme according to the RA-coordinate of the targets.
- computer models of the sample are very useful to understand bias effects because they can provide accurate corrections for known bias effects and an assessment of the uncertainties in the final survey results.

Statistical noise. Low number statistics suffers strongly from statistical noise. Poisson statistics can be used for many simple cases as starting point. For Poisson statistics the uncertainty is

$$\sigma = \sqrt{N}$$

where N is the number of objects in a given statistical bin. For example, if a survey has detected $N = 50$ targets, then the statistical uncertainty is $\sigma = \sqrt{N} \approx 7$ or a fractional uncertainty of 14 %.

2.4.2 Frequency of planets

Statistical results presented in this and the following section are mainly from the review paper of Udry and Santos (2007, Ann Rev. Astron. & Astrophys. 45, 397) and the preprint from Mayor et al. (2011, arXiv:1109.2497v1). Because of the rapid progress in this field many of the presented results will be outdated in a few years.

The most basic statistical quantity of planet search programs is the fraction of detected planets for the surveyed stars.

Giant planets around late type stars. For giant planets around single late type F, G, K main-sequence stars the RV velocity surveys find the following numbers:

- for about 1 % of the stars a close in (<0.1 AU), hot Jupiter with $m_P \sin i > 0.1 \mathrm{M}_J$ is detected,
- for about 15 % of the stars a giant planet with $m_P \sin i > 0.1 \mathrm{M}_J$ out to a separation of 5 AU is present,
- RV-surveys can not say much about the frequency of giant planets at large separation > 5 AU, because the RV-search programs run not yet long enough for detecting planets with periods much longer than 10 years.

These numbers are quite robust because the signal of Jupiter mass objects are relatively easy to detect. Slide 2.10 illustrates how an extrapolation from the detected number of plants to expected planet rates is made. For giant planets, the uncertainties in the correction factor for non-detected systems are smaller than the statistical noise.

One may speculate about the real frequency of giant planets, including also objects with larger separation than 5 AU. Since there seems to be no drop of planet frequency with long periods > 3 years (see period distribution below) one can assume that the distribution does not drop steeply just beyond $P > 10$ years. Thus one can assume that at least 25 % of single, late type stars have a giant planet.

Frequency of low mass planets. It is not easy to derive estimates for the frequency of low mass planets from RV surveys, because the samples are small and usually the detection depends a lot on stellar properties, like intrinsic atmospheric variations and observational limitations. For this reason the numbers derived so far should be considered cautiously.
For very stable, single G and K stars the Geneva group has estimated the occurrence rate of low mass planets, taking bias effects into account. For this the detection probability as function of orbital period and mass was determined in order to extrapolate from the number of detected planets to the number of expected planets (see Slide 2.10).

- the planet/star ratio is about 0.50, counting planets in the mass range from 1 M_E – 0.1 M_J and periods shorter than 100 days,

Note, that the solar system does not fulfill this selection criterion for a planetary system with low mass planets.

Metallicity dependence. Soon after the detection of the first extra-solar planets, it was recognized that the host stars have on average a high metallicity. More detections of giant planets confirmed this planet-metallicity correlation.

- the frequency of giant planets around low metallicity stars is about 5 %, where low metallicity means the metallicity range from about 0.3 – 1.0 times the solar value.
- the frequency of giant planets is much higher for high metallicity stars, about 20 % or even a bit higher for stars with a metallicity twice the solar value.

For the frequency of low mass planets $M < 0.1$ M_J, such a correlation with metallicity is not observed.

Selecting high metallicity stars in a planet search program enhances significantly the detection rate of giant planets.

Metallicity differences of stars is a well known property in our Milky Way but also other galaxies. The variance of the stellar metallicities in the Milky Way is explained partly as an age effect (older stars tend to have lower metallicity) and a general metallicity gradient from the center to the outer regions of the Milky Way.

The fact that the giant planet population is different for high metallicity systems, when compared to low metallicity systems indicates that the planet formation process must depend on metallicity. In a high metallicity system the fraction of dust in the protoplanetary disk may be enhanced. Another widespread effect of a high metallicity medium is a more efficient cooling for the gas through the line emission from heavy elements. More dust and a cooler environment could indeed have strong effects on the planet formation process.

How many planets are there in the Universe? The numbers given above from the RV-surveys indicate that single stars with planets are more frequent (about 2/3) than stars without planets (about 1/3). This results holds at least for F,G,K stars. In addition one can extrapolate the planet frequency including giant planets which are further away than 10 AU from the star, or low mass planets with periods larger than 100 days, or planets with masses less than M_E.

Our solar system does not qualify to be counted in the statistics of Slide 2.10. Jupiter and Saturn have a period which is beyond the considered separation range and Earth and Venus have also a too long orbit to be counted in the low mass planet statistics. Of course, we can not use the solar system for an extrapolation of the planet frequency values, but the solar system hints to the fact that the statistics considered above could miss a large number of existing planets. Thus one can suspect that there are more planets than stars in the Universe and one may now speculate about the average number of planets per star.

An interesting scientific question for the future is, whether there are stars without planets and why they do not harbor planets.

2.4.3 Distribution of planetary masses

Stellar companions to stars are very frequent. About $25 - 30$ % of all stellar system are binary or multiple star systems or almost every second star is part of a binary or multiple system (see Table 2.4). The typical separation of binaries peaks around 20-50 AU, or for periods of about 100 years.

Including planets in this companion distribution then their is a very strong bimodality. There are many giant planet companions and many stellar companions but only few objects with a mass in the brown dwarf regime 20 – 50 M_J. This mass range is often called the brown dwarf desert.

Giant planets. If one considers the mass of the detected RV-planets, then there is a well defined frequency drop-off for high mass planets $m_P \gtrsim 2\,M_J$ (Slide 2.12). The high mass end of the giant planet distribution can be described by

$$\frac{dN}{dM} \propto \frac{1}{M}\,.$$

This strong drop-off is a robust result, because it is much easier to detect giant planets with higher mass.

Ice giants and super-Earths. The RV-surveys for low mass planets are still quite small. The largest sample of low mass planets is the HARPS-sample from the Geneva team. Their sample shows:

- a strong peak in the mass distribution at about 2 M_J for the giant planets,
- a clear minimum in the mass distribution in the range 0.1 – 0.3 M_J,
- another strong peak for masses 10 – 20 M_E (0.03 – 0.06 M_J) comparable to Uranus and Neptune in the solar system,
- perhaps another depression just below 10 M_E,
- a clear increase in the frequency of planets for lower masses but the gradient of the mass distribution of low mass planets is unclear.

The minimum in the mass distribution in the range 0.1 – 0.3 M_J is at least qualitatively a very robust results. Bias effects cannot explain the higher rates of Neptune mass planets (ice-giants), when compared to low mass 0.1 M_J gas giants, which must indeed be quite rare.

2.4.4 Orbital period distribution of extra-solar planets

The orbital period is another key parameter for planets. Already the first detection of 51 Peg b pointed to the fact that orbital periods of extra-solar planet may be very different to what we have in the solar system. Indeed it became rapidly clear that 51 Peg b is just one representative of a larger group.

Period distribution of giant planets. The period distribution of giant planets is well known. There are the following features:

- there is a strong maximum at about 3–5 days orbital period. This group is called the “hot” Jupiters,

- planets with very short periods of less than 2 days are 10 times less frequent than planets in the 3–5 day peak,
- in the range 10 – 300 days there are significantly less planets per $\Delta \log P$ interval, than in the 4 day peak,
- at periods > 300 days the planet are more frequent up to periods of a few 1000 days which marks the limit in the duration of the observing programs.

Period distribution of low mass planets. The statistics for low mass planets are poor. However, one can say that the period distribution in the range 3 - 30 days is rather flat, without a strong preference for a specific period like for the hot Jupiters.

Period-mass diagram. The period-mass diagram shows an interesting property of the short period giant planets. Essentially all giant planets with periods < 100 days have a mass < 2.5 M_J (Slide 2.13). There are a few exceptions, but these are planets in stellar binary systems.

This property indicates that there is a mass selection effect for giant planet with short periods. Only the lighter ones are pushed inwards towards the star.

2.4.5 Orbital eccentricities

The period - eccentricity diagram (slide 2.14) illustrates well the properties of the orbits of extra-solar planets.

- Short period planets $P < 10$ days have circularized orbits with eccentricities $\epsilon \lesssim 0.2$. For many of these planets the eccentricity is $\epsilon = 0$ within the measuring uncertainties. All planets with orbital periods $P < 3$ days have zero eccentricity.
- For longer periods $P > 100$ days the planets show a very broad eccentricity distribution, with about 30 % of all planets with an eccentricity larger than $\epsilon > 0.4$.

2.4.6 Planets in binary systems

Surveys of planet in binary systems are much more difficult to perform because there is already a very strong RV-signal form the orbital motion of the binary. Some studies have been made, but the statistics are still poor because there are many different types of binaries. For example one may select narrow or wide binaries, binaries with stars of equal mass or with a large mass ratio, binaries with eccentric or near circular orbits and so on. For each class the search strategy must be adjusted and the selected sample must be representative of a given subgroup of binary system. It is not surprising that robust statistical data for binaries are missing. Slide 2.15 shows some examples of multi-planet systems which illustrates the large diversity in their structure.

A rough result is that planet also exist in binary systems. About 15 % of all known planet are in binary systems but it is not clear how many binary systems have been really searched for planetary companions.

2.4.7 Multiple planet systems

About 10 % of the detected systems with RV-planets have more than 1 known planet. A few examples of these multi-planet systems are shown in Slide 2.15. In many cases the additional planets were found because the RV fit solution for the first planet showed systematic deviations. The statistical properties of multi-planet systems are not well known because the selection effects are hard to define and no "clean" sample exists. A few finding are described here.

More low mass planet systems? Just counting detected planets in multiple systems indicates that low mass planets tend to be members in multi-planet systems. This could be a selection because low mass planets are usually searched with many high quality observations, which are required for a successful detection. Such high quality data are therefore better suited to detect multiple planets in a system. On the other hand a giant planet search program which surveys a large number of stars but only with a restricted number of measurements per star, say 20, is more successful in detecting only the most prominent planet in a system. Multiple planet detection would require more high quality data per system.

Orbital period resonances? If the orbital periods of the planets are considered then one finds many systems which are close to an orbital resonance.

Systems with planet close to orbital resonances are:

- Jupiter ($P_J = 11.9$ yr) and Saturn ($P_S = 29.5$ yr) have a period ratio of 2.48 or close to a $5:2$ resonance,
- Gl 876 b and c have periods $P_b = 30.12$ days and $P_c = 61.02$ days with a period ratio of 2.03 close to a $2:1$ resonance,
- the pulsar system B 1257+12 has two planet in orbit with $P_B = 66.54$ days and $P_C = 98.22$ days close to a $3:2$ resonance.

There are many more systems where such a coincidence seems to exist. Although the statistical evidence is not strong, there are hints that at least for some systems the dynamical evolution seems to lock some planets into resonant orbital periods. Planet migration could explain such a behavior. If one planet moves inwards, e.g. due to angular momentum transfer to a disk, then this may also force another planet further in to migrate because its orbit is locked in a resonance with the migrating planet.
More data are required to assess the statistical evidence for the preference of planets in resonant orbits.

2.5 Astrometric detection of planets

2.5.1 The astrometric signal induced by a planet

The astrometric signal θ of the reflex motion of a star at the distance D introduced by a planet is equivalent to the apparent angular size of the semi-major axis of the star

$$\theta = \frac{a_S}{D} \approx \frac{m_P}{m_S}\frac{a}{D}\,, \tag{2.22}$$

where we used the relation $a_S m_S = a_P m_P$ and the approximation $a_P \approx a$. With Kepler's 3$^{\rm rd}$ law one obtains:

$$\theta = \Big(\frac{G}{4\pi^2}\Big)^{1/3}\frac{m_P}{m_S^{2/3}}\,\frac{P^{2/3}}{D} \tag{2.23}$$

This can be written in convenient units like

$$\theta = 2.9\,\mu{\rm as}\,m_P[{\rm M}_E]\,\Big(\frac{1}{m_S[{\rm M}_\odot]}\Big)^{2/3}(P[{\rm yr}])^{2/3}\Big(\frac{1}{D[{\rm pc}]}\Big)\,.$$

Some typical values for Jupiter mass and Earth mass planets are given in Table 2.5. From this table and Equation 2.23 the following dependencies are apparent:

- the astrometric signal decreases with the distance of the source $\theta \propto 1/D$, favoring thus strongly nearby systems,
- the astrometric signal is larger for longer orbital periods $\theta \propto P^{2/3}$, or proportional to the semi-major axis $\theta \propto a$,
- the signal is proportional to the mass of the planet $\theta \propto m_P$,
- the astrometric effect of a planet is larger for low mass star $\theta \propto 1/m_S^{3/2}$.

A strong signal is produced for nearby low mass stars, with massive giant planet on long orbits. Very interesting is the fact that the detection bias for astrometry with respect to orbital separation or orbital period is opposite to that of the radial velocity method. Because the current measuring limit is about 1 mas, no planets were detected up to now by the astrometric method. However, in the near future a precision as good as 20 μas is expected with the GAIA satellite and ground based astrometric interferometry.

Table 2.5: Astrometric signature for different planet - star systems

m_P	m_S	a [AU]	P	θ(10pc)	θ(100pc)
M_J	$M_\odot$	5.2	12 y	480 μas	48 μas
M_J	$M_\odot$	1.0	1.0 y	92 μas	9.2 μas
M_J	$M_\odot$	0.1	11.6 d	9.2 μas	0.9 μas
M_J	2.5 $M_\odot$	5.2	7.5 y	190 μas	19 μas
M_J	0.4 $M_\odot$	5.2	18.7 y	1200 μas	120 μas
M_E	$M_\odot$	1.0	1.0 y	0.29 μas	0.029 μas
M_E	0.4 $M_\odot$	5.2	18.7 y	3.8 μas	0.38 μas

2.5.2 Science potential of astrometry

The science goals of astrometric studies of the stellar reflex motion due to planets must consider that many properties of extra-solar planets are already known from RV surveys. Thus the astrometric studies should address questions which are complementary to the results from the RV survey. Let's assume that the next generation of instruments reaches an astrometric precision in the 10 - 100 μas range. This allows the detection of giant planets with orbits longer than a few years within 50 to 100 pc (see Table 2.5). In this case the following science topics can be investigated:

- Astrometry can detect giant planets around more massive stars, magnetically active stars, and young fast-rotating stars, which are hard to detect with the RV method. Astrometry can therefore provide an inventory of extra-solar planets around stars of all types.
- Planet masses m_S can be easily determined for objects detected by the RV-surveys but for which the $\sin i$ factor is not known. A few measurements are sufficient to determine the orbit inclination.
- Many stars show long term trends in their RV-data. Astrometry can clarify the presence of companions at large separation more easily because the astrometric signal increases linearly with a while the RV signal scales with $1/\sqrt{a}$.
- Astrometry can clarify whether multiple systems are coplanar or not. Dynamical interactions with planet can lead to eccentric orbits and tilts between orbital planets. Astrometry can measure the mutual inclination.
- With an astrometric orbit of the central star the position of an unseen planet can be determined. This is a most important information for the search of a planetary signal with high contrast imaging.

2.5.3 Astrometric motion of stars

The astrometric reflex motion of a star due to a planet is very small when compared to other astrometric motion components, which are:

- **the proper motion** of the center of mass of the system which is the velocity component projected on the sky with respect to the center of mass of the solar system. The proper motion of a star is described by the angular motion in right ascension μ_α and the angular motion in declination μ_δ. Typical values for the proper motion are of the order 0.1 – 1 arcsec/yr for nearby stars ($\approx$ 10 pc), and 10 – 100 mas for stars at about 100 pc.
- **the annual parallax** of a target is due to the motion of the Earth around the center of mass of the solar system. The size of this effect is strictly related to the distance:

$$\pi[\text{arcsec}] = \frac{1}{D[\text{pc}]}\,,$$

  and it is by definition 1 arcsec for a distance of 1 pc, 100 mas for 10 pc, 10 mas for 100 pc, etc.. The shape of the annual parallactic motion is an ellipse and its ellipticity depends on the direction of the line of sight with respect to Earth orbit (the ecliptic plane).

Both effects are about 2 to 3 orders of magnitudes larger than the typical signal induced by an orbiting giant planet. Thus, one needs to measure first accurately the proper motion and annual parallax, before one can aim for the detection of the astrometric reflex motion due to a planet.

Slide 2.16 illustrate the different astrometric motion components for a nearby low mass binary system: orbital motion, annual parallax, and proper motion. For a planetary system the orbital motion will be about 2 orders of magnitudes smaller (of the order mas) and there is in general no co-moving secondary component present which can be used as relative reference point.

2.5.4 Projected orbital motion

Astrometric measurements provides after the correction for the proper motion and the annual parallax the orbital motion as projected on the sky. From this one can derive the mass of the planet m_P, if the mass of the star is known and the orbital parameters P, ϵ, i, ω, Ω and T_0. There is no $\sin i$ ambiguity.

For the simple case of an intrinsically circular orbit the projected orbit is an ellipse and the ratio between the projected axes x (major) and y (minor) is $y/x = \cos i$.

For an intrinsically elliptic orbit the situation is more complicated. However, the orbital ellipticity and the "projection ellipticity" can be disentangled because the orbital ellipticity defines the temporal behavior of the observed motion. There remain only two solutions with an ambiguity about the near or far side of the orbit. This ambiguity must be solved with radial velocity measurements.

Fitting the astrometric orbits of visual binaries is a classical topic in astronomy (see e.g. Binnendijk 1960, Properties of double stars. Univ. Pennsylvania Press). Data for planetary systems face the problem that one needs to disentangle potentially the contributions of multiple planets to the reflex motion of the star, considering correctly the noise in the data.

2.5.5 Astrometric measurements

Astrometric measurements of the planet induced reflex motions are difficult. The measuring precision reached up to know is of the order 1 mas, which is comparable to the expected signal for an ideal (best) case of a nearby system with a giant planet with an orbital period of several years.

Due to these restriction, no extra-solar planets have been detected purely based on astrometric measurements up to now. However, it was possible to measure the astrometric reflex motion of a few stars with known RV-planets. A good example for the detection of a planet-induced astrometric motion is the nearby star ν And. The astrometric measurements for this system are shown in Slide 2.17. Thanks to many RV-data points most orbital parameters for the "two astrometric planets" c and d were already well known. Astrometry provided in addition the inclination $\sin i$ for the orbits and the masses of the two planets. Of much interest for the orbit dynamics of the system is the large mutual inclination between the two orbits of about $\Delta i = 30°$.

2.5.6 Expected results from the GAIA mission

GAIA is an all-sky, astrometric satellite which will measure the astrometric parameters of more than 10^9 stars, 10^7 galaxies, 10^5 quasars and 10^5 asteroids in the brightness range from 6 mag to 20 mag. The GAIA satellite was successfully launched in Dec. 2013 and the mission will last for about 5 years.

GAIA will scan the sky with a predefined, regular pattern and each object is observed about 70 times, or about 15 times per year. It will not be possible to adjust the observing strategy for improving the sampling of "interesting targets".

The GAIA instrument is a continuously rotating, double telescope which projects two sky regions separated by 120 degrees onto a huge array of more than 100 detectors with in total of more than 1 billion detector pixels (see Slide 2.18). The instrument peformes besides astrometry, also accurate photometry, and spectroscopy which allows stellar RV measurements with a precision of about 1 km/s. The detectors read continuously the detected signal and a powerful on board computer system, preprocesses the data, in particular it selects and transmits only the scientifically useful data down to Earth.

The expected astrometric precision of the satellite is better than 100 μas per single observation for stars brighter than 15 mag. The end of mission precision, after 50 - 100 measurements will be at the level of 20 μas, again for stars brighter than 15 mag. Thus GAIA is sensitive for the astrometric detection of giant planet around stars closer than about 100 pc with orbital periods in the range 0.5 – 5 years (see Table 2.5).

The GAIA mission will have an important impact in many fields of astronomy from solar system research, stellar astrophysics, galactic astronomy and cosmology.

The expected results for extra-solar planets are:

- many 1000 giant planets will be detected astrometrically,
- the orbits due to the reflex motion of about 500 giant planets will be measured with high precision allowing a determination of planet masses m_p with an accuracy better than 20 %,
- for about 100 planetary systems the orbital parameters of more than one planet can be measured and the orientation of their orbital planes can be investigated.

The GAIA detections will trigger many follow up studies using the RV method or direct imaging for the most interesting systems.

2.5.7 Ground based interferometric astrometry

Interferometry measures the interference pattern of the light of a star collected by two telescope separated by a distance B which is called the baseline. The exact angular position θ of an object in the plane defined by the baseline and the line of sight can be deduced by the external path length difference for the light reaching telescope 1 and telescope 2.

$$D_{\rm ext} = B \cos\theta \,. \tag{2.24}$$

Because a ground based interferometer rotates with respect to sky, different baseline orientation can be measured and the exact position α, δ of the object be determined. With

interferometry one obtains an interferometric wave pattern $I(D)$ as function of the path difference

$$I(D) \propto \sin D \quad \text{where} \quad D = D_{\rm ext} + (D_2 - D_1) + \frac{\lambda}{2\pi}\phi\,,$$

λ is the wavelength of the light, ϕ the phase, and $D_2 - D_1$ is the relative path difference for the light going through telescope 1 and 2, respectively. Figure 2.10 and Slide 2.19 illustrate the basic principle for interferometric astrometry. Important components are the two telescopes, the delay line with moving mirrors which compensate the changes of the external path length difference because of the Earth rotation, and the wave correlation laboratory.

Figure 2.10: Basic principle for interferometry and the double difference method used for astrometry.

In interferometry path differences can be measured with an accuracy of a small fraction of the phase ϕ, equivalent to a small fraction of the wavelength. For example, for a measuring precision of $\lambda/100$, equivalent to the path difference of ± 20 nm for IR light with $\lambda = 2\mu$m the position angle θ can be measured with a precision of

$$\Delta\theta = \pm\frac{\lambda}{100\,B}\,.$$

This yield for a 2-μm interferometer with a baseline of $B = 100$ m an angular precision of $20\text{nm}/100\text{m} = 10^{-10}$ or $2 \cdot 10^{-10}\text{arcsec} \cdot \pi/(180 \cdot 3600) = 10\,\mu$as.

In reality, this measurement is very difficult. Atmospheric turbulence introduces for the two telescopes path length variations which are larger than a wavelength. Also any variation in the path length inside the telescope and interferometer due to unstable air conditions and mechanical instabilities are harmful and must be under control. The following double difference strategy must be applied for a successful measurement of accurate astrometric positions:

– The astrometric position of a target is determined relative to a nearby background or reference star located within a few tens of arcsec. This requires that interferometric

measurements are made simultaneously for the target star and the background reference star. Instead of measuring $D_{\rm ext}$ one measures the relative difference between target and reference star

$$\Delta D = D_t - D_r = (D_{t,2} - D_{t,1}) - (D_{r,2} - D_{r,1}) = B(\cos\theta_t - \cos\theta_r)\,.$$

The big advantage is that both objects show the same path length variations introduced by the atmosphere and the instrument.

It is complicated to perform this double difference measurements. One needs to be able to measure the interference pattern of both stars simultaneously and correct the target interferogram "on-line" for the path length variation seen in the reference star interferometer. The measured "phase difference" between target and reference yields then the position of the target relative to the reference star.

PRIMA is the astrometric instrument at the VLT interferometer which is currently tested. Unfortunately the tests showed that the laser monitoring concept for the interferometer meteorology is not sufficient. This system should measure the differential path length variations of the light beams of the target star and the reference star in the interferometer and the telescopes, because mechanical vibrations, air turbulence and other differential effects between the 4 different beams must be corrected. An improved instrument monitoring system is now build before PRIMA astrometry becomes available.

2.6 Pulsar and transit timing

Timing studies are a third method to search for the reflex motion induced by extrasolar planets. However, this method works only for systems which produce a measurable and well defined periodic signals. Up to know there are only few planetary systems known for which this method could be applied:

- for two pulsars with planets,
- for the short period, eclipsing binaries consisting of a white dwarf and low mass star. A good example is the system HW Vir which harbors probably two circumbinary planets.
- for planetary transit signals mainly from KEPLER light curves.

Only the pulsar planets and the transit timing method are firmly established. The binary eclipse system require further confirmation with more data. In the following subsection two well established examples are given.

2.6.1 Planets around the pulsar B 1257+12

The milli-second pulsar B 1257+12 is famous because the first planetary mass objects were found around this object. Milli-second pulsars are very special astronomical objects. Pulsars are born in supernovae as collapsed cores of the former stellar iron-core of a high mass star which reached the Chandrasekhar mass limit and became unstable. Therefore one can expect that the pulsar B 1257+12 has a mass of about 1.4 $M_{\odot}$ like essentially all binary pulsars with mass determinations.

Pulsars are, very compact $R \approx 10$ km, highly magnetized (stellar magnetic fields compressed to small diameter), and fast rotating (angular momentum conservation!) objects. Because the magnetic axis is usually not perfectly aligned with the rotation axis, they accelerate electrons along the polar magnetic fields to relativistic speeds, so that the emitted synchrotron radiation from the electrons emit strongly pulsed radiation with a pulse period equal or half the orbital period for a bipolar magnetic field. Pulsars are born as extremely hot, highly magnetized, and fast rotating objecs with a rotation period of less than 0.1 sec and strong pulses. They slow down because the relativistic particles extract angular momentum so that pulsars become slower and weaker with time. For example, their pulse period doubles from 0.5 s to 1 s in a few million years. If a pulsar has slowed down to a rotation period of several seconds then their radio emission disappears and they are no more observable.

Pulsars can be "re-born" if they reside in close binary systems. In a compact binary mass from the companion can flow to the pulsar via an accretion disk. This spins up a previously "old", cold and low magnetic field pulsar. If the pulsar's rotation become faster than a rotation period of 20 ms, it starts to have again an radio signal. These reborn "old" milli-sec pulsars are therefore often found in binary systems. In some special cases the pulsar wind can evaporate the close companion and only an old, fast rotating millisecond pulsar is left. It's orbital period is extremely stable, because the pulsar magnetic field is relatively low, the pulsar wind is stable and the interior structure is settled. For some milli-second pulsars the stability of the pulse arrival times is higher than any man-made atomic clock with an evolution of the pulse period at a level of $\dot{P}/P \approx 10^{-16}$ or a stability of the pulse arrival times of about 1 ns over a full year.

The main points of this milli-second pulsar story for planet research are:

- milli-second pulsars are ideal clocks for timing studies,
- planetary mass object around such systems must have survived a SN explosion in a binary system, a strong pulsar wind, the evolution and evaporation of a companion or they were formed during or after one of these events,
- the interpretation of properties of pulsar planets must consider their very special nature.

On the other hand, pulsars are ideal targets for the search of planets with the timing method. Because of the presence of planets the pulsar position oscillates around the center of mass of the system and the pulse arrival time provides an exact position along the line of sight relative to the center of mass. Because pulse arriving times can be measured with a precision of about 0.01 ms any relative displacement of $0.01\text{ms} \cdot c = 3\text{km}$ becomes measurable. This means that even the reflex motion due to objects significantly less massive than Earth can be measured.

Planets around pulsars are rare. Besides the famous system B 1257+12 only one other pulsar B 1640-26 with a planet is known.

Properties of the B 1257+12 pulsar system: B 1257+12 is a milli-second pulsar with a period of 6.3 ms. It was studied in more details because it showed pulse arrival anomalies which turned out to be caused by three planetary mass objects. The measured deviations of the barycentric pulse arriving times from a constant value are shown in Slide 2.20 together with a fit for a three planet system. The residual scatter from this fit are of the order $\pm 10\,\mu$s. The derived parameters for the planets are:

- innermost planet $P = 25.3$ days, $a = 0.19$, $\epsilon = 0.0$, $m_P \sin i = 0.015\,\mathrm{M}_E$,
- second planet $P = 66.5$ days, $a = 0.36$, $\epsilon = 0.018$, $m_P \sin i = 3.4\,\mathrm{M}_E$,
- third planet $P = 98.2$ days, $a = 0.47$, $\epsilon = 0.026$, $m_P \sin i = 2.8\,\mathrm{M}_E$.

Note that the pulsar timing allowed in this case to find an object with the mass of the Moon. The second and third planet are close to an interesting $2 : 3$ orbital period resonance.

2.6.2 Transit timing for KOI 875

More and more transiting planets are detected. For all these objects an accurate transit timing can be used to search for additional unseen planets.

No transit timing variations (TTV) are expected for a single planet system. The transits will be strictly periodic. If a second or more planets are present then the position of the star is altered, it is not just on the other side of the center of mass C with respect to the transiting planet. The star can be displaced from the line through the planet and the center of mass because of other planets in the system. This means that the planet must move a bit less or a bit more than a full orbit until the next transit occurs.
Transit timing variations are particularly large for a transiting planet with a long period, because it moves with slower speed and any lateral displacement of the star from the center of mass will result in a longer transit time difference.

Figure 2.11: Geometry of transit timing variation (TTV).

A good example for the TTV effect is KOI 872. This system was noticed as Kepler Object of Interest (KOI) because it showed eclipses with a period of about $P_b = 33.6$ days. A detailed study of the transit times revealed that the transits vary in time by about ± 1 hour. These deviations are introduced by a second, non-transiting planet with a mass of about $M_c = 0.37 \mathrm{M}_J$ and an orbital period of $P_c = 57$ days (see Slide 2.21). In addition a transiting close-in planet with a radius of 1.7 R_E was found. The main result of the TTV effect are:

- KOI 872 b with a period of 33.6 days shows transit timing variations of up to ± 1 hr,
- KOI 873 c, a non-transiting planet, with a period of 57 days and a mass of 0.37 M_J is responsible for the large timing variations,
- the periodicity of the TTV provide the orbital period of planet c ,
- the mass of component c can be determined from the amplitude of the timing variations,
- the mass of the transiting planet b cannot be determined from the TTV data, however the modelling for the dynamic stability of the system requires $M_b < 6$ M_J,
- the effect of the innermost low mass planet on the transit timing are two small to be detected in the data.

In transiting systems the TTV effect is a basic tool for the investigation of additional planets in a system.

Chapter 3

Transits of planets: mean densities

Close-in (short period) planets have a relatively high chance to transit in front of the star. A transit introduces a small periodic dimming of the star which is for a given star proportional to the size of the planet. The photometric observation of transits provides therefore the orbital period and the radius ratio between planet and star. Because the radii of stars are quite well known, the planet radius can be derived and with the planet mass from the radial velocity method also the mean density. Planet transits have been detected for more than 1000 stars. Such a large sample yields important information, not only on planet frequency and orbit properties, but also on the composition and internal structure of different types of planets based on the mean density determinations.

In this chapter we discuss first the radius and mean density of solar system objects. They serve again as well studied test cases. Then we describe the transit technique for extra-solar planets and describe some important scientific results from transit observations.

3.1 The structure of solar-system planets

3.1.1 Radius, mass and mean density for planets

A rough characterization of a solar system planets, comparable to observations of extra-solar planets, can be based on the size, the mass, and the mean (or bulk) density.

The radii of solar system planets can be derived from angular diameter measurements. Because of their fast rotation the giant planets are flattened significantly (> 1 %). The difference between equatorial and polar radius R_e and R_p are given in Table 3.1. The flattening $f = (R_e - R_p)/R_e$ for the terrestrial planets is less than 1 %. It is useful to express the size of a planet by its volumetric mean radius R_V defined as radius of a spherical body with the same volume.

Masses of solar system planets and large bodies can be determined from the orbits of moons using Kepler's third law.

From the radius R and the mass M one can calculate the mean density $\bar{\rho}$ of a planet:

$$\bar{\rho} = \frac{M}{V} = \frac{M}{(4/3)\pi\, R^3}\,. \tag{3.1}$$

Table 3.1: Giant planet data: rotation period $P_{\rm rot}$, equatorial radius R_e, polar radius R_p, volumetric mean radius R_V, and flattening $f = (R_e - R_p)/R_e$.

Planet	$P_{\rm rot}$	R_e [km]	R_p [km]	R_V [km]	f
Jupiter	9^h55^m	71 490	66 850	69 910	6.5 %
Saturn	10^h39^m	60 270	54 360	58 230	9.8 %
Uranus	17^h15^m	25 560	24 970	25 360	2.3 %
Neptune	16^h06^m	24 770	24 340	24 620	1.7 %

Table 3.2 gives mass M, radius R, and ρ for the large bodies in the solar system. A few points are notable from Table 3.2:

- the objects with the highest density are the terrestrial planets Mercury, Venus and Earth with $\bar{\rho} \approx 5.5\,\mathrm{g\,cm^{-3}}$, followed by Mars with $\bar{\rho} \approx 4\,\mathrm{g\,cm^{-3}}$,
- the Earth's moon and the Galilean moons Io and Europa have intermediate densities of about $\bar{\rho} \approx 3.0$ to $3.5\ \mathrm{g\,cm^{-3}}$,
- all other large solid bodies have a density around $\bar{\rho} \approx 2.0\,\mathrm{g\,cm^{-3}}$,
- the gas planets and the sun have a low mean density of $\bar{\rho} \approx 0.7 - 1.7\,\mathrm{g\,cm^{-3}}$.

The main reason for the different mean densities are the different compositions, but also the different structure. As a very rough statement one can say, that high density objects $\bar{\rho} > 3.0\,\mathrm{g\,cm^{-3}}$ are composed of rocks and iron, intermediate density objects $\bar{\rho} \approx 2.0\,\mathrm{g\,cm^{-3}}$ contain in addition a substantial fraction of water ice, while for low density objects a substantial part of their volume is due to H and He.

3.1.2 Composition of solar system planets

It is not easy to derive the composition of planets because one cannot look deep into their interior. There are two very important sources of information about elemental abundances in the solar system:

- **The solar photosphere**, which is considered to show representative abundances for the convective envelope of the sun and the elemental abundances in the pre-solar nebula.
- **Carbonaceous chondrites**, a special type of primitive meteorite, which were not strongly altered by heating, melting, differentiation and other processes. Therefore, their abundances are considered to represent well the dust particles of the early solar system.

The abundances in the solar photosphere can be determined by the spectroscopic analysis of the solar spectrum. This method is accurate (± 10 %) for abundant elements with many absorption lines and well determined atomic data. Abundances of H, He, noble gases, and other volatile elements can be directly compared to heavy elements (e.g. Mg, Si, Fe). The abundance determination is difficult for rare elements with only few spectral lines in the solar spectrum.

The abundances of meteorites can be determined with very high accuracy with modern laboratory techniques. Abundances of very rare elements can be determined as well as

Table 3.2: Mean radius R, mass M and mean density $\bar{\rho}$ for large solar system objects with $R > 1000$ km. The mass M is given in Earth mass $M_{\rm E} = 5.97 \cdot 10^{24}$ kg.

Planet / moon	R [km]	M [$M_{\rm E}$]	$\bar{\rho}$ [g cm^{-3}]
sun	696'000	333'000	1.41
planets			
Mercury	2 440	0.0553	5.43
Venus	6 050	0.815	5.24
Earth	6 370	1.0	5.51
Mars	3 390	0.107	3.94
Jupiter	69 910	318.	1.33
Saturn	58 230	95.2	0.70
Uranus	25 360	14.5	1.30
Neptune	24 620	17.1	1.76
dwarf planets			
Pluto	1 160	$2.2 \cdot 10^{-3}$	2.0 (?)
Eris	1 160	$2.7 \cdot 10^{-3}$	2.5 (?)
moons			
Moon (E)	1 740	0.0123	3.35
Io (J)	1 820	0.0150	3.53
Europa (J)	1 560	$8.0 \cdot 10^{-3}$	3.01
Ganymede (J)	2 630	0.0248	1.94
Callisto (J)	2 410	0.0180	1.83
Titan (S)	2 580	0.0255	1.88
Triton (N)	1 350	$3.6 \cdot 10^{-3}$	2.06

values taken from en.wikipedia.org/wiki/list_of_Solar_System_objects_by_size.

the different isotopes for the elements. A major disadvantage is that volatile elements are strongly depleted, most notably H, He, C, N, O, S, and the noble gases. Especially for the elements H, C, O, S, there are strong differences between different chondritic meteorites, depending on the individual history of the sample. Abundances of many samples must be compared to exclude the possibility of abundances anomalies due to a special sample.

The solar and chondrite abundances form also the backbone for the determination of the universal "cosmic" abundances of the elements (Slide 3.1). Data from other stars and emission nebulae confirm that the abundances determined in the solar system are representative for the whole Universe.

The mass fraction $f_{\rm mass}$ of the most abundant elements is given in Table 3.3 for the solar photosphere, chondrites, the Earth and Jupiter. The sun, like the entire Universe, is essentially made of 98 % H and He plus about 2 % of "heavy" elements. Heavy elements include first O, C, Ne and N, then less abundant Fe, Si, Mg, and then the other elements. They can be explained by the stellar nucleo-synthesis theory (see also Slide 3.1). The composition for chondritic meteorites is very different because of the strongly reduced abundance of the volatile elements or "gases". For the abundances of Earth and Jupiter only the elements which contribute more than 1 % to the mass are given in Table 3.3 to emphasize the large differences in composition between these two planets. Of course also

Table 3.3: Abundances of the most frequent elements for the sun, chondrites, Earth and Jupiter expressed as mass fraction f_m and for the sun as atomic abundance relative to Si.

Z element	sun N(X)/N(Si)	sun $f_{\rm m}$	Chondrite $f_{\rm m}$	Earth $f_{\rm m}$	Jupiter
1 H	27200	74 %	2.0 %	<	71 %
2 He	2180	24 %	<	<	26 %
6 C	12.1	0.40 %	3.5 %	<	1 %
7 N	2.5	0.10 %	0.3 %	<	<
8 O	20.1	0.86 %	46.6 %	30 %	2 %
10 Ne	3.8	0.20 %	<	<	<
11 Na	0.06	<	0.5 %	<	<
12 Mg	1.08	0.07 %	9.6 %	12 %	<
13 Al	0.08	<	0.9 %	1 %	<
14 Si	1	0.08 %	10.7 %	19 %	<
16 S	0.52	0.05 %	5.2 %	1 %	<
18 Ar	0.10	0.01 %	<	<	<
20 Ca	0.06	<	0.9 %	1 %	<
24 Cr	0.01	<	0.3 %	<	<
26 Fe	0.90	0.14 %	18.5 %	33 %	<
28 Ni	0.05	<	1.1 %	2 %	<
< is:		< 0.01 %	< 0.1 %	< 1 %	< 1 %

estimates for the abundances of the rare elements are available. However, these are often inferred from abundance ratios from the sun or chondrites, because some pairs of elements (e.g. Ni/Fe) are expected to behave very similar. Alternatively the abundance ratio Si/Fe follows from determinations of the size of the planet core based on models for the internal structure.

3.1.3 Differentiation: the example of Earth

Terrestrial planets were formed by the accumulation of solid bodies with elemental abundances similar to (chondritic) meteorites. There are four elements which dominate:

- O - oxygen: mostly bound in silicates X-SiO_x, but also other minerals like MgO,
- Fe - iron: bound in FeS, in silicates $FeSiO_x$ and other minerals, but if melted then it accumulates as iron metal or melt,
- Si - silicon: main components of silicate rich rock, quartz (sand),
- Mg - magnesium: important constituent of silicate rich rock $MgSiO_x$.

The interior of Earth has a temperature of a few 1000 K because of internal heating by the decay of radioactive nuclei. This temperature is high enough that the rock/magma and iron in the Earth interior behave like fluids on geological time scales. This means that Earth had time for sedimentation into layers of progressively higher density towards the center. This explains the formation of an iron core and the silicate-rich mantle. This structure is well established based on the analysis of seismic waves. Due to the

differentiation the surface composition is not representative for the average composition of a planet, especially if predominant constituents have a high density and accumulate in the center.
Thus the material composition and rough structure of Earth is:

- 33 % of the mass are Fe or iron alloys, mainly in the core,
- 67 % of rock-like material mainly based on silicates (Mg,Fe)SiO_x in the mantle,
- the water H_2O in the oceans contributes only 0.024 % to M_E,
- the N_2 and O_2 in the atmosphere is only 0.0001 %.

This very simple structure description for the planet Earth is of course far from a geological description of the inner structure of this planet. However, this is a useful starting point for the comparison of the structure of Earth with other solar system bodies and extra-solar planets.

3.2 Basic equations for the structure of planets

The radial model structure of a spherical planets can be derived from three basic equations:

The hydrostatic equation

$$\frac{dP(r)}{dr} = -g(r)\rho(r) = -\frac{GM_r(r)}{r^2}\rho(r)\,, \tag{3.2}$$

where $P(r)$ is the pressure, $\rho(r)$ the density, $M_r(r)$ the mass inside radius r, and $g(r) = GM_r(r)/r^2$ the gravitational acceleration (G the gravitational constant).

The equation for mass conservation

$$\frac{dM_r(r)}{dr} = 4\pi r^2\rho(r)\,. \tag{3.3}$$

The equation of state (EOS) which describes the density as function of pressure, temperature and composition:

$$\rho(r) = \rho(P, T, \text{composition})\,. \tag{3.4}$$

The EOS cannot be self-consistently derived from the above equations because the temperature cannot be determined without equations for the energy generation and the energy transport. Therefore, assumptions or prescriptions of the temperature are required. Also the composition must be given as input parameter for the modelling.

The following boundary conditions are valid for self-gravitating sphere:

$$M_r(r) = 0 \quad \text{at} \quad r = 0\,, \quad \text{and} \quad P(r) = 0 \quad \text{at} \quad r = R\,,$$

where R is the radius of the planet. For gaseous planets there is not a well-defined outer boundary and often the boundary condition $P(R) = 1$ bar (10^5 Pa) is used.

The modelling can be substantially simplified if the equation of state depends only on the pressure and a predefined radius-dependent composition according to

$$\rho(r) = \rho(P(r), \text{composition}(r))\,.$$

In this case the three equations can be solved iteratively and one obtains density, pressure and composition as function of radius.

3.2.1 Central pressure for a homogeneous planet

First we consider the very simple model case of a homogeneous planet. The assumption $\rho(r) = \bar{\rho}$ is a very rough over-simplification because there are strong composition and therefore density changes in planets and the materials may be compressible. For small solar system bodies, the assumption of a homogeneous structure may be acceptable.

Nonetheless the central pressure $P_0 = P(r = 0)$ of a homogeneous sphere yields a first guess on the typical pressures in planets which must be considered for the equation of state.

The pressure profile and the central pressure follows from an integration of the equation for the hydrostatic equilibrium using $M_r(r) = (4/3)\pi r^3 \bar{\rho}$. The pressure profile is:

$$P(r) = -\int_r^R \frac{dP}{dr} dr = \int_r^R \frac{GM_r(r)}{r^2} \bar{\rho}\, dr = \frac{4\pi G \bar{\rho}^2}{3} \int_r^R r dr = \frac{2\pi G \bar{\rho}^2}{3} (R^2 - r^2)\,.$$

This can be written in the simple form

$$P(r) = P_0 \left(1 - \frac{r^2}{R^2}\right)\,, \tag{3.5}$$

with the central pressure

$$P_0 = \frac{2}{3}\pi G \bar{\rho}^2 R^2 \quad \text{or} \quad P_0 = \frac{3GM^2}{8\pi R^4}\,, \tag{3.6}$$

where for the second equation the $\bar{\rho}$ is replaced according to Eq. 3.1.
For a given mean density $\bar{\rho}$ the central pressure increases like

$$P_0 \propto R^2 \propto M^{2/3}\,,$$

and for a given mass like

$$P_0 \propto 1/R^4\,.$$

Table 3.4 gives P_0 for several solar system objects using Eq. 3.6 and R and M from Table 3.2. The central pressures derived with the homogeneous planet $\bar{\rho}$ approximation are already accurate within a factor of a few for the terrestrial planets. The discrepancy is larger for the giant planets because they have a low density envelope $< 1\,\mathrm{g\,cm^{-3}}$ and a high density core $> 10\,\mathrm{g\,cm^{-3}}$.

Table 3.4: Central pressures P_0 calculated for homogeneous planets and based on detailed model calculations

object	R/R_E	M/M_E	$P_0(\bar{\rho})$	P_0 model[1]	T_0 model[1]
Moon	0.273	0.0123	47 kbar	45 kbar	1 800 K
Mercury	0.383	0.0553	240 kbar	400 kbar	2 000 K
Earth	1	1	1.7 Mbar	3.6 Mbar	6 000 K
Jupiter	11.0	318.	12 Mbar	80 Mbar	20 000 K

1: central pressures and temperatures from detailed modelling taken from the compilation of de Pater and Lissauer

3.2.2 Phase diagrams

The equation of state relates the pressure, density and temperature for the different materials in a planet. For the structure equation, we need a relation for the density as function of radius $\rho(r)$ or as function of pressure $\rho(P)$. The density of a material depends on the phase, which can be described in phase diagrams (Fig. 3.1).

.

Figure 3.1: Generic $P-T$ phase diagram with the different dividing lines between phases.

The triple points, dividing lines, and the different phases in the phase diagram (Figure 3.1) can be characterized as follows:

- **The triple point** in the low T - low P domain defines the areas for the solid, liquid, and gas phases.
- **The critical point** defines the transition between the gas or liquid phase and the super-critical fluid phase.
- **Solid** materials are relatively cold or under high pressure. The density of solids varies between about $\rho \approx 1$ and $10\,\mathrm{g\,cm^{-3}}$. Depending on the pressure the materials can have different solid state structures (crystalline structures), with more compact = higher density configurations under high pressure. The density changes typically less than a factor of 2 between the different solid state structures.
- **Liquid** materials have a temperature above the triple point and below the critical point. Often a liquid becomes a solid if the pressure is strongly enhanced and it evaporates for very low pressures. The density of a liquid changes only little (< 30 %) if pressure or temperatures are changed.
- **Gas** occurs for material under relatively low pressures, below the critical point or below the triple point. If gas is further heated it dissociates and becomes an **atomic gas**, which may become an **ionized plasma** for even higher temperatures. Gas behaves like (ideal gas)

  $$\rho(P) = \frac{P}{\mathcal{R}\mathcal{T}}\,\mu\,,$$

  where μ is the mean particle mass. Thus, a gas is compressible and the density increases for a given temperature like $\rho \propto P$ until a critical density is reached at which a cold gas condenses or a hot/warm gas becomes a super-critical fluid.

– **Super-critical fluids** are at temperatures and pressures which are both above the critical point. Under very high pressures a super-critical fluid goes into a solid state. A super-critical fluid has properties of the liquid (can solve other materials) and the gas phase. For example, a super-critical fluid fills the whole volume of a container homogeneously and does not form a transition between a gas in the upper part and a liquid in the lower part of the container.

Figure 3.2: Behavior of liquid/gas and a super-critical fluid in a container

In particular, super-critical fluids are **compressible** and their density increases if the pressure is enhanced. The dependence $\rho(P,T)$ can not be described by a simple relationship.

– **Dissociation of molecules** in the gas phase or the super-critical fluid phase occurs in the temperature regime of 1000 K or a few 1000 K typically. The bindings between the individual atoms is broken up and the gas contains only atoms.
– **Ionization** of atoms occurs at temperature above a few 1000 K and if the temperature is further enhanced then atoms can be progressively ionized.
– **Pressure ionization** of solids or super-critical fluids occurs in the density regime of the order Mbar (10^{11} Pa). These pressures are high enough to squeeze together the nuclei in the material despite the Coulomb forces. The nuclear potentials start to overlap and the electrons can move freely between the nuclei - the material is pressure ionized. Many properties are similar to metals (e.g. electric conductivity). Pressure-ionized materials are compressible with the approximative dependence

$$\rho(P) \propto P^{1/2} \,.$$

Figure 3.3: Schematic difference for solid and pressure ionized materials.

3.2.3 Phase diagrams for the solar system planets and moons

We discuss in this section strongly simplified (qualitative) phase diagrams for hydrogen, water and iron in order to get the general picture about the EOS for solar system objects (Slides 3.2 to 3.4). The diagrams include points for the central pressure and temperature of objects, as well as lines and a few values which illustrate the run of parameters from the center towards the surface of the highlighted objects given in red. All diagrams cover the identical parameter space for an easy comparison. Thermodynamic parameters for hydrogen, water and iron are given in Table 3.5.

Table 3.5: Thermodynamic quantities of important materials for planets; melting $T_{\rm melt}$ and boiling temperatures $T_{\rm boil}$ and density ρ at $P = 1$ bar and $T_{\rm CP}$ and $P_{\rm CP}$ for the critical point.

material	1 bar properties			critical point	
	$T_{\rm melt}$	$T_{\rm boil}$	ρ	$T_{\rm CP}$	$P_{\rm CP}$
H_2	14 K	21 K	0.09 mg cm^{-3}	33 K	13 bar
H_2O	273 K	373 K	1.0 g cm^{-3}	647 K	220 bar
Fe	1810 K	3130 K	7.9 g cm^{-3}	9250 K	8.8 kbar

Hydrogen phase diagram. Hydrogen is an important constituent for the giant planets. Because it is a light element it will be located outside the planet core which is composed of heavier materials.

In Jupiter and Saturn hydrogen is the dominant constituent. Pressure-ionized (metallic) liquid hydrogen with a density of about 4 g cm^{-3} at a pressure of 10 Mbar is expected outside the central rocky/iron core. The transition to a super-critical fluid (around 1 Mbar) occurs for Jupiter at about 0.8 R_J and in Saturn at about 0.5 R_S. As super-critical fluid the hydrogen has a density of the order 0.7 g cm^{-3}. H_2 gas is only present close to the surface (Slide 3.2).

Neptune and Uranus are also called ice-giants because a substantial fraction of their mass is made up by icy materials, mostly H_2O, but also CH_4 and NH_3 ice. Hydrogen as light super-critical fluid $\rho = 0.7$ g cm^{-3} occurs outside > 0.7 R_N of the ice - rock core and extends up to the hydrogen gas atmosphere.

H_2O phase diagram. Water is an important constituent in the ice giants Neptune and Uranus and in TNOs and the moons of the giant planets. We take Europa as an example for a "solid" icy body.

In Neptune and Uranus a zone of liquid (metallic) pressure-ionized ice in the form of H_3O^+, OH^- is predicted from about 0.2 to 0.7 R_N. The density is about 2.5 to 5 g cm^{-3}. Inside of the ice zone is a rocky/iron core and outside the hydrogen envelope discussed above.

For Europa and other solid bodies in the outer solar system, water is the lightest material and it forms therefore the outermost layer. The internal temperature of Europa is probably high enough that there exists an ocean of liquid water below the surface crust of water ice. For smaller bodies than Europa the temperature might be too low for the presence of water in liquid form.

Fe phase diagram. Iron is the most abundant heavy material in planets and therefore it is a main constituent in the cores of the planets. The melting temperature of iron at low pressure is around 1500 - 2000 K and it depends a lot on the presence of other elements, like sulfur, nickel and others. At high pressure > 1 Mbar $T_{\rm melt}$ is enhanced to values above 4000 K.

For Earth, the largest terrestrial body in the solar system, the iron core is in a pressure-ionized solid (metallic) phase, surrounded by a layer of liquid iron. In smaller bodies the central pressure is not high enough for pressure ionization and the temperature is probably too low for liquid iron. Thus, one may assume that all other terrestrial planets and differentiated asteroids have a solid iron core. The presence or absence of liquid iron may also explain the presence or absence of strong magnetic fields for terrestrial planets.

The giant planets have pressure-ionized cores. It is not clear whether the iron and rocky materials in their cores are differentiated, or whether the core is too stiff for sedimentation.

Rocky materials. Much of the Fe phase diagram applies also for rocky materials. The melting temperatures for rocks depends in a very complex way on the rock composition. However, roughly $T_{\rm melt}$ of rocks are comparable to iron so that one can assume that rock is solid in terrestrial bodies. Exceptions are the largest terrestrial planets Earth (and perhaps Venus). Another exception is Jupiter's moon Io, for which the interior is heated up by tidal forces. When terrestrial planets were young their internal temperature was higher and therefore they had sufficient magma and melts for volcanic activity (e.g. Mars, Venus and the Moon). The rock in the center of giant planets is at such high pressures that it is in the solid pressure-ionized phase.

3.2.4 Simple approximation for the equation of state

Because the observational data for M and R of extra-solar planets are not very accurate, a simplified treatment of the equation of state (EOS) can be chosen for an assessment of the internal structure. According to the study of Seager et al. (2007) the EOS for different solid and liquid materials can be described with a modified "polytropic" equation:

$$\rho(P) = \rho_0 + cP^n\,, \tag{3.7}$$

where ρ_0 is the low pressure density of the considered material, and c and n are material dependent parameters. The $\rho(P)$ curves for some important materials are shown in Slide 3.5

Equation 3.7 and the curves in Slide 3.5 indicates that the planetary material is incompressible for pressures significantly below the critical pressure $P < P_{\rm crit}$ while for $P > P_{\rm crit}$ the material becomes compressible or the density behaves like a polytrope $\rho(P) \approx cP^n$ with $n \approx 1/2$. The critical pressure is defined by $P_{\rm crit} = (\rho_0/c)^{1/n}$. The curves for all materials are similar because the underlying physical cause for the density is the balance between gravitational forces and the Coulomb forces of the electrons and nuclei.

– At low pressures $P < P_{\rm crit}$ the density is essentially constant $\rho(P) = \rho_0$, because iron, rock, water-ice, and carbon are solid or liquid in the planets (Slide 2.15 and 2.16) and the density of the material is defined by atomic Coulomb forces (solid state forces).

Table 3.6: The parameters ρ_0, c and n for the EOS $\rho(P) = \rho_0 + cP^n$ of important materials (Seager et al. 2007). $P_{\rm crit}$ indicates at which pressure the material changes from incompressible to compressible:

material	ρ_0 [kg m^{-3}]	c [kg m^{-3} Pa^{-n}]	n	$P_{\rm crit}$
H_2O	1460	$3.11 \cdot 10^{-3}$	0.513	114 GPa
C (graphite)	2250	$3.11 \cdot 10^{-3}$	0.514	251 GPa
SiC	3220	$1.72 \cdot 10^{-3}$	0.537	479 GPa
$MgSiO_3$ (perovskite)	4100	$1.61 \cdot 10^{-3}$	0.541	693 GPa
$(Mg,Fe)SiO_3$	4260	$1.27 \cdot 10^{-3}$	0.549	770 GPa
Fe(α)	8300	$3.49 \cdot 10^{-3}$	0.528	1192 GPa

- At higher pressure $P < P_{\rm crit}$ the material becomes compressible because the gravitational forces are so large that the atoms are squeezed together to distances which are smaller than the typical size of atoms. This means that the material becomes "pressure ionized". The electrons "float" like in a metal.
- The critical density for pressure ionization depends on the material but lies in the range 1 - 12 Mbar (100 - 1200 GPa).
- For very high pressures $P \gtrsim 300$ Mbar the electron gas becomes degenerate. In this state the electron density is defined by the Pauli exclusion principle. This principle states, that fermions of a given spin can only be stacked in a momentum-location cell of $\Delta p \Delta V = h^3$. This state is called "degenerate gas" and it becomes important for objects more massive than $\approx M_{\rm J}$. We will discuss this equation of state in the next section.

For hydrogen this simple description is not applicable because the hydrogen density is not constant at pressures < 1 Mbar, because it is in then in the phase of a compressible super-critical fluid.

3.2.5 Interior structure of solar system planets

The interior structure of the planets can now be summarized qualitatively.

Earth and the terrestrial planets. The reference model for Earth interior is plotted on Slide 3.6. The diagram gives the PREM density structure, and rough indications about the pressure and temperature within Earth. The other terrestrial planets have a similar internal structure, with an iron core and a rocky mantle. The size of the core is different for Venus, Mars, Mercury and the Moon. Important differences are that the smaller objects Mars, Mercury and Moon are not or much less-pressure ionized and have therefore flatter density profiles for the iron core or the mantle, because the material is not compressed. It is also not clear whether there is liquid iron in Venus and Mercury.

Important properties for the terrestrial planets and the Moon just based on their mean density parameter:

- The mean density of Earth and Venus is above 5 g cm^{-3} because their interior is compressed due to the high pressure.

- Mercury is too small for significant compression and its high density is due to a large overabundance (e.g. with respect to Earth) of iron and a correspondingly large iron core. This property of Mercury is explained by the strong irradiation by the sun at this small separation and the partial evaporation of silicate grains during the planet formation process.
- The Moon has a low mean density indicating that there is only a small iron core present. The low mean density is one important fact pointing to the formation of the Moon by mantle material from Earth after a large collision of Earth with a Mars-sized body.

Jupiter and Saturn. Figure 3.4 gives the structure of Jupiter and Saturn. Both planets have first an outer region of molecular hydrogen which translates at the pressure of about 1 Mbar into metallic hydrogen. There must be a core of high density material, presumably a layer of O, C, and N rich "ices" and then a central rocky core. From the molecular abundances measured at the surface it is impossible to estimate the abundance of heavy elements (Si and Fe) and the size of the central high density core.

Figure 3.4: Interior structure for Jupiter (top) and Saturn (bottom).

Neptune and Uranus For Saturn and Jupiter one can at least assume that H and He dominates for radii larger than > 0.15 R. The situation is more ambiguous for Uranus and Neptune because the molecular H, He, CH_4 envelope extends only down to $r \approx 0.7$ R. It is assumed, but not sure, that there is an extended region of O-, N- and C-rich materials (ices) and then presumably a rocky core within $r < 0.2$ R.

Figure 3.5: Interior structure for Neptune.

3.3 Mass - radius relation for planets

In the previous sections we have discussed in detail the internal structure of solar system planets and the equation of state for the different materials for high pressure. Here we repeat a few basic principles in the context of the mass - radius relationship for substellar objects.

3.3.1 Low mass planets

For a terrestrial low mass planet, the densities of the different radial layers is given by ρ_0 from Table 3.6. The radius of the planet increases then with mass just like $R \propto M^{1/3}$

$$R = \frac{4\pi}{3\bar{\rho}} M^{1/3} \,. \tag{3.8}$$

The mean density $\bar{\rho}$ is given by the composition. For example, a small ($< \mathrm{M}_E$) terrestrial planet with $\bar{\rho} > 4$ g cm^{-3} must contain a substantial fraction of iron and a planet with $\bar{\rho} < 2.5$ g cm^{-3} a lot of ices.

This general relationship is also a good approximation for low mass < 0.1 M$_J$ giant planets made of H and He, because the H and He material are not yet strongly compressed under these conditions.

The situation is a bit more complicated for Neptune like gaseous planets. like Neptune, with a solid core of a few M$_E$. The radius of such objects changes quite rapidely if the mass of the hydrogen envelope is changed. Because the gravitation is not so high, a relatively small amount of H in gaseous form increases strongly the radius by a factor of about 1.5, because the H-gas adds an extended low density $\rho < 1$ g cm^{-3} "envelope" around the core, which reduces significantly the mean density from a mean core value of $\bar{\rho} \approx 3 - 5$ g cm^{-3} to $1 - 2$ g cm^{-3}.

3.3.2 Degenerate high mass planets and brown dwarfs

Under high pressure, like in the centers of high mass ($> \mathrm{M}_J$) planets, the material is pressure ionized and therefore compressible. This state is also called electron degenerate matter. According to the Pauli-principle two spin-1/2 particle (Fermions) cannot exist at the same location (quantum cell dV_x) in the same quantum-mechanical state. Possible states for electrons are two spin orientations and different momenta $p = m_e v$, where the "cell size" in momentum space is d$V_p = h^3$ (h=Planck constant).

For a degenerate gas the density of electrons can be described by the number of electrons in momentum space

$$n_e = 2 \cdot \frac{4}{3\pi} \frac{1}{h^3} \quad p_0^3 \,, \tag{3.9}$$

where p_0 is the Fermi-momentum and $E_0 = p_0^2/2m_e$ the corresponding (kinetic) Fermi-energy. For a cold, fully degenerate electron gas all low energy (or low momentum) states are occupied up to the Fermi-energy or Fermi-momentum:

$$p_0 = \Big(\frac{3h^3}{8\pi}\Big)^{1/3} n_e^{1/3} \approx \frac{h}{2} n_e^{1/3} \,. \tag{3.10}$$

For an ideal gas the pressure P is related to the kinetic energy, and therefore also the momentum, of the gas particles, according to:

$$P = nkT = \frac{2}{3}n\langle E\rangle = \frac{2}{3}n\langle\frac{p^2}{2m}\rangle\,.$$

This relation is also valid for a degenerate electron gas where the mean electron energy is related to Fermi energy by $\langle E\rangle = 3/5 \cdot E_0$. This yields then an equation of state for a degenerate electron gas describing the relation for the electron gas pressure and the electron density

$$P_e = \frac{2}{5}\,n_e \cdot E_0 = \Big(\frac{3}{8\pi}\Big)^{2/3}\frac{h}{5m_e}\,n_e^{5/3}\,. \tag{3.11}$$

This is essentially equivalent to the description of the high pressure regime of Table 3.6 describing the compressibility of matter at $P \gg 100$ GPa (or $\gg$ Mbar)

$$n_e \propto \rho \propto P^{3/5}\,.$$

What happens phenomenologically? An electron degenerate material parcel with a volume V_1 is squeezed under enhanced gravitational pressure to a smaller volume $V_2 < V_1$, so that the Fermi-momentum and the associated pressure of the electron "gas" is enhanced in order to reach a new pressure equilibrium. In this state the gas pressure does not depend on temperature as long as the kinetic momentum of the electrons due to the temperature is smaller than the Fermi momentum. This is the case for all "cold" objects.

Figure 3.6: Relation between volume and Fermi momentum of electron degenerate matter under low and high pressure P_1 and P_2.

The equation of state for degenerate matter has a most important effect on the mass-radius relationship of substellar objects. In a hydrostatic equilibrium the gravitational pressure P_G is equal to the electron pressure P_e:

$$P_G \propto \frac{GM^2}{R^4} \propto P_e \propto n_e^{5/3} \propto \Big(\frac{M}{R^3}\Big)^{5/3} = \frac{M^{5/3}}{R^5}\,.$$

This yields the mass-radius relation for degenerate matter

$$R \propto M^{-1/3}\,. \tag{3.12}$$

The exact relationship requires the consideration of the density profile through the planet, instead of a simple consideration of a mean density. However, the relation given above

describes roughly the functional dependence. For example a white dwarf star is an electron degenerate "cold" objects with $M_{\rm wd} \approx 10^3\,{\rm M}_J$ and $R_{\rm wd} \approx 10^{-1}\,{\rm R}_J$.

The most important message is, that cold substellar object in the regime $M > 3\,{\rm M}_J$ become smaller in radius if more mass is added. The maximum radius of cold substellar object is therefore

$$R_{\rm max} \approx R_J\,,$$

and brown dwarfs are slightly smaller than Jupiter.

Slide 3.7 and 3.8 illustrate the mass-radius relationship. Slide 3.7 is based on an analytic result for spheres with different (homogeneous) composition. This shows that the radius of an object depends on its composition. The curve X=0.75 is the result for a model with 75 % of H and 25 % of He. White dwarf stars are composed of C and O and the curve for C reaches for 1 $\rm M_\odot$ the expected radius of 0.01 $\rm R_\odot$ which is the same radius like for a C rich planet-mass object with $5 \cdot 10^{-5}$ $\rm M_\odot$. The mean density of such an electron degenerate white dwarfs is therefore very extreme with $\bar{\rho}$ $approx$1000 $\rm kg\,cm^{-3}$. Slide 3.8 illustrate the mass-radius relation calculated for stars and substellar gaseous objects. This diagram shows a strong break in the relation at 0.08 $\rm M_\odot$, where the nuclear hydrogen burning of stars sets in. For stars on the main sequence the kinetic momentum of the electron p_e due to the temperature is larger than the Fermi-momentum. These stars behaves like ideal gas spheres and the mass-radius relationship is roughly

$$\frac{R}{\rm R_\odot} \approx \frac{M}{\rm R_\odot}\,. \tag{3.13}$$

In addition the mass-radius relationship in Slide 3.8 is less peaked in the substellar regime when compared to the analytic solution. Detailed calculations take more accurately into account the different phases and the internal thermal energy which is still generated due to the ongoing contraction of brown dwarfs and giant planets.

3.4 Transiting planets

A transits of a planet in front of its parent star occurs if the line of sight is very close to the orbital plane. The transit probability is thereby much enhanced for planets with small separations. The unexpected presence of hot Jupiters made transit observation after the detection of 51 Peg b an attractive and very successful planet search technique (Slide 3.9). Transit data form now a cornerstone for extra-solar planet research.

Transits have the following basic characteristics (see Fig. 3.7 and Slide 3.10):

- Transits occur periodically,
- the stellar intensity is reduced by an amount which is proportional to the size of the planet,
- the planet occults during the transit different regions of the star, what may introduce during the transit photometric and spectroscopic features due to distinct surface structures of the star,
- during the transit some light passes through the outermost atmosphere of the planet what may introduce measurable absorption effects and allow an investigation of the uppermost atmosphere of the planet.

Secondary eclipses occur for most transiting planets about half an orbital period before or after the transit. If the emission of the planet is strong and the sensitivity of the measurement high then one can measure the drop in intensity when the planet goes into secondary eclipse.
Orbital phase curves may also be detected, because the hemisphere facing the star is expected to be brighter than the planet's night side. This requires that the measuring sensitivity is at least accurate enough for secondary eclipse observations.

Figure 3.7: Light curve features for transiting planets.

3.4.1 Approximations for basic transit parameters

We discuss here first approximate transit parameters for planets on a circular orbit and for central transits. This simplified treatment illustrates the basic transit properties. More detailed derivations and dependencies of the transit parameters are described in the following sections.

The transit depth is the fractional reduction of the stellar intensity $\Delta I/I$ due to the planet transit. Assuming a homogeneous stellar disk and treating the planet like a black disk yields

$$\frac{\Delta I}{I} \approx \frac{R_P^2}{R_S^2} \,. \tag{3.14}$$

The transit depth is equal to the ratio of the cross sections of the planet to the star. This measurement yields the absolute radius R_P of the planet, if R_S of the star is known. If also the mass of the planet is known, then one gets the mean density $\bar{\rho}$ of the planet, which is a very important parameter for the planet characterization.

For a Jupiter – Sun system the transit depth is about 1 %, and for an Earth – Sun system about $1.0 \cdot 10^{-4}$ (see Table 3.7).

The transit duration of a planet on a circular orbit across the center of the star is

$$\Delta t \approx \frac{P}{2\pi a} \, 2R_S \,. \tag{3.15}$$

where R_S is the stellar radius. This is an upper limit for circular orbits because non-central transits are shorter. This formulation defines the transit duration from the mid-ingress to the mid-egress phase.

We can express the transit duration as fraction of the orbital period:

$$\frac{\Delta t}{P} \approx \frac{R_S}{\pi a} \,.$$

Table 3.7: Transit properties of solar system planets for an "outside" observer. P is the orbital period, Δt the absolute and $\Delta t/P$ the relative transit duration, $\Delta I/I$ the transit depth, $p_{\rm trans}$ the transit probability, and i the orbit inclination

planet	P [yr]	Δt [hr]	$\Delta t/P$	$\Delta I/I$	$p_{\rm trans}$	i
Mercury	0.241	8.1	$38 \cdot 10^{-4}$	$1.2 \cdot 10^{-5}$	1.19 %	6.33
Venus	0.615	11.0	$20 \cdot 10^{-4}$	$7.6 \cdot 10^{-5}$	0.65 %	2.16
Earth	1.000	13.0	$15 \cdot 10^{-4}$	$8.4 \cdot 10^{-5}$	0.47 %	1.65
Mars	1.880	16.0	$9.7 \cdot 10^{-4}$	$2.4 \cdot 10^{-5}$	0.31 %	1.71
Jupiter	11.86	29.6	$2.9 \cdot 10^{-4}$	1.01 %	0.089 %	0.39
Saturn	29.5	40.1	$1.5 \cdot 10^{-4}$	0.75 %	0.049 %	0.87
Uranus	84.0	57.0	$0.77 \cdot 10^{-4}$	0.135 %	0.024 %	1.09
Neptune	164.8	71.3	$0.49 \cdot 10^{-4}$	0.127 %	0.015 %	0.72

The transit probability describes the chance that a planet shows period transits. For systems with circular orbits a transit will occur for

$$a|\cos i| < R_S\,.$$

This condition considers only full transits and grazing transits where at least half of the planet is in front of the star. For the integration we need to consider that all inclinations in the range $i = 90^\circ \pm \theta$ are considered and this needs to be normalized to a random distribution of orbit orientations according to:

$$p_{\rm trans} = \frac{\int_{90-\theta}^{90+\theta} \sin\vartheta d\vartheta}{\int_0^{180} \sin\vartheta d\vartheta} = \frac{-\cos\vartheta|_{90-\theta}^{90+\theta}}{-\cos\vartheta|_0^{180}} = \cos(90^\circ - \theta)\,,$$

or with the relations for $\cos i$ follows the approximate transit probability:

$$p_{\rm trans} \approx \frac{R_S}{a}\,. \tag{3.16}$$

3.4.2 Detailed transit geometry

The basic transit geometry, which is equivalent to the classical eclipse description of binary stars, is illustrated in Fig. 3.8. The description for the transit is also valid for the secondary eclipse of the planet behind the star.

Figure 3.8: Transit times.

– Four **transit times** or eclipses times are defined: the start and end of the ingress phase t_i and t_{ii} and the start and end of the egress phase t_{iii} and t_{iv} which are also called the $1^{\rm st}$, $2^{\rm nd}$, $3^{\rm rd}$ and $4^{\rm th}$ contact. In addition there is of course the mid-eclipse time $t_{\rm ecl}$.

– For the **transit duration** or eclipse duration one distinguishes between the total (transit or eclipse) phase which is the time span where at least some parts of one object is behind the other object while the full eclipse phase is the time span where one object is located completely in front or behind the other object.

$$\Delta t_{\rm tot} = t_{iv} - t_i \quad \text{and} \quad \Delta t_{\rm full} = t_{iii} - t_{ii}$$

– The **impact parameter** b is the minimum projected distance of the center of the two objects at mid-eclipse.

– A **partial transit** or a partial eclipse occur for $R_S - R_P < b < R_S + R_P$ without full phase $\Delta t_{\rm full} = 0$.

The transit timing can be calculated from the radii of the star R_S and the planet R_P and the orbital parameters. For circular orbits we can write for the transit times relative to mid-eclipse $t_{\rm ecl} = 0$:

$$t_{iv} = -t_i = \frac{P}{2\pi a}\sqrt{(R_S + R_P)^2 - b^2} \quad \text{for} \quad b \le R_S + R_P \tag{3.17}$$

and

$$t_{iii} = -t_{ii} = \frac{P}{2\pi a}\sqrt{(R_S - R_P)^2 - b^2} \quad \text{for} \quad b \le R_S - R_P\,. \tag{3.18}$$

The total and full transit duration are twice these values, respectively. From the full and and total transit duration one can derive the impact parameter (square the equation above and solve for b^2, if the ratio between the radii R_P/R_S is known, e.g. from the transit depths:

$$b^2 = R_S^2\,\frac{(1 - R_P/R_S)^2 - (\Delta t_{\rm full}/\Delta t_{\rm tot})^2(1 + R_P/R_S)^2}{1 - (\Delta t_{\rm full}/\Delta t_{\rm tot})^2}\,. \tag{3.19}$$

For a central transit there is $(\Delta t_{\rm full}/\Delta t_{\rm tot}) = (R_S - R_P)/(R_S + R_P)$ and b becomes zero as it should be.

For small planet radius $R_P \ll R_S$ one can simplify this further and it results for the impact parameter

$$b^2 \approx R_S^2\Big(\frac{R_S - R_P}{R_S}\,\frac{\Delta t_{\rm tot}}{\Delta t_{\rm full}}\Big)$$

Summary: With the measurement of the different eclipse times and the radius ratio one can determine the impact parameter b and derive an accurate estimate on the orbit inclination.

For the simplified treatment given above the following approximations are used, which are for many cases very reasonable:

– The relative trajectory of the transiting or eclipsed object is treated as a straight line what is a good approximation for a large separation $d \gg R_S$.

– The relative transverse motion is considered to be constant during the entire transit, what is reasonable for a large separation $d \gg R_S$.

– For many cases it will be possible to use circular orbits as good approximation. If eccentricity is non-negligible then one can correct the timing-formula with the following correction factors:

$$f_{\rm corr} \approx \frac{1 - \epsilon^2}{1 \pm \epsilon \sin\omega}\,, \tag{3.20}$$

where $\pm$ depends on whether this factor is used for the the transit or eclipse times or whether it is used for the determination of the impact parameter b.

3.4.3 Transit and eclipse light curve

The combined flux of the star and the planet $F(t)$ can be described for the different phases as the sum from the star and the planet $F(t) = F_S(t) + F_P(t)$. These fluxes can vary due to the following reasons:

- $F_S(t)$ and $F_P(t)$ are the intrinsic flux of the star and the planet which can vary with time,
- in addition parts of the star are occulted during the transit phase which can be described by a transit attenuation function $\delta_t(t)$. This yields then the observed stellar flux
$$F'_S(t) = F_S(t)(1 - \delta_t(t)) \,.$$
- Similarly we can describe the observed flux from the planet which is the intrinsic flux which may be attenuated $\delta_e(t)$ during secondary eclipse by the star
$$F'_P(t) = F_P(t)(1 - \delta_e(t)) \,.$$

We are not interested in the intrinsic variability of the star and normalize the total flux to the intrinsic stellar flux:

$$f(t) = \frac{F'_S(t) + F'_P(t)}{F_S(t)} = 1 - \delta_t(t) + \frac{F_P(t)}{F_S(t)}(1 - \delta_e(t)) \,. \tag{3.21}$$

The individual terms can be characterized as follows:

- the attenuation by the transiting planet $\delta_t(t)$ is in Eq. 3.21 a large term of the order 1 % for a giant planet transiting a solar-type star. In many cases only this "transit" term must be considered. Outside a transit there is $\delta_t(t) = 0$.
- the intrinsic flux of the planet normalized to the stellar flux $F_P(t)/F_S(t)$ is usually a small term < 0.1 % and hard to measure. In this ratio the time dependence of the stellar brightness can be neglected $F_S(t) \approx \bar{F_S}$ because the relative phase effects in the planet flux are expected to be much larger. However, strong intrinsic variability of the star on short timescales can introduce problems with the normalization and introduces significant uncertainties in the leading "1-term" which can be larger than $F_P(t)/\bar{F_S}$.
- the attenuation of the planet flux during secondary eclipse is $\delta_e(t) = 0$ out of secondary eclipse and $\delta_e(t) = 1$ during the full eclipse phase. Only during ingress and egress a time dependence is expected which can be approximated with a linear interpolation or another simple fit.

3.4.4 Limb darkening

It is well known from the sun that the solar disk has no uniform surface brightness. The sun is brighter in the center and the surface brightness drops towards the limb. This effect is universal for stars. This implies for the attenuation during a planetary transit:

- $\delta_t > R_P^2/R_S^2$ when the planet is in front of the center of the star,
- $\delta_t < R_P^2/R_S^2$ when the planet is in front of the outer regions of the star.

Thus the transit light curve is not just a simple trapezium defined by the transit times t_i to t_{iv} and a constant transit depth δ_t, but a rounded transit curve depending on the exact transit trajectory and the surface brightness distribution of the star. This level of detail must be taken into account for an accurate derivation of radius ratio R_P/R_S.
Different limb darkening laws are used for the description of stars. A frequently used function has the following radial dependence:

$$\frac{I(r)}{I_0} = 1 - a(1 - \mu(r)) - b(1 - \mu(r))^2 \,. \tag{3.22}$$

The used term $\mu(r) = \cos\theta = 1 - \sqrt{1 - (r/R_s)^2}$ in this quadratic fit equation describes the emissivity of the atmosphere as function of the angle θ between surface normal and direction of the emission. For the center of the disk $r = 0$, there is $\mu = \cos\theta = 1$ and $I(r = 0) = I_0$. For the limb $r = R_S$ there is $\mu = 0$ and $I(R_S)/I_o = 1 - a - b$. The emitted intensity $I(\mu)$ is a typical result of radiative transfer calculation for stellar atmospheres. The fit formula given above is only one of many different fit models used in the literature.

Some general properties of stellar limb darkening are:

- the limb darkening law depends on the spectral type of the star,
- for a given star the limb darkening is a function of wavelengths, for solar type stars it is stronger for short wavelengths,
- the transit light curves provide currently the best measurements for the limb darkening of stars.

Slide 3.11 shows the wavelength dependence of the limb darkening as function of wavelength measured for the transit of HD 209458 b with the spectroscopic mode of HST.

Small scale stellar surface structure. The transit depth $\delta_t(t)$ is a measure of the intensity of the hidden region on the stellar photosphere. Small scale inhomogeneities on the stellar structure can therefore produce bumps in the transit light curve if they are hidden by the planet on its trajectory in front of the star.

- the transit intensity is higher (the attenuation less deep) if the planet is in front of a dark spot,
- the transit intensity is lower (more attenuation) if the planet is in front of a bright region.

Different examples for surface structure were found in well observed transit light curves, like for the active star HD 189733 in Slide 3.12. For extra-solar planet research it is important to be aware of this effect so that it can be taken into account in the transit light curve analysis.

3.5 Observational data for transiting planets

First transit observations. The first successful transit observations were made in 1999 for HD 209458 b with a small telescope and a simple camera system by a group led by David Charbonneau (see slide 3.09). They had a systematic program of follow-up photometry of all close-in giant planets which were detected by the RV-method. The detected transit depth for this object is about $\delta_t \approx 1.5$ %. The fact, that planetary transits can be detected with simple equipment demonstrated that this technique has a huge potential for extra-solar planet research.

3.5.1 Requirements for transit photometry.

Transit photometry requires observations with a high photometric accuracy during at least the duration of the transit which lasts for close-in planets typically a few hours. For different science goals the following rough photometric sensitivities $\Delta I/I$ must be achieved (3σ detection):

- $\Delta I/I \approx 1$ % during several hours for the detection of transits of giant planets around solar type stars,
- $\Delta I/I \approx 10^{-3}$ during several hours for the detection of the secondary eclipse of hot Jupiters in the mid-IR wavelength range $\lambda > 3\,\mu$m; if this is precision can be obtained for about a full period then one can also measure the phase curve of the planet,
- $\Delta I/I \approx 10^{-4}$ during several hours for the detection of the transits of terrestrial planets around solar type stars,
- $\Delta I/I \approx 10^{-4}$ during several hours in narrow band or spectroscopic mode for the measurement of the spectral dependence of the transit depths of giant planets.
- $\Delta I/I \approx 10^{-6}$ for the measurement of the transit spectrum of an Earth like planet around a solar-type star.

Ground based observations are limited by the atmospheric turbulence which produces photometric fluctuation at a level of about $\Delta I/I = 0.2 - 0.5$ %. This limit depends not on telescope size and therefore one can measure the 1 %-transits of giant planets around bright stars ≈ 10 mag with small 10 cm telescopes, and around faint stars ≈ 20 mag with large 8 m telescopes.

Space observations have the huge advantage that there is no disturbing atmosphere. The achievable measuring precision is then limited by instrumental effects and the photon noise. Systematic effects can easily reach a level of more than 0.1 % if the space telescope is not designed for high precision photometry or if there are unforeseen disturbing effects. On the other side it has been demonstrated with HST and the KEPLER satellite that a precision of $\Delta I/I = 10^{-4}$ can be achieved, and new instruments with a higher precision are currently built or planned.

Photon noise is a severe limit for small space instruments, because it is so expensive to build large telescope in space. The photon noise is given according to the Poisson statistics by

$$\sigma = \sqrt{N}$$

where N is the number of collected photons for the transit measurement. This means that a 3σ measuring precision of $\Delta I/I = 10^{-4}$ requires about $N \approx 10^9$ photons per hour. This

can be achieved for about a 13 mag star with the KEPLER space telescope collecting all photons from 400 - 900 nm.

3.5.2 Results from the transit search programs from the ground.

There are two major types of transit searches form the ground, the RV follow-up and the wide field transit search.

RV follow-up: The follow-up search for transits of planets detected by the RV-method is very useful. The chance of making a successful detection is quite high because the presence of a planet and its approximate mass and therefore size is known. Further the transit search can be well scheduled because the epoch of a possible transit follows also from the RV curve.

Well known transiting systems which were detected by follow-up observations are:

- HD 209458 b, the first transiting extra-solar planet discovered. It is a hot Jupiter on a 3.5 day period around a bright ($m_V = 7.5$ mag) G0 V star. This is one of the brightest, and best studied transiting systems (Slides 3.9, 3.11).
- HD 189733 b, a hot Jupiter on a 2.2 day period around a nearby (19.5 pc), bright ($m_V = 7.7$ mag) and quite active K2 V star. This is one of the brightest and best studied transiting systems.
- HD 80606 b is a $M_P = 4M_J$ hot Jupiter planet on an $P = 111$ days orbit with the extreme eccentricity of $\epsilon = 0.93$ around a G5 V star. The irradiation for this planet changes by a factor of 800 from apostron to periastron.

Wide field search for close-in planets. With small wide field cameras stars of magnitude 8 to 12 mag can be monitored for periodic transits by giant planets. The HAT (Slide 3.13) and WASP search programs found up to now more than 100 transiting planets. Most of the found objects are giant planets on short orbits, the majority with $P = 2$ to 5 days. The found planets tend to have large radii, between 1.1 to 1.3 R_J. Such large radii are not expected for intrinsically cold (=old) giant planets. The large radii are suspected to be induced by the strong irradiation and possibly also by tidal effect. Large, giant planets on short orbits are of course the most easy objects to detect and therefore this sample is strongly biased in this direction.

The transit search surveys confirms strongly the predominance of orbital periods of $3 - 5$ days for giant planets. This must be a particularly stable configuration.

The search for transits around M-stars is a very interesting variant for ground-based planet searches. Because the radii of M-stars are only 0.1 to 0.4 $R_\odot$ transits of smaller planets produce a transit signal of more than 0.1 %. A well known example for a successful transit detection of GJ 1214 b, a Neptune-sized planet with $R_P = 2.7R_E$ on a 1.6 day orbit. Because the star is a M4.5 V star with a mass of 0.15 $M_\odot$, and only a radius of 0.21 $R_\odot$ also a Neptune sized planets is detectable from the ground. The mean density of this planet is 1.9 g/cm^3 again similar to the ice giants in the solar system. The equilibrium temperature is about 450 K.

3.5.3 Result from the follow-up observation with space telescopes

Space telescopes are the ideal follow-up instruments for transit and secondary eclipse observations. Because these instrument are in space a higher precision than with ground based observations is achievable. Follow-up observations provide successful measurements because the best observing epoch can be selected and the very valuable observing time can be invested optimally. Most important follow-up space instruments for transit measurements are HST and SPITZER. Some results from these observations will be discussed in the section on hot jupiters.

HST follow-up. A few examples of HST transit observations are shown in Slides 3.9, 3.11, 3.12. Key characteristics of HST transit observations are:

- HST achieves the highest quality transit data with low resolution spectroscopy in the visual. This mode provides transit light curves for the 300 - 900 nm range which an accuracy up to 10^{-4}.
- HST is with 2.5 m diameter a large space telescope which can also observe planetary transits for fainter stars.
- HST is a versatile observatory with many different observing modes from the far-UV to the near-IR.
- A disadvantage of HST is the high pressure factor for observing time. Therefore one needs to have a very strong science case to obtain observing time.
- Because HST orbits Earth in about 2 hours the transit observations are always interrupted. During each orbit the spacecraft goes from the night side to the day side and after an integration of about 1 hour there is always a gap of one hour. The day - night changes introduce also strong thermal effects which cause difficulties for accurate photometric measurements.

SPITZER follow up . Two typical results from SPITZER secondary eclipse and phase curve observations will be discussed in chapter on hot jupiters. Key characteristics of the SPITZER spacecraft are:

- The SPITZER spacecraft is an 85 cm telescope which provides photometric measurements in the mid-IR, at wavelength of about 3.6 μm, and 4.5 μm, 5.8 μm, and 8 μm and longer wavelength. After the consumption of the cryogen the two short wave channels 3.6 μm, and 4.5 μm can still be used for transit studies.
- the expected planet signal is large in the SPITZER range and therefore relatively easy to detect,
- SPITZER is located at the Lagrange point 4 what allows uninterrupted observations during several hours.
- a big disadvantage of SPITZER is, that it was not designed for high precision photometry. Despite this it reaches a rather good sensitivity but requires many instrumental corrections which induce quite large uncertainties.

The most important results are the measurements of the thermal emission from close-in planets via secondary eclipse observations and even phase curve measurements. This will be discussed in the Section on hot jupiters.

3.5.4 The KEPLER satellite mission

The KEPLER satellite produced with a systematic transit search a scientific revolution for the study of statistical properties of extra-solar planets. It found more than 2000 planet candidates most of which will be confirmed in the coming years. Thanks to this mission we have now a good understanding on the statistics of planets with periods of less than about 1 year.
The key characteristics of the KEPLER mission are:

- KEPLER is a modified Schmidt telescope with a entrance corrector lens of 0.95 m aperture and a 1.4 m diameter f/1 primary mirror,
- the instrument has a large field of view of $10.5° \times 10.5°$ or 115 square degrees,
- the satellite is located on an Earth trailing orbit and can therefore point continuously at the same sky region in the constellations Cygnus and Lyra (see Slide 3.15),
- the detector system monitors the brightness of about 150'000 main sequence stars in the brightness range 11 – 15 mag,
- the system is sensitive enough to detect a single transit of an Earth-sized planet ($\Delta I/I \approx 10^{-4}$) in front of a 12 mag G2 V star.
- KEPLER is working now without major interruptions since May 2009 or more than 1400 days.

Confirmation of planet candidates. The main problem of the KEPLER mission is the verification of transit candidates. There are many variable stars in the field which produce photometric signals which look like transits. Several test must be applied to verify the presence of a real planet transit:

- the transit signal must be periodic,
- the light curve should look like a transit,
- the photo-center should remain stable between transit and out-of transit phases, to exclude the presence of blended background or foreground target, e.g. an eclipsing binary, which may introduce a transit like disturbance (so called false positive),
- a mass determination using the RV-method or transit timing variation measurements can confirm the presence of a planetary mass object.

Basic detection statistics. The following numbers were derived based on 16 months of Kepler measurements (Batalha et al. 2013, ApJS 204, 24):

- about 190'000 stellar light-curves were measured, about 130'000 during the entire period,
- about 5000 stars show periodic, transit like light curves,
- about 2300 viable planet transit candidates were identified,
- about 400 systems show transits from multiple planets, giving about 900 detected planet candidates in multiple systems,
- about 200 planets are firmly confirmed with mass determinations and many more will follow in the coming years,
- the richest system found so far is Kepler-11 with 6 transiting planets (Slides 3.17 and 3.18).

3.5.5 Main results from the KEPLER mission

The Kepler mission is not yet completed and only preliminary results are available. However, they give already a very good impression about statistical properties of extra-solar planets with $R_P > 2\,\mathrm{R}_E$ and $P < 50$ days (see Howard et al. 2012, ApJS 201, 15).

Frequency of extra-solar planets. The frequency of extra-solar planet can be estimated taking into account the transit probability p_{trans} which is a function of stellar radius and orbital separation. A study selecting planets with $P < 50$ days and $R_P > 2\ \mathrm{R}_E$ (the sample of easy to detect short period, "large" planets) obtained the following planet to star ratios N_P/N_S (see Slide 3.19 and 3.20):

- there are about $N_P/N_S = 0.18$ planets per star with $R_p > 2\mathrm{R}_E$ and $P < 50$ days.
- and about $N_P/N_S = 0.01$ giant planet per star with $P < 50$ days,
- planets with small radii are much more frequent indicating that Earth-sized planets $R_P \approx 1\,\mathrm{R}_E$ and a period $P < 50$ days could be present around every third or second star or even more frequently.

Note, that our solar system with its 8 planets is a system which does not qualify to be counted in this sample ($P < 50$ days). The RV-results from the HARPS instrument are in agreement with these statistics from KEPLER.

Period distribution. The period distribution of planets follows from the transit detection rates, taking the strong period dependence of the transit probability into account. For long periods the number of detections gets small and the correction factors large. Therefore, the KEPLER transit survey provides good results for planets on orbits with short periods but not for planet with orbits > 1 year. The main findings are (see Slide 3.20):

- planets with very short periods $P < 2$ days are very rare, less than 0.1 % of all stars have such a planet,
- hot Jupiters ($R_P > 8\,\mathrm{R}_E$) with $P < 10$ days occur around 0.5 % of all stars, and the same number applies also for the period range $P = 10 - 50$ days.
- There is, like in the RV-data, a pile up of Jupiters around $P = 3 - 5$ days and a clear minimum for $P = 5 - 10$ days.
- planets with radii intermediate between Neptune and Jupiter $R_P = 4 - 8\,\mathrm{R}_E$ have essentially the same frequency and period dependence like the giant planets, but there lacks the clear period minimum at $P = 5 - 10$ days in the distribution.
- small planets $R_P = 2 - 4\,\mathrm{R}_E$ have a similar frequency for very short periods $P = 2 - 4$ days like giant planets, but they are 5 to 10 times more frequent than giant planets for $P > 10$ days.

Frequency of planets for different stellar types. The KEPLER sample is large enough to investigate the planet occurrence as function of stellar type. In Slide 3.21 the results are plotted for planets with $R_P > 2\,\mathrm{R}_E$ and $P < 50$ days. The main points are:

- for small planets, $R_P = 2 - 4\,\mathrm{R}_E$, there is a very strong dependence of the planet occurrence with stellar type with a planet to star ratio $N_P/N_S \approx 0.25$ for K and early M dwarfs (1 Neptune-sized planet with $P < 50$ days per 4 stars)

- for G-stars (the majority of the objects in the KEPLER sample) the derived planet to star ratios for small planets $R_P = 2 - 4\,\mathrm{R}_E$ is about $N_P/N_S \approx 0.15$, and for F stars about $N_P/N_S \approx 0.10$.
- larger planets $R_P > 4\,\mathrm{R}_E$ show no preference for stellar types, they seem to occur with equal ratios $N_P/N_S \approx 0.02$ around F, K, G, and M dwarfs.

Multiple planets. Thanks to the many multi-planet detections by KEPLER one can investigate the properties of multiple systems. First of all, systems with multiple transiting planets are frequent among the 1405 stars with transits (Fabricki et al. 2013, ApJ 768, 14):

- for 1044 stars (74 %) only one transiting planet was detected,
- for 242 stars (17 %) transits of two planets were detected,
- 85 stars (6 %) show transits of three planets,
- 25 stars (1.8 %) show four planets with transits,
- 8 stars (0.6 %) show five planets,
- 1 star shows six planets (Kepler-11).

These are of course only lower limits since the analysis of more data and more studies will reveal more transiting planets. Form these statistics one can infer:

- systems with multiple transiting systems are frequent, indicating first that multiple planetary systems are frequent and second that the orbits are often close to coplanar,
- for close-in Jupiters there are essentially no additional second planet with transits found indicating that there is rarely a second co-planar planet in these systems (see Slide 3.22),
- in systems with multiple transiting planets the innermost planet tends to be a small planet (Slide 3.22),
- period ratios between planets are larger than $P_{\mathrm{out}}/P_{\mathrm{in}} > 1.25$ defining an observational limit for the "planet packing" density (Slide 3.23).
- planets are typically not in orbital resonance, but they prefer period ratios which are just a bit larger than 3:2 or 2:1 and avoid ratios just below this value (Slide 3.23).

3.6 The empirical mass-radius relation for planets

Transiting planets with radius determinations and mass determination can be used for empirical mass radius $M_P - R_P$ and radius-density $R_P - \bar{\rho}$ diagrams. Such diagrams are shown in Slide 3.24. Ground-based data provide mainly a sample of giant planets with short periods while KEPLER provides data for smaller planets.

For the giant planets the masses are determined with the RV method. Masses for some KEPLER planets are derived based on transit timing variations (see Section 3.6. Errors in the mass and radius determinations are small for the giant planets while the errors for the masses for low mass planets are quite large. Additional, but smaller error sources are the uncertainties in the mass and radius estimates for the host stars. The sample was devided into lower mass planets $M_P < 150\,\mathrm{M_E} = 0.5\,\mathrm{M}_J$ and higher mass planets $M_P > 0.5\,\mathrm{M}_J$. There is also a systematic difference between planets with higher and lower incident flux, where the border line is just the median irradiation value. This median value corresponds to an irradiation per unit area which is 800 times stronger than for Earth.

The following properties for the mass M_P, radius R_P and mean density $\bar{\rho}$ can be derived from the diagrams in Slide 3.24:

- giant planets with masses $> 100\,\mathrm{M}_E$ have without exception radii in the range $R_P = 10 - 20\,\mathrm{R}_E = 1 - 2\,\mathrm{R}_J$,
- the strongly irradiated giant planet have clearly a larger radius,
- lower mass planets $M_P < 150\,\mathrm{M}_E$ show a strong radius-mass dependence which can be described roughly by the relation

$$\frac{R_P}{R_E} \approx \Big(\frac{M_P}{M_E}\Big)^{0.5},$$

- the scatter around this curve is large since there are $10\,\mathrm{M}_E$ planets with small radii $R_P \approx 1.5\,\mathrm{R}_E$ and such with large radii $R_P \approx 6\,\mathrm{R}_E$,
- strongly irradiated Neptune-mass planets tend to be small,
- the density for giant planets increases from about $\bar{\rho} = 0.3\,\mathrm{g/cm}^3$ at $M_P \approx 0.3\,\mathrm{M}_J$ to $\bar{\rho} = 10\,\mathrm{g/cm}^3$ for $M_P \approx 10\,\mathrm{M}_J$,
- for low mass planets there is a huge spread in density which can be as low as $\bar{\rho} \approx 0.4\,\mathrm{g/cm}^3$ or as high as $\bar{\rho} = 10\,\mathrm{g/cm}^3$ for planets with $M_P = 10\,\mathrm{M}_E$
- clearly, the strongly irradiated low mass planet are the ones with very high densities.

Interpretation. The available data can be interpreted as follows. Giant planets become not larger with higher mass because their matter is simply more compressed if the pressure increases. The enhanced radii due to strong irradiation is an important process for close-in planets. Lower mass planets show a strong variety of densities. This might point to the fact that irradiation and evaporation has a strong impact on these planets. Low mass planets with mean densities of $\bar{\rho} \approx 10\,\mathrm{g/cm}^{-3}$ could be large rocky / iron cores which have lost their lower density H envelope. Of course it would be interesting to compare the currently available data with planets at larger distances, which are only slightly affected by irradiation. This may be possible in the future.

Chapter 4

Radiation from planets

We consider first basic, mostly photometric radiation parameters for solar system planets which can be easily compared with existing or future observations of extra-solar planets. In the next section we consider in more detail the physics of planetary atmospheres which is important for the interpretation of the thermal or reflected spectral radiation from planets.

4.1 Equilibrium temperature

The equilibrium temperature $T_{\rm eq}$ of a planet is a theoretical parameter which assumes that the irradiated flux $F_{\rm in}$ from the star is equal to the thermal back-body emission luminosity of the planet $L_{\rm out}$. The following assumptions are made for the derivation of $T_{\rm eq}$:

- the irradiated radiation is either reflected or absorbed,
- the absorbed radiation energy is re-emitted as thermal radiation,
- there is no internal energy source,
- the planet is isothermal (same temperature on the day and night side!).

The irradiated flux is:

$$F_{\rm in} = \frac{L_\odot}{4\pi\, d_P^2}\, \pi R_P^2 (1 - A_B)\,, \tag{4.1}$$

where $L_\odot$ is the luminosity of the sun, d_P the separation and R_P the radius of the planet, and A_B is the Bond albedo. A_B is the fraction of the total irradiated energy which is reflected and which does not contribute to the heating of the planet. A Bond albedo $A_B = 1$ means that all light is reflected, while $A_B = 0$ indicates a perfectly absorbing (black) planet. Both cases are not realistic. Expected values for the Bond albedo are in the range $A_B = 0.05$ to 0.95.

The luminosity of the planet, which is assumed to radiate like a black body, is

$$L_{\rm out} = L_P = 4\pi\, R_p^2\, \sigma T_{\rm eq}^4\,, \tag{4.2}$$

where σ is the Stefan-Boltzmann constant and $T_{\rm eq}$ the equilibrium temperature of the planet.

The equilibrium temperature $T_{\rm eq}$ follows from $F_{\rm in} = L_{\rm out}$:

$$T_{\rm eq} = \left(\frac{L_\odot (1 - A_B)}{16\,\pi\,\sigma}\right)^{1/4} \frac{1}{\sqrt{d_P}} \tag{4.3}$$

This indicates that $T_{\rm eq}$ decreases with distance from the sun for solar system objects or from the star for extra-solar planets. An important feature of this equations is, that it does not depend on the radius of the irradiated body which can be as small as a dust particle (mm-sized) or as large as a giant planet.

Temperatures for solar system planets. The equilibrium temperatures $T_{\rm eq}$ for the solar system planets is given in Table 4.1 using the indicated Bond albedos A_B and the planet separation $d_P = a$ from Table 2.1. The Table compares $T_{\rm eq}$ also with the measured ground temperature $T_{\rm ground}$ for terrestrial planets and the effective temperatures of the emitted thermal radiation $T_{\rm eff}$. $T_{\rm eff}$ is for Jupiter, Saturn and Neptune higher than the equilibrium temperature, because these planets have a substantial intrinsic energy source.

Mercury is a special case because this planet has no atmosphere and only a slow rotation. For this reason there are very large temperature differences between the irradiated (725 K) and the non-irradiated (100 K) hemisphere. For Mercury the assumption of an isothermal planet is not appropriate. However, averaged over all direction the effective temperature of the emitted thermal radiation agrees quite well with the equilibrium temperature.

Table 4.1: Radiation parameters for solar system planets: A_B is the Bond albedo, $T_{\rm eq}$, $T_{\rm ground}$, $T_{\rm eff}$ the equilibrium, ground and effective temperature, and $L_P/F_{\rm in}$ the flux ratio between thermal emission and irradiation, $L_p/L_\odot$ the luminosity contrast, and $F_p/F_\odot$(IR) the flux contrast at long wavelengths $\lambda \gg \lambda_{\rm max}$.

Planet	A_B	$T_{\rm eq}$	$T_{\rm ground}$	$T_{\rm eff}$	$L_{\rm th}/F_{\rm in}$	$L_p/L_\odot$ 10^{-10}	$\lambda_{\rm max}$	$F_p/F_\odot$ 10^{-6}
Mercury	0.12	448 K	725/100[1] K	448 K	1	4.4	6.5 μm	0.95
Venus	0.75	328 K	730 K	328 K	1	7.7	8.8 μm	4.3
Earth	0.31	279 K	290 K	279 K	1	4.5	10.4 μm	4.0
Mars	0.25	227 K	225 K	227 K	1	0.56	12.8 μm	0.93
Jupiter	0.34	110 K	–	124 K	1.6	21.	23.4 μm	220.
Saturn	0.34	81 K	–	95 K	1.9	5.0	30.5 μm	110.
Uranus	0.30	59 K	–	59 K	1	0.14	49.2 μm	13.
Neptune	0.29	47 K	–	59 K	2.5	0.14	49.2 μm	13.

1: 725 K is for the irradiated hemisphere and 100 K for the "night" hemisphere. For the sun the adopted temperature is $T_{\rm eff} = 5800$ K.

Greenhouse effect for terrestrial planets. For the planets Earth and Venus the ground temperature T_{ground} is a significantly higher than T_{eq} due to the greenhouse effect.

Figure 4.1: Energy flow diagram for the greenhouse effect on Earth.

In the greenhouse effect (e.g. for Earth) the visual light from the sun penetrates through the atmosphere down to the surface and heats efficiently the ground. However, the thermal IR-radiation from the ground can only escape in certain spectral windows without strong molecular absorptions (H_2O, CO_2), while the rest is absorbed in the atmosphere (see Slide 4.1). Energy transport from the warm/hot ground to higher cold layers occurs therefore through convection and radiation until the thermal radiation can escape to space. T_{eq} represents the temperature of the atmospheric layers from which the thermal radiation can escape. Therefore, the ground temperature is higher than T_{eq}. The effect is stronger on Venus because of its much thicker atmosphere (90 bar) when compared to Earth (1 bar).

4.2 Thermal radiation from planets

Intrinsic energy for the giant planets. Table 4.1 gives the ratio between flux irradiation $F_{\rm in}$ and the total thermal emission L_P which can be deduced from the equilibrium and effective temperatures according to

$$\frac{L_P}{F_{\rm in}} = \Big(\frac{T_{\rm eff}}{T_{\rm eq}}\Big)^4 .$$

A ratio > 1 for Jupiter, Saturn, and Neptune indicates that these planets emit significantly more energy than they receive from the sun. This can be explained by the ongoing contraction, and differentiation, of these three planets. For Uranus, it is expected that there is also a small intrinsic flux but only at a level of about $5-10$ % of the irradiated flux. This effect is hard to measure due to uncertainties in the effective temperature determination. The presence of the internal energy source indicates that the central temperature of the giant planets is of the order $\approx 10'000$ K. Intrinsic energy sources can be neglected for the terrestrial planets in the solar system.

Black body radiation. The spectral intensity of the thermal radiation of an object at temperature T can be described by the Planck or the black body intensity spectrum:

$$B(T,\lambda) = \frac{2hc^2}{\lambda^5}\frac{1}{e^{hc/\lambda kT}-1}\,, \tag{4.4}$$

where h, k and c are Planck constant, Boltzmann constant and speed of light. The Planck intensity is given in unit of e.g. [J m^{-2} sr^{-1} s^{-1} μm^{-1}] or [erg cm^{-2} sr^{-1} s^{-1} Å^{-1}]). Black body radiation is isotropic so that the black body flux through a unit surface area is $\pi B(T,\lambda)$. It is assumed that the properties of the black body radiation are known and we remind here only some important facts:

- the black body spectrum can also be expressed as function of frequency

$$B(T,\nu) = \frac{2h\nu^3}{c^2}\frac{1}{e^{h\nu/kT}-1}\,,$$

- conversion between $B(T,\nu)$ and $B(T,\lambda)$ must use the factor $d\nu = -c/\lambda^2 d\lambda$,
- the peak of the black body spectrum $B_{\rm max}(T,\lambda)$ is according to the Wien law at the wavelength:

$$\lambda_{\rm max} = \frac{2.9{\rm mm}}{T[{\rm K}]}\,, \tag{4.5}$$

  which is at 10 μm for a planet with $T = 290$ K ($\approx$ Earth),
- for low frequency or long wavelengths the Planck radiation can be approximated by the Rayleigh-Jeans law:

$$B(T,\nu) = \frac{2\nu^2}{c^2}kT \quad \text{or} \quad B(T,\lambda) = \frac{2c}{\lambda^4}kT\,, \tag{4.6}$$

- the total luminosity of the spherical black body (planet) with radius R and effective temperature $T_{\rm eff}$ is

$$L_P = 4\pi R_P^2\,\sigma T_{\rm eff}^4\,, \tag{4.7}$$

  where σ is the Stefan-Boltzmann constant (identical to Equation 4.2 except that $T_{\rm eff}$ is used instead of $T_{\rm eq}$ which does not account for intrinsic energy sources).

Thermal luminosity and flux contrast between planet and sun. The thermal luminosity L_P of an irradiated planet without intrinsic energy source is given by Equations 4.1 or 4.2. This can be expressed as thermal luminosity contrast $C_{\rm th}$ between the planet and the sun

$$\frac{L_P}{L_\odot} = \frac{R_P^2}{R_\odot^2}\frac{T_{\rm eq}^4}{T_\odot^4} = \frac{R_P^2}{d_P^2}\,\frac{1}{4}(1-A_B)\,. \tag{4.8}$$

For solar system planets this ratio is very small, of the order 10^{-9} to 10^{-11} (see Table 4.1).

Equation 4.8 for the luminosity contrast is also valid for extra-solar systems. For hot Jupiters the ratio $L_P/L_\odot$ is much larger than for solar system planets.

The flux contrast as function of wavelength is important for observational studies. For long wavelengths, in the Rayleigh-Jeans part of the Planck function of the planet, one can use equation 4.6 which yields:

$$\frac{F_P(\lambda \gg \lambda_{\rm max})}{F_\odot(\lambda)} = \frac{R_P^2}{R_\odot^2}\,\frac{T_{\rm eq}}{T_\odot}\,. \tag{4.9}$$

The factor for the temperature ratio between planet and sun (or star) $T_{\rm eq}/T_\odot$ is of the order $\approx 10-100$. Thus the flux contrast at long wavelengths $\lambda \gg \lambda_{\rm max}$ is several orders of magnitudes ($10^3 - 10^6$) larger than the total luminosity contrast (see Table 4.1). On the other hand, the planet to star flux contrast at short wavelengths $\lambda < \lambda_{\rm max}$ decreases rapidly to very small values because the thermal radiation of the planet drops-off exponentially. At short wavelengths the scattered light will therefore dominate.

4.3 Reflection from planets

Reflection by a Lambert surface. A Lambert surface is used as reference in many technical and scientific studies on reflectivities. A Lambert surface reflects all incident light and the surface brightness is the same for all viewing angles. However, for viewing directions with an angle θ with respect to the surface normal the apparent reflecting area and therefore also the reflected flux is reduced $\propto \cos\theta$. Thus the reflected intensity $I_{\rm Lam}$ of a flat Lambert surface per unit solid angle is

$$I_{\rm Lam}(\theta) = F_i \frac{\cos\theta}{\pi} \quad \text{for} \quad 0^\circ \le \theta < 90^\circ \tag{4.10}$$

where F_i is the incident flux onto the considered surface. Thereby, the reflection from a Lambert surface does not depend on the direction of the irradiation. A sheet of white paper, a with screen or a white wall are close to a Lambert surface.

Figure 4.2: Reflection from a Lambert surface.

The factor $1/\pi$ in Equation 4.10 is the normalization factor because energy conservation requires that the reflected intensity $I_{\rm Lam}$ integrated over all direction is equal to F_i

$$\int_0^{2\pi}\int_{-\pi/2}^{\pi/2} I_{\rm Lam}(\theta)\sin\theta\, d\theta\, d\phi = \int_0^{2\pi}\int_0^{\pi/2} F_i\frac{\cos\theta}{\pi}\sin\theta d\theta d\phi = F_i\,.$$

An observer at a distance D (much larger than the linear dimension of the surface area) measures a reflected flux $F_{\rm Lam}$ per unit area of

$$F_{\rm Lam}(\theta) = \frac{I_{\rm Lam}(\theta)}{D^2} = \frac{F_i\cos\theta}{\pi D^2}$$

Normal retro-reflection of a Lambert disk irradiated by the sun. Based on the reflection law for a Lambert surface we can derive the normal retro-reflection (= normal irradiation and normal reflection) of solar light by a round Lambert disk with radius $R_{\rm disk}$ at the distance $d_{\rm disk}$ from the sun:

$$F_{\rm disk}(\lambda,\theta=0) = \frac{F_i(\lambda)}{\pi D^2} = \frac{L_\odot(\lambda)}{4\pi d_{\rm disk}^2}\pi R_{\rm disk}^2\,\frac{1}{\pi D^2}\,,$$

where F_i is replaced by the explicit formula for the sunlight intercepted by the Lambert disk.

Geometric albedo for solar system planets. The geometric albedo $A_g(\lambda)$ is the spectral reflectivity of a planet at zero phase angle $\alpha = 0$ (full phase) relative to the reflectivity of a Lambert disk with the same cross section as the planet

$$A_g(\lambda) = \frac{F_P(\lambda)}{F_{\rm disk}(\lambda)}\,. \tag{4.11}$$

Thus, the geometric albedo of a planet can be determined by measuring the magnitude of that planet at opposition (normal retro-reflection), which is then compared to the calculated reflection of a Lambert disk with the same cross section.
It is convenient to express the theoretically reflected flux from a Lambert disk relative to the flux of the sun measured from Earth $F_\odot(\lambda) = L_\odot(\lambda)/4\pi d_{\rm E}^2$ (where $d_{\rm E} = 1$ AU), because this ratio is independent of wavelength:

$$R = \frac{F_{\rm disk}}{F_\odot} = \frac{d_{\rm E}^2}{D^2}\,\frac{R_{\rm disk}^2}{d_{\rm disk}^2}\,. \tag{4.12}$$

Opposition $\alpha = 0^\circ$ occurs for the outer planets almost every year. Because the phase angles for the giant planets is never really large, $\alpha \lesssim 12^\circ$ for Jupiter, $\lesssim 5^\circ$, and less for Uranus and Neptune, one can correct for the small deviations for an "ideal" geometric albedo measurement.

Example Jupiter: As example we calculate with Equation 4.12 the case for Jupiter for which the distance to Earth at opposition is $D = 5.2 - 1$ AU, $R_{\rm disk} = R_J = 69910$ km and $d_{\rm disk} = d_J = 5.2$ AU with $1\,{\rm AU} = 1.510^8\,{\rm km}$. The ratio between the flux of a Lambert disk with a cross section equivalent to Jupiter and the solar flux is

$$R = \frac{F_{\rm disk}}{F_\odot} = 4.55 \cdot 10^{-10} \quad \text{or} \quad m_{\rm disk} - m_\odot = -2.5 \log R = 23.36\,,$$

where the result is also given as magnitude difference. The apparent V-band magnitude for the sun is $m_\odot({\rm V}) = -26.74$ mag and for Jupiter at opposition about $m_J({\rm V}) = -2.70$ mag. This yields an opposition contrast of $m_J - m_\odot = 24.04$ mag or about $\Delta m = 0.7$ mag more than expected for a Lambertian disk. The geometric albedo of Jupiter is this magnitude difference Δm expressed as ratio $A_g = 10^{-0.4\cdot\Delta m} = 0.52$ in good agreement with available literature values.

Figure 4.3: Typical constellation for the geometric albedo measurement of Jupiter or another outer planet.

Table 4.2: Reflection properties of solar system planets: geometric albedos for the V-band and the IR, phase integral q and calculated spherical albedos. The last columns give the factor R_p^2/d_p^2 and the flux contrast for the scattered light at quadrature phase assuming $f(90^\circ) = 0.3$.

planet	$A_g(V)$	A_g(IR)	q	$A_s(V)$	$A_s(IR)$	A_B	R_p^2/d_p^2 10^{-10}	$F_P/F_\odot$ 10^{-10}
Mercury	0.142		0.48	0.07		0.12	18.	0.77
Venus	0.67					0.75	31.	6.2
Earth	0.367					0.31	18.	2.0
Mars	0.170					0.25	2.2	0.11
Jupiter	0.52	0.27	1.25	0.65	0.34	0.34	77.	12.
Saturn	0.47	0.24	1.40	0.66	0.34	0.34	16.	2.3
Uranus	0.51	0.21	1.40	0.71	0.29	0.30	0.78	0.12
Neptune	0.41	0.25	1.25	0.51	0.31	0.29	0.29	0.036

Geometric albedo of a Lambert sphere. It is important to note that a Lambert sphere has a geometric albedo of $A_g = 2/3$. The surface brightness of a Lambert disk of normalized radius $R = 1$ is constant over the whole disk and one can write for the normal retro-reflection ($\theta = 0$):

$$I_{\rm disk}(r) = \frac{F_i}{\pi} \quad \text{and} \quad \int_0^1 I_{\rm disk}(r)\, 2\pi r\, dr = 2F_i \int_0^1 r dr = F_i$$

A sphere (not a disk) at zero phase angle has a surface brightness distribution with a limb darkening which behaves for the normalized radius $0 \le r \le 1$ like

$$I_{\rm sph}(r) = F_i \frac{\cos\theta'(r)}{\pi} = F_i \frac{\sqrt{1-r^2}}{\pi}\,.$$

Figure 4.4: Schematic difference of the geometric albedo of a Lambert disk and a Lambert sphere.

The angle θ' is the angle of incidence with respect to the surface normal which depends on the radial distance $r = \sin\theta'$ measured from the center of the illuminated hemisphere (apparent disk). The sub-solar point reflects like a disk (surface brightness F_i/π) but the irradiation of the more an more inclined surface towards the limb results in a reduced back-scattering because the strongest scattering occurs along the surface normal.

Integration for a fully illuminated Lambert sphere yields:

$$\int_0^1 I_{\rm sph}(r)\, 2\pi r\, dr = 2F_i \int_0^1 r\sqrt{1-r^2}\, dr = 2F_i\Big(-\frac{(1-r^2)^{3/2}}{3}\Big)\Big|_0^1 = \frac{2}{3}F_i\,.$$

A Lambert sphere reflects only 2/3 of the light for phase angle $\alpha = 0$ when compared to a Lambert disk because a substantial fraction of light is scattered into direction $\alpha > \pi/2$ what does not occur for an illuminated disk. Lambert disk and Lambert sphere scatter both all light and have a Bond albedo (or spherical albedo) of $A_B = 1$ but the angular distribution of the scattered light is different.

It is not surprising that the solar system planets have geometric albedos $A_g \lesssim 0.7$ when considering the case of the perfectly reflecting Lambert sphere. Averaged over all wavelengths the A_g should be smaller (about 2/3) than the Bond albedo A_B. This is roughly the case for Venus and Mars (see Table 4.2).

For Earth and the giant planets the situation is different. The geometric albedo in the visual is higher than the Bond albedo $A_g(\mathrm{V}) > A_B$. This indicates that the geometric albedo must be low at other wavelengths, what is the case for the IR wavelength regime because of molecular absorption by H_2O for Earth and CH_4 for the giant planets (see Slides 4.1 and 4.3).

Spherical albedo and Bond albedo. The spherical albedo $A_s(\lambda)$ gives the reflection in all direction and not only the normal retro-reflection as measured for the geometric albedo $A_g(\lambda)$. The spherical albedo is required for an accurate derivation of the Bond albedo A_B. A_B, which is used for energy budget calculations, is the flux weighted wavelength average of the spherical albedo:

$$A_B = \frac{\int_0^\infty F_i(\lambda) A_s(\lambda)\, d\lambda}{\int_0^\infty F_i(\lambda) d\lambda}\,. \tag{4.13}$$

With a scattering model of a planet it is easy to calculate the geometric albedo and spherical albedo. Observationally, one needs to know the scattering in all direction, what is a very difficult to achieve. For example, the reflection $f(\alpha)$ of Earth for a phase angle $\alpha = 90°$ will be different if mainly the white polar regions are seen from a polar direction when compared to the dark oceans as seen from equatorial directions.

One simple way to address the problem of the reflection into different directions is the phase integral q defined by

$$q = 2\int_0^\pi \frac{F_{\rm ref}(\alpha)}{F_{\rm ref}(\alpha=0)} \sin\alpha\, d\alpha\,,$$

where $F_{\rm ref}(\alpha)$ is a rotationally symmetric phase angle dependence of the reflected radiation normalized to the geometric albedo $F_{\rm ref}(\alpha = 0) = A_g$. With this definition the phase integral, geometric albedo, and spherical albedo are related by

$$A_s = A_g\, q\,. \tag{4.14}$$

It should be noted that this approach is only a first order approximation which is formally only correct for rotationally symmetric reflection from planets.

The phase integral q for special cases is:

- $q = 1$ for a Lambert disk,
- $q = 3/2$ for a Lambert sphere,
- $q = 4$ for (a theoretical) isotropically scattering body.

Some values of the phase integral for solar system planets are given in Table 4.2.

Reflectivity phase curves. The phase angle dependence of the reflected radiation from a planet is important for the analysis of observations. In general, planets are not observed at phase angle $\alpha = 0°$. For example, the inner planets, Mercury and Venus, are behind the sun for $\alpha = 0°$, and extra-solar planets are behind "their star". With direct imaging of extra-solar planets only data in the range $30° < \alpha < 150°$ can probably be obtained in the near future. For this reason one needs to study the reflectivity phase curves $F_{\rm ref}(\alpha)$ or the phase dependence of the reflection normalized to the geometric albedo:

$$f(\alpha) = \frac{F_{\rm ref}(\alpha)}{F_{\rm ref}(\alpha = 0)}\,. \tag{4.15}$$

Phase curve for a Lambert sphere. The phase curve for a Lambert sphere can be derived analytically by integrating the $\cos\theta$ reflection law of the visible part of the illuminated sphere as function of phase angle α. The solution is:

$$f(\alpha) = \frac{1}{\pi}(\sin\alpha + (\pi - \alpha)\cos\alpha)\,. \tag{4.16}$$

The phase curve for a Lambert sphere is plotted in Slide 4.4.

Flux contrast for reflecting extra-solar planets. Equation 4.12 is also valid for a very distant observer outside of the solar system or for the observations of extra-solar planets from Earth. In this case the distance of the observer to the central star $d_{\rm star}$ (which was $d_{\rm E}$ for an Earth-based observer looking at a solar system planet) and the distance from the planet to the observer D are equal and very large $d_{\rm star} = D \gg 1$ AU. Thus the contrast of a reflecting planet $C_{\rm ref}$ with respect to its illuminating star is

$$C_{\rm ref} = \frac{F_{\rm P}}{F_{\rm star}} = A_g(\lambda) f(\alpha) \frac{R_P^2}{d_P^2}\,, \tag{4.17}$$

where $A_g(\lambda)$ is the geometric albedo and $f(\alpha)$ a normalized phase function as described by Equation 4.15 which takes into account that the reflected light depends on the angle star - planet - observer. Table 4.2 gives the factors R_P^2/d_P^2 for the solar system planets and also estimates for the contrast of the reflected light for a phase angle $\alpha = 90°$.

4.4 Atmospheres of solar system planets

For spectroscopic studies of planets one needs to understand the net emission of the radiation from the surface or the atmosphere. Radiative transfer in planetary atmosphere is therefore a very important topic for the analysis of solar system objects but also for direct observations of extra-solar planets. In this section we discuss some basic properties of planetary atmospheres.

4.4.1 Hydrostatic structure of atmospheres

The planet structure equation from Section 3.2 apply also for planetary atmospheres. One can often make the following simplifications:

- the atmosphere can be calculated in a plane-parallel geometry considering only a vertical or height dependence z,
- the vertical dependence of the gravitational acceleration can often be neglected for the pressure range 10 bar – 0.01 bar and one can just use $g(z) = g(z = 0) = g(R) = g = GM_P/R_P^2$.
- the equation of state can be described by the ideal gas law

$$P(\rho) = \frac{\rho kT}{\mu} \quad \text{or} \quad \rho(P) = \frac{\mu P}{kT}\,,$$

  where μ is the mean particle mass (in [kg] or [g]),
- a mean particle mass which is constant with height $\mu(z) = \mu$ can often be used in a first approximation,
- a temperature which is constant with height $T(z) = T$ can often be used as first approximation.

Pressure structure. The differential equation for the pressure gradient is:

$$\frac{dP(z)}{dz} = -g(z)\rho(z) = -g(z)\frac{\mu(z)P(z)}{kT(z)}$$

which yields the general solution:

$$P(z) = P_0 e^{-\int_0^z 1/H_P(z)dz} \quad \text{with} \quad H_P(z) = \frac{kT(z)}{g(z)\mu(z)}\,.$$

For a homogeneous, isothermal atmosphere and $g(z) = g$, $\mu(z) = \mu$ a simple exponential pressure law is obtained

$$P(z) = P_0\, e^{-z/H_P} \quad \text{with} \quad H_P = \frac{kT}{g\mu}\,, \tag{4.18}$$

where H_P is the pressure scale height. This is the vertical length scale over which the pressure decreases by a factor $e^{-1} = 0.368$.

Density structure. Very similar equations apply for the density structure and one obtains for an homogeneous, isothermal atmosphere a density structure equivalent to the pressure:

$$\rho(z) = \rho_0 \, e^{-z/H_\rho} \quad \text{with} \quad H_\rho = \frac{kT}{g\mu} \, . \tag{4.19}$$

where H_ρ is the density scale height and ρ_0 the density at the reference point $z_0 = 0$. In the simple case described here the pressure and density scale heights are identical:

$$H_P = H_\rho = H \, .$$

Table 4.3: Basic atmospheric parameters for planets with atmospheres and Titan.

object	$T_{\rm ground}$	$P_{\rm ground}$ [bar]	$T_{\rm eff}$	$\mu/\mu_{\rm H}$	g [m s^{-2}]	H [km]	$dT/dz\|_{\rm ad}$ [K/km]	$v_{\rm esc}$ [km s^{-1}]
Mercury		10^{-14}	448 K		3.7			4.4
Venus	730 K	92	328 K	44	8.9	6.9		10.4
Earth	288 K	1.01	263 K	28	9.8	8.4		11.2
Mars	215 K	0.006	227 K	44	3.7	11.		5.0
Jupiter			124 K	2.3	23.1	19.	1.9	59.5
Saturn			95 K	2.3	9.0	38.	0.84	35.5
Uranus			59 K	2.3	8.7	24.	0.85	21.3
Neptune			59 K	2.3	11.0	19.	0.86	23.5
Titan	93 K	1.46	80 K	28	1.4	17.		2.6

Scale heights for planets. The equation for the scale height indicate the following relationships:

- The scale height depends on the atmospheric properties. For a planet with given radius R_p and bulk density $\bar{\rho}$ (or mass) the scale height is proportional to the atmospheric temperature $H \propto T$ and inverse proportional to the mean particle mass $H \propto 1/\mu$,
- The scale height depends for given atmospheric temperature and composition on the planet properties. The scale height is inverse proportional to the surface gravity $H \propto 1/g = R_P^2/GM_P \propto 1/R_p\bar{\rho}$.

The scale height can be particularly large for hot planets, with a hydrogen atmosphere and a small gravitational acceleration (large radius and low mean density).

For solar system planets the scale heights are given in Table 4.3. H was calculated with the indicated $T_{\rm eff}$ and mean particle mass $\mu/\mu_{\rm H}$. The scale heights are in a narrow range of 5 – 40 km. From the composition, one would expect much smaller scale heights for the terrestrial planets when compared to giant planets because of the much larger particle mass (≈ 30 in terrestrial planets but only 2.3 in giant planets). But this effect is compensated by the higher atmosphere temperature and lower gravity for the terrestrial planets.

Column density. The column density $\Sigma(z_0)$ gives the total density of gas per unit area above a certain height z_0 (e.g. defined as $z_0 = 0$). $\Sigma(z_0)$ is an important quantity for the calculation of the optical depth. Since the density drop-off with height is exponential $\Sigma(z_0)$ for is proportional to the density at $\rho(z_0) = \rho_0$

$$\Sigma(z_0) = \rho_0 \int_0^\infty e^{-z/H}\, dz = -\rho_0 H e^{-z/H}|_0^\infty = \rho_0 H\,.$$

This can be directly linked to the pressure

$$\Sigma(P_0) = \frac{\mu P_0}{kT} H = \frac{P_0}{g}\,. \tag{4.20}$$

All solar system planets have a gravitational acceleration at the surface of the order $g \approx 10\,\mathrm{m\,s^{-2}}$. Thus for order of magnitude estimates one can use a surface density of $\Sigma(1\mathrm{bar}) \approx 1\mathrm{kg\,cm^{-2}}$.

Chemical composition for atmospheres of solar system planets An important input parameter for the analysis of atmospheres is their composition which is given in Table 4.4. We will discuss later the interpretation of these abundances.

Table 4.4: Abundances by mass of the most important chemical spezies for solar system objects.

object	dominant molecule	secondary constituents	minor constituents
Venus	96.5 % CO_2	3.5 % N_2	0.01 % SO_2
Earth	78.1 % N_2	20.1 % O_2	0.93 % Ar, 0.03 % CO_2
Mars	95.3 % CO_2	2.7 % N_2	1.6 % Ar, 0.27 % N_2
Jupiter	85 % H_2	15 % He	0.24 % CH_4
Saturn	94 % H_2	6 % He	0.3 % CH_4
Uranus	85 % H_2	15 % He	1 % CH_4
Neptune	85 % H_2	15 % He	1 % CH_4
Titan	92 % N_2	4 % CH_4, 4 % Ar	

4.4.2 Thermal structure of planetary atmospheres

The vertical structure of planetary atmospheres can be characterized by their thermal structure which depends on the heating processes and energy transport mechanisms. For our discussion of these processes we take the Earth atmosphere as a guideline. The atmospheric temperature profile is used as basis to distinguish different atmospheric layers. The vertical structure of Earth atmosphere is illustrated in Figure 4.5 and described in Table 4.5.

Figure 4.5: Vertical structure of Earth atmosphere.

Heating processes. There are the following important heating processes for planetary atmospheres. Starting from the top to the bottom of the atmosphere these are:

- **Ionization.** Neutral atoms in the high atmosphere absorb easily the solar far-UV radiation. Each ionization by a photon with energy $h\nu$ above the ionization energy $h\nu_0$ of an atom will produce an energetic electron with the "excess energy" $\Delta E = h(\nu - \nu_0)$. Ionization is only important in the uppermost thermosphere, because further down there will be no ionizing photons left (see Slide 2.23).
- **Particle radiation** and plasma processes related to the planet magnetosphere can contribute to the heating of the upper atmosphere. For Earth, the impact of these effects depends a lot on the solar activity cycle and the associated enhancement of solar mass ejections and magnetic storms.
- **Photodissociation** of molecules by UV-photons is an important heating process in the stratosphere. In the Earth atmosphere the main processes are the dissociation of O_3 and O_2. Below the stratosphere there are no UV-photons left (see Slide 4.5).
- **Light absorption.** The optical and near-IR light gets absorbed in the troposphere at pressure levels around 1 – 10 bar. At these pressures collision induced absorption sets in, and the density of absorbing molecules becomes high enough for significant

Table 4.5: Parameters and boundaries of the different atmospheric layers in the Earth atmosphere

layer or boundary	z [km]	T [K]	P [bar]	comment
troposphere	0–12	290–215	1–0.1	heated by the surface, with a decreasing temperature gradient due to convection
tropopause	12	215	0.1	vertical temperature minimum and upper limit of the convection layer
stratosphere	12–50	215–270	0.1–10^{-3}	temperature increases due to absorption of UV radiation by O_3
stratopause	50	270	10^{-4}	intermediate temperature maximum
mesosphere	50–85	270–190	10^{-4}–10^{-6}	decrease in temperature due to the lack of heating processes
mesopause	85	190	10^{-7}	absolute temperature minimum in the atmosphere
turbopause	100	200	10^{-8}	below this level the composition is quite homogeneous, above it the particles are stratified according to their weight
thermosphere	85–500	190–1000	10^{-7}–10^{-10}	the gas is heated due to ionization by solar far-UV photons
exobase	~ 500	1000	$\approx 10^{-11}$	above this limit particles can escape, the height changes with solar activity
exosphere	$\gtrsim 500$	$\gtrsim 1000$	$< 10^{-11}$	composed mainly of H and He particles which escape to space

absorption of optical/near-IR light. The absorbed photon energy is converted first into intrinsic rotational or vibrational energy of the molecule, which is then transferred to thermal motion of the gas.

– **Surface heating.** The optical and near-IR light of the star can also be absorbed by the ground surface which is heated up. This provides a hot bottom for terrestrial objects.

– **Internal energy.** Gas giants are still contracting and therefore they have a steady upward energy flow which heats the troposphere from below.

Energy transport processes. The three major energy transport mechanisms for planetary atmosphere are radiation, convection and conduction. Usually one process dominates for the definition of the temperature structure in the atmospheres.

- **Convection** is the energy transport by vertical gas flows. It only sets in if the conditions for convection are favorable. There must be a dense gas and a fast temperature decrease with height for convection. All solar system planets show convection in their troposphere because the upper tropospheric layers cool efficiently by radiation, while the low troposphere is strongly heated by internal energy or irradiation (see Slide 4.6). Convection is discussed in detail in the following section.
- **Radiation** energy transport is important for optically thin atmospheric layers. Optical light from the sun (central star) is efficiently deposited in the lower troposphere while UV light is absorbed in the upper atmosphere. Thermal radiation emitted in the IR wavelength band can easily escape from the upper troposphere and all layer above causing always a cooling. Deep in the troposphere the emission of IR-light is not efficient, because the gas is optically thick in the IR range and the radiation energy is essentially "trapped".
- **Conduction** is the energy transport by collision between particles. Conduction is the mechanism which transfers the energy from a hot surface to the gas because the surface particles have a high kinetic motion. Conduction is also the process which transports the energy in the thermosphere and exosphere. Because of the large mean free path length between collisions ($\Delta s \gtrsim$ H) these outermost layers are homogeneous in temperature.

Temperature structure for solar system planets.

- **The giant planets** show also a troposphere and a tropopause with a temperature between 50 and 100 K. The stratosphere reaches a temperature of 150 K. The heating is due to photochemical absorption by haze (photo-chemical smog), while the cooling is mainly due to emission lines of C_2H_2 (acetylene) and C_2H_6 (ethane). Above comes the thermosphere where the temperature reaches about 800 – 1200 K. There is no well defined mesosphere for the giant planets.
- **Venus** has a troposphere which extends up to the tropopause at 70 km where the pressure is about 0.1 bar and which marks also the top of the cloud layers. Above this follows a constant temperature region up to about 100 km. Further above there exists a strong difference between a 300 K thermosphere on the day side and a much colder, only 100 K so-called cryosphere, on the night side. The temperature in the thermosphere is relatively low because CO_2 is a molecule which can efficiently cool the higher atmosphere.
- **Mars** has only a very thin atmosphere, lacking the density of "normal tropospheres and stratospheres". Thus, the temperature decreases from the surface temperature of about 220 K to 120 K above 50 km. The temperature increases then above 120 km to about 160 K. Mars has like Venus "no really hot" thermosphere because of the efficient CO_2 cooling.

4.4.3 Tropospheric Convection

Convection is the energy transport by gas flows and it is a dominant energy transport process in the troposphere. Convection will occur if the following conditions are fulfilled

- a gas parcel, which is slightly hotter, and therefore slightly less dense and lighter than its surroundings will start to rise,
- the ambient pressure decreases and the parcel expands, and cools adiabatically (heat transfer to the surroundings can be neglected),
- if the parcel is, after some upwards motion and adiabatic expansion (and cooling), still hotter and less dense than the surroundings then it will continue to rise in a convective flow.

Convection stops if the parcel is after some upward motion colder and denser than the surroundings. The condition for convection is determined by the relation of two temperature gradients:

- the adiabatic temperature gradient $dT/dz|_{\rm ad}$, which follows from the first law of thermodynamics (see below),
- the surrounding atmospheric temperature gradient $dT/dz|_{\rm atmos}$ which is determined by all heating, cooling and energy transport processes.

If the adiabatic temperature gradient is shallower than the atmospheric temperature gradient

$$-\frac{dT}{dz}\Big|_{\rm ad} < -\frac{dT}{dz}\Big|_{\rm atmos} \tag{4.21}$$

then the atmosphere is unstable with respect to convection and convection will set in. The above relation is often also given with a different sign $dT/dz|_{\rm ad} > dT/dz|_{\rm atmos}$. Equation 4.21 is equivalent to the following statements:

- convection may occur if the temperature decreases fast with height,
- convection does not occur if the temperature is almost constant with height or when the temperature rises with height.

Figure 4.6: Temperature gradients which are stable or unstable with respect to convection.

Derivation of the adiabatic lapse rate. The adiabatic temperature gradient or adiabatic lapse rate follows from the first law of thermodynamics which describes energy conservation:

$$dU = dQ - dW\,.$$

The change in internal energy dU of a gas is equal to the change in thermal energy (added or lost) of the gas dQ and the work done by or put into the system dW. The following relations are valid:

- $dQ = 0$: there is not heat exchange to the surroundings in an adiabatic expansion or compression process,
- $dU = mc_V\,dT = m(c_P - R_s)\,dT = mc_p\,dT - mR_s\,dT$ describes the change in internal energy, where m is the mass of the gas parcel, c_V and c_P the specific heat capacities at constant volume and constant pressure, and R is the specific gas constant,
- $dW = P\,dV = mR_s\,dT - (m/\rho)\,dP$ is the work put into the gas by compression or done by the gas by expansion. In addition there is $P\,dV = mR_s\,dT - (m/\rho)\,dP$, which follows from $P\,dV + dP\,V = mR_s\,dT$, the total derivative of the ideal gas law $PV = mR_sT$, and $V = m/\rho$.

Now we can rewrite the energy equation

$$dU = mc_p\,dT - mR_s\,dT = -mR_s\,dT + (m/\rho)\,dP = -dW$$

and obtain the adiabatic temperature-pressure gradient

$$\frac{dT}{dP} = \frac{1}{\rho c_p}$$

which yields with the hydrostatic pressure law $dP = -g\rho\,dz$ the **adiabatic lapse rate** as final result

$$\frac{dT}{dz} = -\frac{g}{c_p}\,. \tag{4.22}$$

This lapse rate is valid for a dry atmosphere where no condensation occurs.

Lapse rates for Earth atmosphere. For the Earth the adiabatic lapse rate is about -10 K/km. If condensation occurs then the "moist" adiabatic lapse rate should be used which is slightly different with a value of -5 K/km. The average atmospheric lapse rate is -6.5 K/km. This means:

- the Earth atmosphere is stable against convection, if no condensation occurs and the dry adiabatic lapse rate applies,
- the Earth atmosphere is unstable if condensation occurs and convection will take place as soon as the moist lapse rate is appropriate,
- cloud formation and condensation are closely connected to convection.

Convection in other solar system planets. All solar system planets with a substantial atmosphere have a troposphere where convection dominates. Convection is a very efficient mechanism for transporting energy whenever the temperature gradient is superadiabatic (temperature decreases faster than the adiabatic temperature gradient). This places a firm limit how fast the temperature increases with depth in planetary atmospheres.

The troposphere extends in all solar system planets from > 1 bar to a pressure level of about 0.1 bar (see Slide 4.6). This is the range where most of the thermal radiation is escaping from the planetary atmosphere. The main heat source due to the absorption of stellar radiation (and the internal heat for the giant planets) is below the troposphere. Thus, the troposphere is characterized by a strong heat source at the bottom and strong radiation losses at the top which leads naturally to the observed "convective" temperature structure.

4.4.4 Atmospheric escape

Particle escape from an atmosphere involves three steps. First a gas particles must be transported from the lower to the upper atmosphere. Then the particles must be transformed from an atmospheric gas particle, usually molecules, to neutral or ionized atoms which can then in a third step be accelerated to high speed and escape.

The basic process for escape is thermal or hydrostatic escape, where particles in the high atmosphere have thermal velocities which are large enough for escape. In addition the density must be low enough that the particle does not collide on its escape trajectory.

The escape velocity $v_{\rm esc}$ for a particle is reached if its kinetic energy is equal to its potential energy (the energy required to leave the planet):

$$\frac{1}{2} m v_{\rm esc}^2 = \frac{m M_P G}{R_P} \tag{4.23}$$

which yields:

$$v_{\rm esc} = \sqrt{\frac{2GM_P}{R_P}} \,. \tag{4.24}$$

Figure 4.7: Schematic shape of the Maxwell-Boltzmann velocity distribution function.

The particle velocity for a gas in thermal equilibrium can be described by the Maxwell-Boltzmann velocity distribution which gives the number of particles per velocity bin dv

$$n(v)dv = \frac{4n}{\sqrt{\pi}}\Big(\frac{m}{2kT}\Big)^{3/2} v^2 e^{-mv^2/2kT} dv\,,$$

where m is the particle mass, and n the total number density of particles and T the temperature of the gas.

We may use the most likely velocity $\bar{v}$ of this distribution for an estimate on the particle velocity:

$$\bar{v} = \sqrt{\frac{2kT}{m}} \tag{4.25}$$

For T one should use the temperature of the thermosphere which is for giant planets and Earth around $T \approx 1000$ K. The velocity distribution for large velocities decays exponentially. This means that in a gas there are always a small fraction of particles with velocities which are a factor of a few higher than $\bar{v}$. If always a small fraction of particles of a certain kind can escape then after some time (millions or billions of years) a substantial amount of particles may escape.

Considering the formulas for $v_{\rm esc}$ and $\bar{v}$ one can easily derive the following dependencies:

- light particles, in particular hydrogen, escape much easier than heavy particles such as C, N, or O. This explains why Venus, Earth and Mars have essentially no H gas but still CO_2, N_2, O_2 gas made of the elements C, N, or O.
- a planet with high escape velocity (essentially a planet with a large mass) can keep much better an atmosphere. This explains why the Earth has an atmosphere and the moon has none.
- A planet with a cold exosphere will have lower thermal velocities and keep more easily an atmosphere. This may explain why the strongly irradiated planet Mercury has no atmosphere while Titan has one.

4.4.5 Evolution of the chemical composition of planetary atmospheres

The giant planets have a composition which is close to the solar composition. Hydrogen and helium dominate strongly. Atmospheres with substantial amounts of hydrogen are called "reducing atmospheres". They contain a lot of methane CH_4, ammonia NH_3, and also water vapor H_2O and hydrogen sulfide H_2S are expected. But essentially only CH_4 and NH_3 were detected while the other constituents are expected to be trapped in the deeper layers. Not much evolution in composition seems to have been taken place for the solar system giants since their initial formation. The situation is very different for the atmosphere of terrestrial planets.

Primitive and secondary atmospheres of terrestrial planets. The atmospheres of the terrestrial planets evolved strongly in the past. If Earth ever had a H and He rich atmosphere then it was lost completely. Also Ne is strongly underabundant indicating that it never existed in large quantities in Earth's atmosphere or that it was lost during a very early epoch. Therefore, it is assumed that all volatile elements in the Earth atmosphere originate from volcanic processes and outgassing. This provides, as observed, the elements

H, C, N, O, and S but not the noble gases He and Ne. Ar is quite abundant in the atmosphere of Earth because of the radioactive decay of ^{40}K.

For a given elemental composition we may expect abundances as expected from chemical equilibrium. The following reactions are important for the composition of the terrestrial planets:

$$\begin{aligned} CH_4 + H_2O &\leftrightarrow CO + 3H_2 \\ 2NH_3 &\leftrightarrow N_2 + 3H_2 \\ CO + H_2O &\leftrightarrow CO_2 + H_2 \end{aligned} \tag{4.26}$$

The escape of hydrogen shifts the equilibrium to the right and this can explain the predominance of CO_2 and N_2 in Venus, Mars and Titan.

In the Earth atmosphere there is a significant lack of CO_2. Carbon dioxide is dissolved in the water oceans and blocked as calcium carbonate $CaCO_3$ as rock. Volcanic outgassing brings then partly the CO_2 back into the atmospheres.

The abundance of O_2 in the Earth atmosphere is mainly due to photosynthesis of green plants. According to chemical equilibrium O_2 would disappear rapidly (on geological atmospheres) through oxidation from the atmosphere. This non-equilibrium chemistry is an important signature of life on Earth.

4.5 Spectra of substellar objects

4.5.1 New spectral types

M-type main sequence stars were for about a century the lowest mass objects known outside the solar system. The first substellar object, GD 165B, was detected in 1988. At that time it was a "strange" companion to a white dwarf star. The revolution came around 1997 after the development of instruments with large and sensitive IR array detector. Near-IR sky surveys, like 2MASS (2-micron all sky survey), found nearby very cool objects. Many of the new objects have a spectrum like GD 165B which changed its status from a strange object to a prototype of the new class of brown dwarfs. New spectral classes based on red and near-IR spectra had to be introduced for extending the standard spectral sequence of stellar objects. Dominant absorptions are molecular bands from H_2O water vapor and CH_4 methane, which are key features for the definition of the new spectral types L, T. Up to now (2013) many hundred L and T dwarfs were detected. The introduction of another class for even cooler objects, the Y dwarfs, is currently discussed.

We provide here a brief characterization of the spectral classes for low mass objects. Slide 4.7 and 4.8 show the transistion of spectral features from M- to L- and T-type objects in the 700 – 900 nm range and in the near-IR respectively. Table 4.6 summarizes some key properties of these systems.

Spectral class M: Many bright stars are of spectral type M, like e.g. α Sco or α Ori, but these are all evolved high or intermediate mass giant stars. M-stars are red objects indicating that they must be cool < 4000 K. No M-star on the main-sequence is visible to the naked eye, because they are too faint and emit their radiation mainly in the near-IR range.

Their spectral classification was traditionally based in the visual wavelength range. The red spectrum of M-stars is dominated by molecular bands of TiO. These absorption become stronger for cooler objects providing a well established scheme for the association of spectral types to surface temperatures. The TiO-band at 705 nm is an important spectral feature for the classification of M-stars. Other absorptions, like CaH or VO, are used to refine the classification.

Spectral type L: The spectral type L describes objects with strong H_2O absorptions in the near-IR, strong resonance lines of the alkali atoms NaI, KI, RbI and CsI, absorptions form metal-hydrates CrH and FeH. The warmer (early) L dwarfs show still the TiO like the M-stars but their strength decreases rapidely with decreasing photospheric temperature. The water vapour absorptions of L-stars are difficult to observe with ground-based observations because the same absorptions are present in the Earth atmosphere and they often prevent accurate measurements in the corresponding spectral bands. Typical temperatures for spectral type L are in the range 1300 to 2000 K.

Spectral type T: T-dwarfs have a surface temperature of about 700 to 1300 K and are characterized by strong absorptions of methane in the near-IR. They emit a lot of light in a few band between 1 and 1.6 μm, because the CH_4-bands block the emission of radiation from the photosphere efficiently around 1.1 and 1.4 μm and in the range between 1.7 - 3.5 μm. T-dwarf are therefore blue objects in the colors J - H, J - K, and J - M. This is one of the reasons why they were initially not found because according to their IR-colors they looked like "uninteresting" blue background stars.

Table 4.6: Characteristics of the spectral types for low mass objects.

spectral type	V–K	J–K	T [K]	spectral features
M0	3.9	0.8	3800	weak TiO (e.g. 705 nm), CaH, ...
M5	6.0	0.9	2800	strong TiO, ...
M9	7.5	0.9	2400	strong TiO, FeH (870 nm), weak H_2O(near-IR)
L2		1.3	2200	strong H_2O (near-IR), KI, CrH, weak TiO
L8		1.8	1500	strong H_2O, FeH, CrH, ...
T2		0.8	1200	CH_4, H_2O, ...
T6		–0.2	900	strong CH_4
Y			<700	

according to Kirkpatrick, 2005, Annu.Rev.Astron.Astrophys.,43,195

Chapter 5

Hot jupiters

Close-in giant planets, which are also called hot jupiters, form a group of well studied extra-solar planets. They are relatively easy to detect with the radial velocity or the transit method and their mass, radius, and orbital parameters are well known. With additional, more precise measuring methods, more information can be obtained and it is now possible to compare physical models of planets with observations of at least some prototype objects. This is particularly interesting, because hot Jupiter have system parameters, which differ strongly from what we know from solar system objects. In this chapter we address a few key findings about hot jupiters.

5.1 Origin and evolution of close-in planets

5.1.1 Inward migration

The detection of 51 Peg b and other giant planets in very close orbits was unexpected. According to planet formation theories it is not possible that a planet forms at such a location. Circumstellar gas close to a star, is far to hot to form a compact body or to be accreted by a compact body, ie a planet core, to build up an extended envelope. Therefore, there must exist processes which lead to the migration of giant planets from their birth place at a separation > 1 AU to their current position at $d_p \approx 0.1$ AU. Two migration scenarios are often discussed in the literature: disk migration and dynamical interactions.

Disk migration: For disk migration it is assumed that a newly formed planet is interacting with the planet-forming gas disk. The planet opens a gap in the disk and may transfers angular momentum to the disk gas outside the gap. This leads to a reduced inflow of gas. On the other side the angular momentum loss of the planet produces an inward drift and the accretion of lower angular momentum gas from inside the disk gap. Such scenarios were calculated and several detailed models were described.

Disk migration is a well established scenario which will be active in protoplanetary disk. However it is unclear whether this leads only to small or also to strong changes in orbital separation. It is also not clear how this process could stop the migrating planet from falling into the star.

Disk migration would lead to close-in planets with nearly circular orbits and small obliquities between planet orbit and stellar rotation axis. It is possible to explain some of

the close-in planets with this disk migration, but certainly not those with highly eccentric orbits or planets on orbits "counter-rotating" with respect to the stellar rotation.

Dynamical interaction. Dynamical interactions are expected to occur in young planetary systems. It is well known from the solar system that the "normal" planet formation process in a protoplanetary disk produces initially many asteroids and proto-planets. A large number of planets circling around a star have not enough space to remain on stable orbits without mutual gravitational interaction. This must lead to an evolution of the planet configuration. It is well possible for planetary system with an initial configuration similar our solar system that they went through a violent dynamical re-organization. For example if the orbital separation between Jupiter and Saturn would be smaller, then this could lead to a two-body interaction. A possible outcome is, that Saturn loses a lot of angular momentum and enters a shorter period, highly elliptical orbit around the sun which evolves then through tidal interaction with the sun into a tight circular orbit. Jupiter would then acquire more angular momentum and obtain a longer period orbit.

Several observational facts support such a planet interaction scenario:

- The radial velocity surveys show that the typical close-in planet are lower mass giant planets $\lesssim$ M$_J$. If there is a second planet further out in the system then this second planet is typically more massive (see Slide 2.13).
- Many giant planets with rather short periods $P < 3$ years have highly eccentric orbits (Slide 2.14). They may have gone through an interaction event but they do not interact enough with the central star or other planets to evolve rapidly towards a tight, circular orbit.
- The Kepler satellite found for transiting hot jupiters essentially no second transiting planet. This indicates that systems with hot jupiter have a different orbital configuration than the many systems with multi-planet transits (Section 3).
- With the Rossiter-McLaughlin effect it was demonstrated that hot Jupiter have often orbital orientations which are not aligned with the stellar rotation. For example the orbit of the planet can be retrograde with respect to the stellar rotation. This is probably the strongest argument for the dynamical interaction scenario and not explainable with disk migration. Observations of the Rossiter-McLaughlin effect are described in the next paragraph.

5.1.2 Rossiter-McLaughlin effect

For a rotating star the obscuration of a surface region by the planet during transit produces a small RV-effect. Because the planet reduces the signal from one side the rotationally broadened line becomes asymmetric and a net RV effect result. About 100 years ago this effect was described for the first time for binary stars by Rossiter and McLaughlin.

For extra-solar planet the impact of the Rossiter-McLaughlin (RM) effect on the RV is of the order

$$\Delta v_{\rm RM} = \frac{R_P^2}{R_S^2}\, v_{\rm rot} \sin i\,.$$

For a rotation velocity of $v_{\rm rot} \sin i = 1$ km/s the effect of a Jupiter-sized planet is at the level of 10 m/s (see Slide 5.1 for an example).

Figure 5.1: Illustration of the Rossiter-McLaughlin effect.

Depending on the mutual orientation of stellar rotation with respect to the orbital rotation the RV-excursion due to the transit attenuation looks different:

- For the same orientation of the stellar rotation and the orbital motion of the planet (or mutual inclination orbit - rotation $i_{\mathrm{o-r}} \approx 0°$) the approaching hemisphere is first attenuated by the planet and then the receding hemisphere. It results a RV-excursion which shows first a redshift (positive deviation) and then a blueshift.
- For a counter-rotating planet with respect to the stellar rotation (mutual inclination $i_{\mathrm{o-r}} \approx 180°$) the RV-excursion during the transit is first blue-shifted (negative) and then redshifted.
- For a stellar rotation which is inclined with respect to the orbital rotation of the planet the situation is more complicated. In this case we might have an inclined view onto the stellar rotation. The shape of the RV-excursion depends then on the orientation of the stellar rotation, the mutual inclination between planet orbit and stellar rotation and the impact parameter b. Thus the transit may occur mainly in front of the approaching or the receding hemisphere. In general there is a trend that the RV-excursion becomes more blue-shifted during the transit if $i_{\mathrm{o-r}} < 90°$ and more red-shifted if $i_{\mathrm{o-r}} > 90°$.

The Rossiter-McLaughlin effect can be measured like a RV-measurement for the determination of the reflex motion. Of course the sampling of measurements must be particularly high during the eclipse phase for a detection of the RV-excursions (see Slide 5.1).

5.1.3 Evolution of close-in planets

Not much is known about the evolution of close-in planets. It is likely that there exists an evolution from longer orbital periods to shorter orbital periods. From the period distribution, one can distinguish three regimes for the orbital period.

- Close-in planets with a period > 10 days are often on (highly) elliptical orbits which may result from previous dynamical effects. The evolution into short period circular orbits can be induced or at least accelerated by *gravitational interaction* with a second planet or a binary companion star located further out in the system. An alternative is the *tidal interaction* with the star, which damps the orbital eccentricity while reducing the orbital separation.
- Many close-in planets "pile-up" in the period-distribution between $P = 3 - 5$ days. This must be a more stable configuration. It is unclear how stable such orbits are but they could be stable for several Gyr. If this is true, then the lifetime of a close-in planet is terminated by the stellar radius evolution, which slowly expands during its main-sequence life-time.
- Systems with periods shorter than 2 days are very rare. This suggests that planets on such tight orbits are rapidly spiraling into the star. The most likely reason is that the planet induces strong tidal effects on the star which break the orbital motion. It is unclear how long the time scale for such an infall is. Estimates range from 10 Myr to 1 Gyr for a close-in giant planet with a period of 2 days.

Unfortunately, the physical processes responsible for the orbit evolution are not well understood yet. One key issue is, that it is difficult to determine the age of the parent stars for an observational estimate of the typical time scales for different evolutionary processes.

5.2 Atmospheres of hot jupiters

With very high precision measurements of transit light curves and secondary eclipse depths it is possible to probe spectral features from the atmosphere of hot jupiters. This is useful for a better understanding of planetary atmospheres. However, one must also consider the special conditions for hot jupiters. For example, it is well known that the strong irradiation from the star is responsible for enhanced planet radii (Slide 3-24).

5.2.1 Secondary eclipse amplitude

If the secondary eclipse of the planet by the star can be detected then one can measure the intrinsic brightness $F_P(t_e)$ of the planet for the eclipse phase. $F_P(t_e)$ is equivalent to the drop in brightness due to the eclipse (see Fig. 3.7). With secondary eclipse detections one measures really the photons from the planet. This is strictly speaking a direct detection of the planet. The secondary eclipse depth can be used for the following measurements:

- for the reflected light the secondary eclipse gives the geometric albedo of the planet $F_P(t_e) = A_g$,
- for the thermal light the secondary eclipse provides the opposition brightness, or the brightness $F_P(\alpha)$ for phase angle $\alpha = 0^\circ$ which is equivalent to the brightness of the illuminated atmosphere,

- during secondary eclipse only the light from the star is seen. If the system is fainter during secondary eclipse than before and after the transit (Fig. 3.7) then one can also measure the brightness of the "backside" of the planet $F_P(t_t)$ or $F_P(180°)$.
- particularly interesting are secondary eclipse measurements for different wavelengths because this provides the spectral energy distribution for the planet.
- if even the phase curve of the planet is detected $I_P(t)$ or $I_P(\alpha)$ then one can investigate the surface brightness of the planet as function of longitude.

Spectral energy distribution for planets: If the spectral energy distribution (SED) of the planet can be attributed to the thermal radiation then one can compare the relative secondary eclipse signal with Planck-curves for the star and the planet and derive the temperature of the illuminated hemisphere. As discussed in Sect. 2.4 the flux ratio between planet and star is:

$$\frac{F_P}{F_S} = \frac{R_P^2}{R_S^2} \frac{B_\lambda(T_P)}{B_\lambda(T_S)} \,. \tag{5.1}$$

For hot Jupiters, $T_P > 1000$ K, and mid-IR wavelengths in the Rayleigh-Jeans tail of the Planck spectrum $\lambda \gg \lambda_{\max}$ this reduces to

$$\frac{F_P(\lambda \gg \lambda_{\max})}{F_S} = \frac{R_P^2}{R_S^2} \frac{T_P}{T_S} \,.$$

The planet SED gives the temperature on the illuminated and eventually also for the "backside" of the planet.

Spectral analysis: In recent years a lot of effort has been put into the analysis of the spectral dependence of the secondary eclipse depths in order to find spectral features of hot jupiters in the infrared spectral range. Most of these studies are based on data from the SPITZER satellite. The analysis is very difficult, because this instrument was not designed for high precision photometry. Slide 5.2 shows an example of such observations. The detections of various spectral features have been reported, but often these claims remained controversial. It is expected that future infrared satellites, like the JWST, will provide a break-through in this field.

Phase curves: The measurement of phase curves is also very interesting. The analysis of HD 189733 b revealed an offset between the substellar point and the hottest spot on the illuminated hemisphere (Slide 5.3). This indicates that some kind of gas circulation must be present. The offset angle provides strong constraints on the time-scales for atmospheric cooling and atmospheric circulation. From the measured offset in HD 189733 b it was possible to estimate the speed of the atmospheric flows which must be close to the sound speed or supersonic. Because it was also possible to measure the temperature difference between the illuminated and non-illuminated side of the planet, one can construct already quite realistic models about the heat transfer from the illuminated hemisphere to the night side and the overall energy budgets.

5.2.2 Transit spectroscopy

The transit depth may depend on wavelength because the uppermost atmosphere of the transiting planet absorbs the light from the star for certain wavelength more efficiently than for others. This means that the measured transit radius or the size of the planet's dark silhouette depends on wavelength.

Figure 5.2: Geometry for transit spectroscopy.

The same method can be used with higher spectral resolution. Because some light from the star passes during the transit through the uppermost atmosphere certain wavelength may be absorbed stronger than others because they coincide with a strong line or band absorption from the planetary atmosphere. This method has the potential to provide atomic and molecular abundances for the uppermost atmosphere of extra-solar planets.

The strength of this effect is given by the cross-section of the partly transparent upper atmosphere. Giant planets with a large scale height are the best candidates for the detection of spectroscopic transmission features. An estimate for the differential atmospheric absorption can be obtained from the circumference of the planetary disk times the atmospheric scale height:

$$\delta_{\rm atm} = 2\pi R_P \, \frac{kT_P}{\mu g} \frac{1}{R_S^2} \, .$$

Hot, giant planets with a mass like Saturn (low surface gravity = large scale height) are the best targets for such transit spectroscopy.

It is quite difficult with current instrumentation to measure the transit spectrum or the spectral dependence of the effective radius for the planet transit. The result of a study for HD 189733 b is shown in Slide 5.4. The derived transit spectrum is featureless, with a small but steady decrease in effective radius of about 1 % from the UV to the red spectral range. This wavelength dependence can be explained by Rayleigh scattering from molecules and/or photochemical haze in the uppermost layers of the atmosphere, which absorb blue light (400 nm) more efficiently than red light (700 nm).

Chapter 6

Direct imaging of extra-solar planets

Direct imaging for extra-solar planets means that emission from the planet can be spatially resolved from the emission of the bright central star. The two key requirements for a detection of extra-solar planets are

- a high contrast,
- a high spatial resolution.

A good example is the detection of the planetary system around the star HR 8799 (see Slide 6.1).

If a successful detection is achieved then the observation may provide the following information for extra-solar planets:

- the apparent brightness of a planet as function of time and spectral pass band,
- from the IR-brightness one can determine the spectral energy distribution of the thermal emission and derive the surface temperature,
- from the intensity and polarization of the reflected light one can derive albedos and surface scattering properties,
- variability studies give indications about phase effects, seasonal effect, rotational effects and weather changes,
- a spectral analysis allows to gain information about the atmosphere or surface composition, and one may also search for biosignatures.

In principle the direct imaging can also be used to measure the reflex motion of the star and derive the mass of the planet if the astrometric precision of the instrument is good enough.

6.1 Science requirements

Depending on the planet type and the science goals the instrument must achieve different requirements. We distinguish three cases:

- young planets for which the thermal radiation produced by self-contraction is the dominant energy source,

- the thermal radiation of old planets for which the irradiated and reprocessed energy from the central host star is the main energy source; old just means that internal energy sources are not dominant,
- the reflected light of planets.

Thermal radiation from young planets. Newly formed planet will be hot because of the potential energy which is transformed during the formation and contraction phase into thermal energy. In general the evolution of the thermal luminosity is a function of planet mass and age (as will be discussed in a following chapter). An approximate description is

$$L_P(M_P,t)\,[L_\odot] \approx 10\,(M_P\,[\mathrm{M}_J])^2\,\frac{1}{t[\mathrm{yr}]} \quad \text{for} \quad t > 1\,\mathrm{Myr}\,.$$

The luminosity of a planet L_P shows roughly an exponential decay and L_P is larger for higher mass planets. The luminosity evolution goes together with a radius and surface temperature evolution, and the planet to star luminosity contrast is given by:

$$C_{\mathrm{young}} = \frac{L_P(M_P,t)}{L_S} = \frac{R_P^2(M_P,t)}{R_S^2}\frac{T_P^4(M_P,t)}{T_S^4}\,.$$

For long wavelengths, in the Rayleigh-Jeans limit, the flux contrast is

$$C_{\mathrm{young}}(\lambda \gg \lambda_{\mathrm{max}}) = \frac{L_P(M_P,t)}{L_S} = \frac{R_P^2(M_P,t)}{R_S^2}\frac{T_P(M_P,t)}{T_S}\,.$$

The brightness and temperature of young planets are independent of the separation d_P and the luminosity of the star L_S. Therefore, young hot planets at large separation are relatively easy to detect, like in the case of the HR 8799 system. Also, a young contracting giant planet around a low mass star would be an easy target for direct imaging. Unfortunately, there are not many young stellar systems in the solar neighborhood which may harbor young, bright, self-contracting planets.

Thermal radiation from irradiated planets. The energy emitted by the thermal radiation of an old planet is assumed to be equal to the irradiated energy. Thus one can neglect internal energy sources. For planets with $T < 1000$ K the maximum of the thermal radiation is in the mid- or far-infrared spectral region. The contrast between planet and sun was already derived for solar system objects in Section 4.2. The contrast is less extreme for the Rayleigh-Jeans part of the Planck-spectrum of the planet at long wavelengths in the mid-IR or far-IR:

$$C_{\mathrm{old}}(\lambda \gg \lambda_{\mathrm{max}}) = \frac{F_P(\lambda \gg \lambda_{\mathrm{max}})}{F_S} = \frac{R_P^2}{R_S^2}\,\frac{T_{\mathrm{eq}}}{T_S} = \Big(\frac{1-A_B}{4}\Big)^{1/4}\Big(\frac{R_P}{R_S}\Big)^2\Big(\frac{R_S}{d_P}\Big)^{1/2},$$

where we have used the following relation for the equilibrium temperature

$$T_{\mathrm{eq}} = \Big(\frac{1-A_B}{4}\Big)^{1/4}\Big(\frac{R_S}{d_P}\Big)^{1/2} T_S.$$

For shorter wavelength $\lambda < \lambda_{\mathrm{max}}$ there is an exponential drop-off of the planet brightness and the detection becomes very difficult. The flux ratio at long wavelength in the Rayleigh-Jeans regime depends on $1/\sqrt{d_P}$. The peak of the Planck curve moves also like $\lambda_{\mathrm{max}} \propto$

$1/\sqrt{d_P}$. Thus at shorter wavelength there is a good chance to pick-up the shorter period planets. Thus, the detection of the thermal emission of irradiated planets requires high contrast capabilities and high spatial resolution. For a planet with $T_{\rm eq} \approx 300$ K (in the habitable zone) one needs to observe in the mid-IR to far-IR range at wavelength of about 5 μm or longer.

Reflected radiation from planets. The reflected light is according to Section 4.3

$$C_{\rm ref} = \frac{F_{\rm P}}{F_{\rm star}} = A_g(\lambda) f(\alpha) \frac{R_P^2}{d_P^2} \, .$$

The reflected light from a planets depends strongly on the separation $C_{\rm ref} \propto 1/d_P^2$. Further there are also the phase dependence described by $f(\alpha)$ and the spectral dependence of the reflectivity or geometric albedo $A_g(\lambda)$ which need to be considered.

Figure 6.1: Contrast as function of wavelength for young and old planets.

Typical planet to star contrast ratios for a system at 10 pc. Table 6.1 recalls some values from Section 4 for a solar system analog at 10 pc. The contrast is for a given system configuration independent of the distance D but the apparent separation behave like $\propto 1/D$ and the flux of the star and the planet like $\propto 1/D^2$.

Table 6.1: Rough estimates for the expected contrast for Earth-like and Jupiter-like extrasolar planets and Earth-sized and Jupiter-sized young, hot planets.

planet	separation at 10 pc	$C_{\rm ref}$	$C_{\rm near-IR}$	$C_{\rm far-IR}$
"old planets"				
exo-Earth	0.1 arcsec	$2 \cdot 10^{-10}$	$2 \cdot 10^{-10}$	$4 \cdot 10^{-6}$
exo-Jupiter	0.52 arcsec	$1 \cdot 10^{-9}$	$1 \cdot 10^{-9}$	$2 \cdot 10^{-4}$
"young, hot planets"				
1000 K, R_E	0.1 arcsec	$2 \cdot 10^{-10}$	$5 \cdot 10^{-6}$	$1 \cdot 10^{-5}$
1000 K, R_J	0.52 arcsec	$1 \cdot 10^{-9}$	$5 \cdot 10^{-4}$	$1 \cdot 10^{-3}$

For stars with low luminosity the planet to star contrast is less extreme for a self-luminating, young planet. For "old" planets around low luminosity stars the contrast depends mainly on the separation and is therefore like for bright stars if the separation is the same. However, planets with the same surface temperature like Earth will have a smaller separation and therefore the contrast is more favourable for a detection.

Requirements on the telescope size from the spatial resolution. The angular separation of a planet is equal to d_P/D, where D is the distance to a planetary system. For a planet at 1 AU the angular separation is only 0.1 arcsec for a system at 10 pc and only 0.01 arcsec at 100 pc. Of course planets further out, at 10 AU, or 100 AU will have a correspondingly larger angular separation. In any case a high angular separation is desirable to resolve also the inner regions of planetary systems.

There is the fundamental diffraction limit for the spatial resolution of a telescope $\theta =$, which is given by the observing wavelength and the diameter of the telescope

$$\theta = \lambda/\mathcal{D}\,. \tag{6.1}$$

The inner working angle (IWA) of a high contrast imager is then the minimum angular separation at which a faint object can be detected near a bright star $\theta_{\rm IWA} \approx 2\ (\lambda/\mathcal{D})$. The factor 2 applies for the best high contrast instruments available today. Many instruments are not optimized for this task and then this factor is 3 or 5 with a correspondingly larger IWA. Table 6.2 gives some examples for existing and future telescopes for the inner working angle just considering wavelengths and telescope sizes. This table shows that the VLT has in principle enough spatial resolution to search for scattered light at 0.6 μm of a Sun-Earth analog out to a distance of 30 pc, while for the thermal radiation at 5 μm the object must be closer than 4 pc to be resolved from the hot star. However there are only 4 solar type stars within this distance. Therefore a 38 m telescope is required to find the thermal radiation of an Sun-Earth analog within 10 pc.

Table 6.2: Inner working angle (IWA) in milli-arcsec [mas] for different telescopes and wavelengths λ.

telescope	$\mathcal{D}$	IWA 2 λ/D			
		0.6 μm	1.6 μm	5 μm	10 μm
HST	2.5 m	96 mas	260 mas		
JWST	6.5 m		100 mas	310 mas	620 mas
VLT	8 m	30 mas	80 mas	250 mas	500 mas
E-ELT	38 m	6.3 mas	17 mas	53 mas	105 mas

The requirements on contrast and separation for the detection of an extra-solar Earth-like planet or a Jupiter-like object are shown in Fig. 6.2. The inner working angle of the telescopes given in Table 6.2.

Figure 6.2: Contrast vs. separation for Earth-Sun and Jupiter-Sun analogs and young, self contracting planets at 10 pc.

6.2 High contrast instrumentation

The science requirements for the direct imaging put very strong constraints on the planet to star contrast and the spatial resolution to be achieved by a "planet finder" instrument. Several key techniques needs to be used which differ a bit whether an instrument is used in space or on the ground.

Current "planet finder" instruments on the ground use the following basic concept to achieve the detection goal (Slide 6.2):

- a large telescope which provides a high spatial resolution,
- a powerful adaptive optics systems which corrects the wavefront distortions (the seeing) introduced by Earth atmosphere,
- a stellar coronagraph which suppresses the light from the very bright host star,
- differential detection techniques which disentangles efficiently the light from the planet from the residual light from the bright host star.

For a space instrument the concept is quite similar, except that the wavefront distortions from the atmosphere does not need to be corrected. Despite this, still a slow adaptive optics systems might be required to corrected for wavefront corrections due to aberrations introduced by the instrument. The following paragraphs discuss in more detail the different techniques.

Figure 6.3: Illustration of the wave-front aberrations introduced by the atmospheric turbulence.

Atmospheric turbulence and seeing. Atmospheric turbulence produces cells of different size scales ranging from about 0.01 m to 100 m. The cells have a distribution of temperatures and therefore also of densities with corresponding differences in the refractive indices. The irregular refraction produces for astronomical sources tilted wavefronts and light rays which deviate from a strictly straight line (Fig. 6.3). These phenomena are summarized under the term "astronomical seeing".

The seeing has the following effects (Slides 3.44 and 3.45):

- the images of point sources are split up into speckles,

Table 6.3: Characteristic parameters of the atmosphere which are relevant for seeing corrections with an adaptive optics system.

parameter	definition	dependencies	typical values
Fried parameter r_0	diameter of an atmospheric cell producing a phase error of 1 rad	$\propto \lambda^{6/5}$ $\propto \mathrm{am}^{-3/5}$	$r_0 \approx 0.2$ m (R-band)
coherence time τ_0	time interval for phase variation of 1 rad	$\propto \lambda^{6/5}$ $\propto \mathrm{am}^{3/5}$ $\approx r_0/v_{\mathrm{Wind}}$	$\tau_0 = 1 - 7$ ms (R-band) $\tau_0 = 4 - 20$ ms (near-IR)
isoplanatic angle Θ_0	angular distance with phase error less than 1 rad	$\propto 0.3 r_0/(h \cdot \mathrm{am})$	few arcsec for R-band, few tens arcsec in near-IR
seeing FWHM_s	width of the PSF of a point source	$\approx \lambda/r_0$ $\propto \lambda^{-1/5}$	$\lambda = 1$ μm, $r_0 = 0.3$ m, $\mathrm{FWHM}_s = 0.69''$
diffraction limit	width of the diffraction limited PSF	$\approx \lambda/D$	$\lambda = 1$ μm, $D = 8$ m, $\mathrm{FWHM} = 0.026''$

am: airmass, h: height of the turbulent layer,

- the number of speckles N increases for more turbulent atmospheres (smaller cell scale r_0) and larger telescope diameter $\mathcal{D}$ like $N \propto \mathcal{D}^2/r_0^2$,
- the angular size of the speckles is determined roughly by the diffraction limit of the telescope ($\sim \lambda/\mathcal{D}$),
- the speckle pattern changes rapidly with time τ_0 (within ms),
- in long exposures the changing speckle pattern results in a blurred image and the angular diameter of a point source is $\approx \lambda/r_0$. The blurred point source image is called the *seeing disk*,
- the source brightness shows scintillation.

6.2.1 Adaptive Optics

The goal of an Adaptive Optics (AO) system is the correction of the wavefront deformations introduced by the atmosphere and the instrument. AO can provide diffraction limited ground-based observations.

The basic concept of adaptive optics consists of a wavefront sensor (WFS) which measures the wavefront distortions, a fast real-time computer (RTC) which calculates the corrections to be applied by deformable mirrors (DM) to the light beam coming from the telescope (see Fig. 6.7). The wavefront analysis and corrections are usually carried out in the pupil planes.

.

Figure 6.4: Block diagram for an adaptive optics system.

Basic requirements of the AO system are:

- the availability of a light source suitable for a wave-front analysis,
- measurement of the wave-front distortion with a wave front sensor (WFS) with a precision of about 1/20 λ,
- correction for the wave front distortion with deformable mirrors or other active optical components to a level of about 1/20 λ,
- good correction for atmospheric seeing requires corrections for the wavefront distortion on a spatial scale of about the Fried parameter r_0 with a speed of about the coherence time τ_0.

Strehl ratio. The Strehl ratio S is a measure for the performance of an AO system. The Strehl ratio is the peak intensity of the AO corrected point source relative to the peak intensity of a perfect, only diffraction limited PSF. The Strehl ratio can be related to the residual (rms) wave front aberrations σ (after the wave front correction):

$$S = \exp^{-2\,(2\pi\sigma)^2} \tag{6.2}$$

where σ is in units of the wavelengths of the radiation considered. For a good AO system with $S > 0.67$ the aberrations are at a level of $\sigma < \lambda/14$. To achieve this performance for $\lambda = 1$ μm means that the residual wavefront aberrations are less than 70 nm rms.

The requirements on the AO system are much more demanding for large telescopes and shorter wavelengths. For the same AO performance or Strehl ratio the number of required sub-apertures increases like $\propto D^2$ and $\propto \lambda^{-6/5}$.

Wave front sensor. A wave front sensor measures the tilt of the wave-front for sub-apertures in the pupil plane. A perfectly plane wave would show no angular gradient or tilt over the entire pupil. Most popular devices are the Shack-Hartmann wavefront sensor and the Pyramid wavefront sensor. We discuss here only the Shack-Hartmann sensors in more detail.

In a **Shack-Hartmann wavefront sensor** the pupil is divided into many sub-apertures using a micro-lens array which forms for each sub-aperture a point on a detector (Fig. 6.5). A wavefront with a local tilt (or gradient) induces then an $\Delta x, \Delta y$ shift of the point on the detector which is proportional to the wavefront gradient.

Figure 6.5: Principle of a Shack-Hartmann wave-front sensor.

For a good AO correction the Shack-Hartmann wave-front sensor should be able to measure the point offsets $\Delta x, \Delta y$ for each r_0-sub-aperture every ≈ 1 ms. One should note that enough photons must be collected per sub-aperture and exposure to determine the centroids.

Deformable mirrors The wave front correctors must compensate for the measured wavefront deformations to a precision of about $1/10 - 1/20\ \lambda$ for many sub-apertures within about a millisec. Essentially all wave front correctors are based on the deformable mirror concept. Important parameters of deformable mirrors are:

- the number of actuators which should match the number of atmospheric cells as defined by the Fried parameter in front of the telescope pupil (about 40 x 40 for a 8 m telescope),
- the actuator spacing which defines the size of the system,
- the dynamic range of the actuators defining the maximum wavefront correction,
- the response time.

Different types of deformable mirrors are used including $\approx 5-30$ cm mirrors fixed to piezo actuators with a spacing of a few mm, ≈ 1 cm sized micro-electro-mechanical devices with an actuator spacing of about 0.1 mm, or m-sized secondary mirrors of telescopes with magnetic actuators (like in loud speakers) with a spacing of several cm.

AO guide star. For the analysis of the wavefront deformations a light source is required which passes through the same atmosphere area as the target. For high contrast imaging of planets and circumstellar material the central host star is in principle an ideal light sources for the wave front analysis, because it is usually bright and ideally located in the middle of the target field. Using the central stars provides a good AO correction for a field of view of the size of the isoplanatic angle Θ_0 (Table 6.3). This choice sets the limit that the star must be bright enough to provide a star center measurement per sub-aperture and AO-loop period ($\approx$ coherence time τ_0). For extreme-AO systems with sub-aperture of 20 cm diameter this guide star limit is about 10 mag.

6.2.2 Stellar coronagraphs

The basic concept of coronagraphy was introduced by B. Lyot around 1930 for observations of the weak emission from the corona of the sun. Since that time, the concept has evolved to stellar applications. Stellar coronagraphy strongly differs from solar coronagraphy because the central source is point-like and includes therefore a strong diffraction pattern.

A stellar coronagraph is a starlight suppression device designed to reduce the on-axis starlight as much as possible. The basic concept of a Lyot coronagraph consists of

- an amplitude mask, the so-called Lyot mask, in the image plane to block the central (on-axis) star,
- a pupil mask, the so-called Lyot stop, in a pupil plane located after the Lyot stop.

Figure 6.6: Principle of a Lyot coronagraph.

The principle of the coronagraph can be understood by using Fourier optics, which relates the distribution of light in the image and pupil planes via Fourier transformations.

A simple round telescope pupil (or aperture) is described by a hat-function $f(r) = 1$ for $r < \mathcal{D}/2$ and $f(r) = 0$ for $r > \mathcal{D}/2$ produces an light wave amplitude dependence of a Bessel function $\propto 2J_1(r)/r$ which is the 2-dimensional equivalent for the diffraction pattern of a slit $I(x) \propto \sin x/x$. The intensity distribution in the image plane is given by

$$I(r) = I_0\Big(\frac{2J_1(r)}{r}\Big)^2,$$

describing the typical diffraction pattern consisting of a prominent central peak and many so-called diffraction rings. The PSF of a telescope with a central obscuration due to the secondary mirror differs not much from the simple unobscured pupil case. The central narrow peak is mainly due to light wave interferences of the outermost region of the pupil aperture.

The Lyot mask blocks now the central region of the focal plane. This provides a light distribution in the following pupil planes which is quite different from the input pupil. Instead of a uniform distribution inside the geometric aperture, the residual light is mainly located just inside and outside the nominal border of the relayed pupil.

The action of the Lyot stop is to reject the bright regions at the rim of the pupil suppressing the light in the strong low order diffraction rings.

An off-axis object missing the central opaque field mask, in our example the planet, is re-image in the final detector plane. In addition the relay pupil illumination by the planet light is still smooth and the Lyot pupil mask does not attenuate much of the light of the planet. The final image will then provide the point-like planet with a strongly reduced contribution from the central host star.

The stellar point source can be reduced to ≈ 1 % with a Lyot coronagraph and even better for more sophisticated systems. In the real world there are always quite substantial phase aberrations in the light wave due to the non-perfect system or the not perfectly corrected atmospheric turbulence which introduces a light halo with an integrated fractional intensity of the order $1 - S$, where S is the Strehl ratio achieved by the optical system at the position of the focal plane (Lyot) mask.

6.2.3 Differential imaging

The non-perfect correction of the atmosphere by the AO-system produces a residual, strongly variable speckle halo in the coronagraphic image. In addition there are quasi-static speckles which originate from aberrations in the instrument. For the currently available AO-systems and instruments there remains a halo of light from the bright host star in the final coronagraphic image. Typically, this background is (much) stronger than the expected signal of an extra-solar planet. For this reason one needs to apply differential imaging technique to extract the target signal. There are several types of differential techniques which are used or will be offered in new instruments.

Subtraction of the static or quasi-static instrumental features is sufficient for the detection of a target signal which is stronger than the variable, residual speckle halo from the atmospheric turbulence in the final focal plane image.

– **PSF-subtraction** is a very basic technique were the PSF-structure including all fixed instrumental features are subtracted from an observation using a PSF of a reference star. What remains after the subtraction is a difference image in which

faint sources in the surrounding of a bright star may be detectable. The reference star should have similar properties (brightness, color) as the target star and its PSF should be taken with exactly the same instrument configuration.

– **Angular differential imaging** is a more sophisticated version of the PSF subtraction. Observations of the target are taken with different sky orientations, where the sky orientation is rotated with a rotation of the telescope with respect to the sky, or with a rotation of the incoming beam with respect to the instrument. In the image plane the instrumental features remain stable while an off-center target moves. Subtraction corrects then well for instrumental effects while the signal from the target are preserved (see Slide 6.5).

If the signal of a faint companion or other circumstellar target is lower than the variable speckle halo then the correction of the static or quasi-static instrumental pattern is not sufficient. In this case one needs to distinguish between photons from the target and the bright star based on the physical properties of the photons.

– **Spectral differential imaging** is one way to search for planets, e.g. by the search of molecular absorption features which are present in the planet spectrum but not in the spectrum of the star. Young giant planets exhibit strong methane bands in the 1.0 μm - 1.7 μm region which are well suited for differential measurements. Slide 6.6 (left) shows the spectrum of a brown dwarf and of Saturn which illustrates the CH_4 bands which are expected to be also present in many young giant planets. Observations taken in two filters, one in the absorption band and one outside the absorption band, will therefore allow to search for a differential signal due to the presence of a cold objects which has molecular bands. If both images are taken simultaneously, with a double imaging system, then even the variable speckle pattern can be subtracted (Slide 6.7). An alternative is the use of an integral field spectrograph which takes spectra for each point in the field of view and which can then be searched for spectral features from a planet.
– **Polarimetric differential imaging** is a differential method for the search of scattered light from planets (Slide 6.6, right), or from the dust in circumstellar disks. Because the integrated light from the central star is unpolarized it produces no differential signal while the scattering from a planet or a circumstellar dust disk will produce a differential polarization signal.

6.3 The SPHERE "VLT planet finder"

Many high contrast instrument are currently built or available at major observatories. Competitive high contrast observations require large telescopes, good AO systems, and sophisticated differential techniques. The high contrast instrument currently available at the ESO VLT is NACO, an AO system with a high order deformable mirror with about 400 actuators, coronagraphs, and several instrument modes, including imaging, spectroscopy and polarimetry.

SPHERE Project overview. SPHERE is the abbreviation for the Spectro-Polarimetric High contrast REsearch project. SPHERE has been build during the past years and is currently tested for the installation and use at the VLT telescope next year as a 2nd generation high contrast instrument (Slide 6.8). The goal of the SPHERE instrument, is the discovery and study of extra-solar planets orbiting nearby stars by direct imaging of their circumstellar environment. The scientific requirements are very demanding because new planet detection will only be possible if the instrument achieves:

- a very large contrast between host star and planet, larger than 12.5^m or more than 10^5 in flux ratio,
- the high contrast must be achieved at a very small angular separation of $0.1'' - 0.5''$, inside the seeing halo.

SPHERE has different focal plane instruments for the detection of young and evolved planetary systems. Young planets are still contracting and therefore "hot" ($\approx$ 1000 K) and they emit a lot of thermal radiation in the IR which will be measured with differential imaging and integral field spectroscopy in the near-IR. Evolved planets are "cold" and their main emission is reflected stellar light which will be investigated with the differential polarimeter ZIMPOL (Zurich Imaging Polarimeter) in the visible.

The main components of the SPHERE experiment are (Fig. 6.7):

- an 8.2 m VLT telescope providing a diffraction limited resolution of 20 mas at 0.8 μm and 40 mas at 1.6 μm,
- an extreme AO system providing a high Strehl ratio,
- two coronagraphic systems which block the light from the bright host star, one coronagraph is installed in the near-IR science arm, the other in the visible science arm,
- three differential imagers: IRDIS, the infrared dual imaging spectrograph, IFS a near-IR integral field unit, and ZIMPOL, a visual (500 - 900 nm) high precision imaging polarimeter.

.

Figure 6.7: Block diagram for the SPHERE VLT "planet finder" instrument.

The SPHERE AO system. The SPHERE AO system is a so called extreme AO system for high Strehl ratio ($S = 0.85$ in H-band). A bright star, brighter than $R = 10^m$, is required for the AO guide star, in order to provide enough light for the very accurate wave front sensing. Key properties for the AO system are:

- Strehl ratio of $S \approx 0.5$ in the visible (600 nm) and $S \approx 0.85$ in the H-band.
- a good suppression of diffraction and halo stray light out to a radius of about $0.3''$ in the R-band and $0.5''$ in the H-band.

Technical properties of the individual AO components (Slide 6.9):

- fast, 1.2 kHz, tip-tilt mirror for the correction of the overall gradients in the wavefronts,
- fast, 1.2 kHz, deformable mirror with 41×41 actuator for the correction of the small scale wave front aberrations,
- a slow, 0.1 Hz movable pupil tilt mirror which corrects for slow pupil shifts due to the tracking by the telescope,
- a 10 Hz tip-tilt plate which corrects for the differential effects like atmospheric dispersion between the visible path of the WFS and the infrared science path,
- a 40×40 lenslet Shack-Hartmann wavefront sensor covering the wavelength range from 0.45 to 0.96 μm using a 240×240 pixel electron multiplying CCD which achieves a temporal sampling of > 1.2 kHz with a read-out noise smaller than 1 e^-,
- a differential tip-tilt camera which measures offsets of the IR-beam with respect to the visible beam.

Despite the very good AO correction the "typical" planet signal will still be fainter than the variable Speckle halo from the central star. For this reason differential measuring methods are required to compensate for the "speckle noise". Most dangerous are systematic instrumental speckles which drift in a uncontrolled way. For this reason the SPHERE system is optimized for stability and it will be located on the Nasmyth platform.

Slide 6.10 shows the different stages of the stellar point spread function (PSF) for SPHERE: the initial seeing PSF, the AO-corrected PSF-peak with a low and high contrast representation, and finally the coronagraphic PSF which serves as input for the differential imaging instruments.

ZIMPOL imaging polarimetry A very high precision is required with imaging polarimetry to detect a faint, point-like polarization signal of a planet in the residual seeing halo of a bright star. Aperture polarimetry reached already decades ago a very high precision better than 10^{-5} based on fast polarization modulators and a lock-in or photon counting detector.

A similar precision can be achieved with array detectors using the ZIMPOL technique. ZIMPOL (Zurich IMaging POLarimeter) uses also a fast polarization modulator and a special CCD camera performing the on-chip demodulation of the modulated signal. The fast modulator, e.g. a ferroelectric liquid crystal working at a modulation frequency of 1 kHz, and a polarizer converts the polarization signal into a fractional modulation of the intensity signal. This intensity modulation is converted back into a polarization signal by a special ZIMPOL CCD camera which measures for each active pixel the intensity difference between the two modulation states (Slide 3.40). For this every second row of the CCD is masked so that charge packages created in the unmasked row during one half of the modulation cycle are shifted for the second half of the cycle to the next masked row, which is used as temporary buffer storage (the CCD can be equipped with cylindrical micro-lenses which focus the light onto the open CCD rows). After many thousands of modulation periods the CCD is read out within about one second. The sum of the two images is proportional to the intensity while the normalized difference is the polarization degree of one Stokes component. Because the measurement is fully differential, systematic error sources are reduced to a very low level. Key advantages of this technique are:

- images for the two opposite polarization modes are recorded practically simultaneously (the modulation is faster than seeing variations),
- both images are recorded with the same pixels,
- there are only very small differential aberrations between the images for the two opposite polarizations due to the atmosphere or the telescope / instrument.

ZIMPOL is equipped, besides the polarimetric optics, with coronagraphs, exchangable filters, and calibration optics for the 520 nm – 880 nm range. The detector image scale provides one pixel per 7 mas, or about $140 \times 140 \approx 20000$ pixels per arcsec^2 for the search of point-like sources with a resolution of 15 mas at 600 nm.

Chapter 7

Planet formation

The observations indicate that planets form in circumstellar disks around young stars. Therefore planet formation is strongly linked to the star formation process which we describe in the introductory section of this chapter. In addition we discuss basic properties of interstellar and circumstellar material and then we treat circumstellar disks and the planet formation within disks.

7.1 Star formation

7.1.1 Components in the interstellar medium

Molecular clouds are a basic prerequisite for star formation because gravitational collapse of diffuse gas to stars occurs only in dense, cold clouds. Molecular clouds are located in the mid-plane of the Milky Way disk. The different types of gas components of the interstellar medium (ISM) are listed in Table 7.1.

Table 7.1: Components of the interstellar medium (ISM) in the Milky Way disc.

	T [K]	$N(\mathrm{H})[\mathrm{cm}^{-3}]$	gas type	main particles
1.	$10-100$	10^3-10^6	molecular clouds	H_2, dust, CO, ...
2.	$100-1000$	$\approx 1-10$	diffuse atomic gas	H^0, dust, C^+, e^-, N^0, O^0, ...
3.	≈ 10000	$10-10^4$	H II-regions	H^+, e^-, dust, X^{+i}, ...
4.	≈ 10000	≈ 0.1	diffuse, photo-ionized gas	H^+, e^-, dust, X^{+i}, ...
5.	$\gtrsim 10^6$	$\approx 10^{-3}$	diffuse, collisionally ionized gas	H^+, e^-, X^{+i}, ...

Pressure equilibrium. The diffuse cold (100 K), warm (10'000 K) and hot components (10^6 K) are in rough pressure equilibrium

$$p_{\mathrm{ISM}} = nkT \qquad \frac{p}{k} = nT \approx 1000\,[\mathrm{K/cm}^3]\,. \tag{7.1}$$

Cold, warm and hot phases. The interstellar gas exists predominantly in three temperature regimes:

- cold gas with $T < 1000$ K,
- warm gas with $T \approx 10'000 - 50'000$ K,
- hot gas with $T > 10^6$ K.

The existence of these three phases is due to the cooling function $\Lambda(T)$ for astrophysical gas. Gas radiates due to the conversion of thermal energy via particle collisions into radiation energy which escapes from the region. The cooling is proportional to the particle density squared, because two particles are required for a collision. The cooling rate is

$$C = n^2 \Lambda(T)$$

which is given e.g. in [erg/cm^{3}s] for the energy lost per unit volume and time. Most efficient **cooling mechanisms** are:

- emission of CO molecular lines for cold molecular clouds and far-IR atomic fine-structure lines from C II and O I for cold atomic gas,
- emission of nebular lines from ionized atoms, like O II, N II, S II, O III, C IV and others for warm gas,
- bremsstrahlung for the hot gas.

Figure 7.1: Specific cooling function and the predominant cooling and heating processes.

The heating of the gas can be due to very different processes. For certain temperature and density regimes there are the following predominant **heating processes**:

- collisions by cosmic rays, photodissociations of molecules, and gas turbulence for cold, dense gas, and photo-ionization for cold, diffuse (UV-transparent) gas,
- photo-ionization by stellar UV radiation for warm gas,
- adiabatic shocks from supersonic gas motions for hot gas.

Important heating processes per volume element are proportional to the particle density $\propto nH$. H is the heating function which depends on the gas temperature and density and the incident energy in the form of radiation or gas motions from outside the gas region. For a rough pressure and temperature equilibrium we can write:

$$nH = n^2\Lambda(T) \quad \text{and} \quad p = nkT$$

or with $n = H/\Lambda(T)$

$$\frac{\Lambda(T)}{T} = \frac{kH}{p}\,. \tag{7.2}$$

Temperature equilibria are reached for a broad range of heating conditions around $\approx 10 - 1000$ K for cold gas, $\approx 10'000 - 100'000$ K for warm gas and $> 10^6$ K for hot gas, because for these temperature regions the specific cooling rate $\Lambda(T)/T$ has a positive gradient (see Fig. 7.1).
A positive gradient is required, because if the heating and the gas temperature is slightly enhanced, then also the cooling must be more efficient, so that the gas temperature remains in an equilibrium. On the other hand the cooling must be less efficient for less heating and lower temperatures, so that the gas temperature remains close to the initial state.
If the gradient of $\Lambda(T)/T$ is negative, then the cooling is less efficient for a slightly enhanced heating and the gas starts to heat up. Contrary if the heating is slightly lower, then the cooling becomes more efficient and the gas starts to cool down.

Distribution of the different gas components. The diffuse atomic gas and the collisionally ionized hot gas (components 2 and 5 in Table 7.1) fill most of the space in the Milky Way disk, while the molecular gas and the cool atomic gas (components 1 and 2) make up about 90 % of the baryonic mass.

The molecular clouds and H II regions (components 2 and 3) are overdense regions in the interstellar medium. Molecular clouds are the locations where star formation can occur. They are localized in the mid-plane of the galactic disk, preferentially but not exclusively in the spiral arms. High mass stars $> 20\,M_\odot$, which are newly formed in molecular clouds, are very hot and they emit enough UV radiation to ionize their parent cloud. In this way the bright H II regions, like the Orion nebula, are formed. In external galaxies the H II regions trace often nicely the star forming regions located along the spiral arms.

7.1.2 Molecular clouds.

Molecular clouds are overdense regions in the Milky way disk made of molecular H_2, CO and dust predominantly. Because they are dense, their dust and gas is self-shielding the cloud from stellar optical and UV-light from the outside. Because of this, the molecular clouds can not be seen in the visual except for the fact that they obscure the object behind the cloud. The best way to see molecular clouds are CO line observations at $\lambda = 2.6$ mm in the radio range (see Slide 7.1). The dark irregular bands of absorption in the Milky Way are due to these absorbing clouds. The following types of molecular clouds are distinguished:

– Bok globules are small, isolated, gravitational bound molecular clouds of $\lesssim 100\,M_\odot$ in which at most a few stars are born,

- molecular clouds have masses of $10^3 - 10^4\,M_\odot$ distributed in irregular structures with dimensions of ≈ 10 pc consisting of clumps, filaments bubbles and containing usually hundreds of new-born stars,
- giant molecular clouds are just larger than normal molecular clouds with a total mass in the range $10^5 - 10^7\,M_\odot$, dimensions up to 100 pc, and thousands of young stars.

Molecular clouds in the solar neighborhood. The sun resides in a hot (10^6 K), low density bubble with a diameter of ≈ 50 pc. The nearest star forming clouds are located at about 140 pc and because of their proximity they are important regions for detailed investigations of the star and planet formation process (see slide 7.2). Well studied regions are:

- The **Taurus molecular cloud** at a distance of about 140 pc is a large, about 30 pc wide, loose association of many molecular cores with a total mass of about $\approx 10^4\,M_\odot$ and several hundred young stars. Because of its proximity there are many well known prototype objects, like T Tau or AB Aur in this star forming region.
- The ρ **Oph cloud** at a distance of 130 pc has a denser gas concentration than Taurus with a main core and several additional smaller clouds and about 500 young stars with an average age of about 0.2 Myr. The total gas mass is about $\approx 10^4\,M_\odot$.
- the **Orion molecular cloud** complex has a distance of about 400 pc and a diameter of 30 pc. Orion is the nearest high mass star forming region with in total about 10'000 young stars with an age less than 15 Myr. The Orion molecular cloud complex includes the Orion nebula M42 (H II region), reflection nebulae, dark nebulae (Horsehead nebula), an OB associations mainly located in the Belt and Sword of the Orion constellation. The Orion nebula is ionized by the brightest star in the Trapezium cluster (see Slide 7.3).

7.1.3 Elements of star formation

Stars form in dense, molecular clouds. If regions in clouds become dense enough then they may collapse and form under their own gravitational attraction a sphere which evolves into a star. The star formation process is very complex involving many different physical phenomena like the interaction of gas with radiation, hydrodynamics, magnetic fields, gas chemistry, dust grain evolution, gravitation and more.

Key parameters of the gas must be strongly changed for a transition from a cloud to a star:

- the density of a cloud must be enhanced from $\sim 10^{-20}\,\mathrm{g\,cm^{-3}}$ to about $1\,\mathrm{g\,cm^{-3}}$ in a star,
- the specific angular momentum (per unit mass) of the gas must be lowered from $\sim 10^{22}\,\mathrm{cm^2 s^{-1}}$ to about $\sim 10^{20}\,\mathrm{cm^2 s^{-1}}$ for a binary system or $\sim 10^{17}\,\mathrm{cm^2 s^{-1}}$ for a single star with a planetary system,
- and the magnetic energy per unit mass must be lowered from about $\sim 10^{11}\,\mathrm{erg\,g^{-1}}$ to about $\sim 10\,\mathrm{erg\,g^{-1}}$.

Thus, star formation means that the gas is strongly compressed by self-gravity, that it must loose essential all its angular momentum by fragmentation and magnetic breaking, and it must be strongly de-magnetized by processes like ambipolar diffusion.

Gravitational equilibrium and Jeans mass. Because molecular clouds live for long times there must exist, besides an equilibrium for the temperature and the pressure, also a hydrostatic equilibrium. The virial theorem is valid for systems in a gravitational equilibrium.

$$2\,E_{\rm kin} + E_{\rm pot} = 0 \tag{7.3}$$

If we consider a homogeneous (constant density) and isothermal cloud then we can write for the kinetic (or thermal) energy $E_{\rm kin} = E_{\rm therm} = 3kTM/2\mu$. This yields for the virial theorem:

$$2 \cdot \frac{3}{2}\,\frac{k}{\mu}\,T\,M - \frac{3}{5}\frac{G\,M^2}{R} = 0 \tag{7.4}$$

This can be rearranged into $kT/\mu = GM/5R$. The third power of this equation and inserting the mean density of a homogeneous sphere ($\rho = 3M/4\pi\,R^3$) provides an estimate for the equilibrium density or equilibrium mass for a given gas temperature T. These quantities are called **Jeans-mass**

$$M_{\rm J} = \Big(\frac{375}{4\pi}\Big)^{1/2} \Big(\frac{k}{G\,\mu}\,T\Big)^{3/2} \frac{1}{\rho^{1/2}}\,.$$

or **Jeans-density**

$$\rho_{\rm J} = \frac{375}{4\pi} \Big(\frac{k}{G\,\mu}\,T\Big)^{3} \frac{1}{M^2}$$

Example: The Jeans-density for $M = {\rm M}_\odot$, $T = 10$ K and $\mu = 2.7$ is $\rho_{\rm J} \approx 7 \cdot 10^{-19}\,{\rm g\,cm^{-3}}$ equivalent to a particle density (H_2) of $2 \cdot 10^5\,{\rm cm^{-3}}$.

The Jeans mass gives for a fixed cloud temperature and density the minimum mass required for being in gravitational equilibrium. The Jeans mass is smaller for cold, high density clouds. Similarly, the Jeans density describes for a given cloud mass and temperature the minimum density which must be achieved to be in a gravitational equilibrium. The density can be rather low for high mass, cool clouds.

The Jeans-density and Jeans-mass are parameters for an interstellar cloud in a gravitational equilibrium. However it is not clear whether this equilibrium state is stable or whether already a small disturbance yields a collapse to a star or an expansion and diffusion of the cloud.

For a **closed box model** the cloud remains in a hydrostatic equilibrium. If the cloud is slightly compressed then the liberated potential (or gravitational) energy is converted into thermal energy, which enhances the gas pressure and the system goes back into the equilibrium state.

Contraction by radiation. A contraction is possible if energy is radiated away. If contraction occurs then potential energy is converted into kinetic energy $\Delta E_{\rm kin} = -\Delta E_{\rm pot}$ and if part of this thermal (or kinetic) energy is radiated away then the system can find a more compact quasi-equilibrium configuration. According to the virial theorem half of the liberated potential energy must be radiated away, while the other half is converted into thermal energy

$$L_{\rm cloud} = -\frac{1}{2}\Delta E_{\rm pot} \quad \text{and} \quad \Delta E_{\rm kin} = -\frac{1}{2}\Delta E_{\rm pot}. \tag{7.5}$$

The contraction speed depends on the radiation or cooling time-scale:

$$\tau_{\rm cooling} \approx \frac{3}{2} nkT \, \frac{1}{n^2 \Lambda} = \frac{3}{2} \frac{kT}{n\Lambda} \, .$$

The cooling time scales becomes shorter during the collapse because the particle density increases steadily.

- the contraction is rapid in the **optically thin** case, because then the radiation can escape from the entire cloud volume,
- the contraction is slow if the cloud is **optically thick**, because the radiation can only escape from the surface.

The virial theorem requires that the cloud temperature raises during contraction if not radiation is emitted. Thus, contracting clouds heat up. But, because warmer gas emits more efficiently (as long as it is below $T < 1000$ K) for higher temperatures (see Fig. 7.1), the luminosity and therefore the loss of radiation energy of the contracting object becomes higher until the fast contraction changes into a slow quasi-static contraction when the cloud becomes optically thick.

Stabilization mechanisms must exist for self-gravitating clouds because else all existing clouds would collapse in a short timescale. Mechanisms which can stabilize a cloud against collapse are:

- cloud **heating** processes, like radiation from external stars, cosmic rays, magneto-hydrodynamic turbulence and waves, which enhance the gas temperature and the gas pressure so that the cloud expands,
- **angular momentum** conservations may inhibit collapse because of enhanced centrifugal forces for more compact and therefore more rapidly rotating clouds,
- **magnetic fields**, if there are ions in the molecular cloud so that the magnetic fields are frozen into the plasma and contraction enhances the magnetic pressure like $p_{\rm magn} \propto B_0^2 / r_{\rm cloud}^2$.

Star formation is complicated because so many different processes play a role and from observations it is often hard to get detailed information about the cloud geometry, heating processes, specific angular momentum, and magnetic properties of a gas. A few important aspects of star formation follow from the stabilizing processes discussed above.

Star formation feedback is the influence of new-born stars on their environment. Young stars have strong outflows and emit energetic radiation which both can heat the surrounding cloud and stop the star formation process. On the other hand, this heating produces over-pressurized bubbles, like the Orion nebula (Slide 4.3), which expand and which may compress the adjacent gas and trigger the collapse of a cloud. Depending on the details positive or negative feedback occurs and there is strong observational evidence that both mechanisms occur. However, many aspects of the star formation feedback are still unclear.

Fragmentation is linked to the Jeans mass. If a cloud contracts isothermally (loss of energy through radiation) then the density increases and the Jeans mass becomes smaller like $M_{\rm J} \propto 1/\sqrt{\rho}$. Thus, a large contracting cloud can decay in smaller clouds so that many stars are formed simultaneously in a big cloud complex. Typically there are many low mass stars formed $M < 1\,{\rm M}_\odot$ but only a few high mass stars $M > 1\,{\rm M}_\odot$.

.

Figure 7.2: Schematic illustration of the fragmentation process.

The specific angular momentum of the gas in a molecular cloud is very large when compared to a contracted proto-stellar clouds. Therefore the angular momentum barrier inhibits a global contraction of a cloud. However, if subunits can collapse into proto-stars then the global angular momentum with respect to the entire cloud is preserved as motion of the proto-stars around the center of gravity. The remaining specific angular momentum of the gas with respect to the individual proto-stellar cloud unit is then much smaller. The angular momentum barrier is a second important aspect in favor of cloud fragmentation and the quasi-simultaneous formation of many stars from big molecular cloud.

Proto-stellar disks and binaries are a further result of the angular momentum conservation. A contracting pre-stellar cloud core needs still to get rid of angular momentum. Angular momentum transfer via magneto-hydrodynamic processes helps to transport angular momentum away from the contracting cloud.

Another option is the formation of a binary star or a circumstellar disk. Both are configurations which can "store" more angular momentum than a rapidly rotating star.

Figure 7.3: Schematic illustration of ambipolar diffusion.

Ambipolar diffusion can solve the problem of the magnetic field pressure. A contracting cloud with charged particles contracts also the galactic magnetic field and will therefore "feel" soon the magnetic pressure which acts against further contraction. The magnetic field can move out of a neutral molecular clouds by the so-called ambipolar diffusion. This leaves in the end compact, demagnetized, cloud cores. The fact that stars form predominantly in dense, cool, neutral clouds could be due to the lack of magnetic pressure.

7.1.4 Initial mass function

Collapsing interstellar clouds form stars in the mass range from 0.1 to 100 $M_\odot$. The **initial mass function (IMF)** describes the mass distribution for the formed stars. According to the classical work of Salpeter (1955), this distribution can be described for stars of about solar mass and above with a potential law of the form:

$$\frac{dN_S}{dM} \propto M^{-2.35} \quad \text{for} \quad M > \mathrm{M}_\odot \,. \tag{7.6}$$

This relation is often given as a logarithmic power law of the form

$$\frac{dN_S}{d\log M} \propto M^{-1.35} \quad \text{because} \quad \frac{dN_S}{dM} = \frac{dN_S}{d\log M}\frac{d\log M}{dM} = \frac{1}{M}\frac{dN_S}{d\log M} \,.$$

This is equivalent to a linear fit with slope -1.35 in $\log M$-$\log N_S$ diagram (Figure 7.4). This law indicates, that the number of newly formed stars with a mass between 1 and 2 $M_\odot$ is about 20 times larger than the stars with masses between 10 and 20 $M_\odot$. If we consider the gas mass of the molecular cloud, then about twice as much gas ends up in stars between 1 and 2 $M_\odot$ when compared to stars with masses between 10 and 20 $M_\odot$. The initial mass function seems to be valid for many regions in the Universe, for the star formation in small molecular clouds, larger cloud complexes, and the largest star forming regions in the local Universe. Up to now no star forming regions have been found for which the Salpeter IMF is a bad description.

For low mass stars the mass distribution shows a turn over. Since M-stars $M < 0.5$ $M_\odot$ have a main-sequence life time which is longer than the age of the universe we can just use as first approximation the frequency of stars with different spectral types as rough description for the IMF of low mass stars (see also Table 2.4). This distribution shows a maximum in the range of M3V to M5V stars. Thus the mass distribution has a turn-over at a mass of about 0.4 $M_\odot$ followed by a rapid drop-off towards substellar objects. The minimum is around 0.01 $M_\odot$ or 10 M_J at the lower end of the brown dwarf regime. This is the so-called "brown dwarf desert". In the planet regime the frequency of object starts to raise again strongly towards lower masses at least down to M_E as demonstrated by the planet transit statistics from the Kepler satellite.

Figure 7.4: Schematic illustration of the initial mass function for stellar and substellar objects.

Considering the complex physics involved in the star formation process it is surprising that the initial mass function is such an universal law which seems to be valid everywhere in the Universe. There must be one essential process which dominates the outcome of the stellar mass distribution. This could be the fragmentation process. Further it seems to be clear that there are different regimes of formation between stars and planets. The low frequency of substellar object in the mass range 0.01 - 0.1 $M_\odot$ indicates that such objects are not easily formed via the normal star forming process, perhaps because the formation of small fragments or their survival in molecular clouds is rather unlikely. On the other side the planets are very frequent but seem to form predominantly around stars.

This indicates that there exists a bimodal formation mechanism of hydrostatic astronomical objects.

- stars are formed by the collapse and fragmentation of clouds,
- planets are the result of a formation process in circumstellar disks.

7.1.5 Types of proto-stars

There are different phases in the star formation process, from a collapsing cloud, to a pre-stellar core, to a proto-star, and a pre-main sequence star. Some of these phases have specific observational characteristics in the spectral energy distribution (SED). The SED show the signatures of the following components:

- the Planck-spectrum of the main energy source, with its characteristic peak flux wavelength for the temperature of the object,
- an infrared excess, if optical to near-IR radiation is absorbed by the circumstellar material and re-radiated at longer wavelength,
- an UV-visual excess because of energetic processes due to gas accretion onto the star,
- emission lines if the energetic processes are strong enough to dissociate and ionize gas.

Figure 7.5: Schematic illustration of the spectral energy distribution for the different types of young stellar objects.

According to the presence and characteristics of these features different types of young stellar objects are distinguished:

- **Class 0**: The SED peaks in the far-IR or sub-mm part of the spectrum near 100 μm (30 K), with no flux in the near-IR. These are the dense, pre-stellar cloud cores.
- **Class I**: They have a flat or rising SED from about 1 μm towards longer wavelengths indicating that a hot source ($\approx$ 1000 K) is still embedded in a cloud, so that most radiation from the young stellar object is absorbed and re-radiated as IR-emission by the circumstellar dust.
- **Class II**: They have falling SED into the mid-IR and the underlying objects have the characteristics of so-called classical T Tauri stars or Herbig Ae/Be stars. They exhibit strong emission lines and often a strong UV excess from the accretion process. These are the systems with extended circumstellar disks which are strongly irradiated by the central proto-star.
- **Class III**: These are pre-main-sequence stars with little or no excess in the IR, but with still some weak emission lines due to gas accretion. One of the subgroups of this class are the weak-lined T Tauri stars.

Class II and Class III objects can be placed into the Hertzsprung-Russell diagram if the temperature and luminosity are corrected for the contribution from the accretion processes. Compared to normal, main-sequence stars the Class II and Class III objects are located above the main sequence. These object evolve then "down" to the main-sequence (see Slide 7.4).

The **pre-main-sequence** time scale, which describes the quasi-static contraction of a young star follows from the Virial theorem. The Virial theorem requires that half of the potential energy gained by the gravitational contraction is radiated away as described by

$$\tau_{\rm KH} \approx \frac{E_{\rm pot}}{L} \approx \frac{G\,M^2}{R\,L}\,. \tag{7.7}$$

This time-scale is also called the Kelvin-Helmholtz timescale. For solar parameters there is $\tau_{\rm KH} \approx 30$ Myr. Pre-main sequence stars start as relatively large ≈ 3 $R_\odot$ and luminous objects ≈ 10 $L_\odot$ with correspondingly shorter time-scales.

7.2 Circumstellar disks

Before 1980 there existed only indirect evidence about the presence of circumstellar disk around young stars. In the 1980's collimated outflows, so-called jets, could be clearly associated with young stars (Slide 7.5). Jets are a well known phenomenon of accretion disks in active galactic nuclei and binaries were mass flows from one component through an accretion disk onto the companion.

Direct imaging of accretion disk became only possible with high resolution observations using HST. Edge-on disks could be imaged in nearby star forming regions and in Orion one could see the dark, light absorbing silhouettes of dusty disks in front of the extended nebular region (Slide 7.6 and 7.7). Thanks to coronagraphs on HST and ground based AO instruments with high contrast capabilities it is now possible to take images in scattered light for face-on proto-planetary disks. Some of the best images of such images were taken by our ETH group (Slide 7.8). Well resolved images of the gas in proto-planetary disks are now also possible in the mm-wavelength range using the new ALMA interferometer.

Another type of disk was discovered with the first far-IR satellites. It was recognized that many (10 %) A-stars exhibit an infrared excess. β Pic is the prototype of these stars and it shows a disk of dust, similar to the zodiacal dust disk in the solar system (Slide 7.9). The dust in these disk is due to colliding solid bodies (e.g. asteroid and meteoroids) or evaporating material from comets. Such disk are frequent around young stars and they are signposts of the early evolution of a planetary system.

All these data show, that disks are an important feature in the star and planet formation process.

7.2.1 Constraints on the proto-planetary disk of the solar system

We can use the solar system to infer some limits on the angular momentum and the mass of the circumstellar disk which formed "our" planetary system.

Angular momentum. The angular momentum budget of the solar system can be split into the angular momentum of the sun and the angular momentum of the planets.
The angular momentum of the Sun can be calculated as for a rotating, spherically symmetric sphere

$$L_S = \int_V (\vec{r} \times (\vec{\omega} \times \rho(r)\vec{r}))dV = k\, M_\odot\, R_\odot^2 \omega_\odot = 3 \cdot 10^{48} \mathrm{g\,cm^2\,s^{-1}}$$

where $k = 0.1$ accounts for the radial density distribution $\rho(r)$ of the sun ($k = 2/5$ for homogeneous sphere), and $\omega_\odot = 2.9 \cdot 10^{-6} \mathrm{s^{-1}} = 2\pi/25\,\mathrm{days}$ is the angular velocity of the solar rotation.
This can be compared to the angular momentum of Jupiter, which is

$$L_J = a_J \frac{1}{2} M_J v_J = M_J \sqrt{G M_\odot a_J} = 2 \cdot 10^{50} \mathrm{g\,cm^2\,s^{-1}}\,.$$

The total angular momentum of Jupiter is almost 100 times larger than for the sun. For the specific angular momentum the contrast is even more dramatic with the material in the planets having about 10^5 times more specific angular momentum than the gas in the sun. Thus there must exist an efficient process which can segregate angular momentum during the formation of the solar system.

Mass distribution. The sun contains more than 99 % of the mass of the solar system. We can deduce how much mass was present in the circumstellar disk for the formation of the planets. A rough budget for the gas mass and the mass of heavy elements is given in Table 7.2. This budget provides

- the mass of the heavy elements present now in the planetary system,
- an estimate on the total gas mass available in the pre-solar nebula for planet formation, if the dust to gas ratio was $m_d/m_g \approx 0.01$.
- a comparison with the gas mass and heavy element mass of the sun.

Table 7.2: "Minimum mass limit" for the circumstellar disk required for the formation of the planetary system.

total mass budget		heavy element (h.e.) budget	
mass of the sun	$2 \cdot 10^{33}$ g	1 % h.e. in the sun	$2 \cdot 10^{31}$ g
mass of giant planets	$3 \cdot 10^{30}$ g	10 % h.e. in giant planets	$3 \cdot 10^{29}$ g
mass of terr. bodies	$3 \cdot 10^{28}$ g	100 % h.e. in terr. bodies	$3 \cdot 10^{28}$ g
		h.e. in all planets	$3.3 \cdot 10^{29}$ g
mass required for planets for $m_g/m_d = 100$	$3.3 \cdot 10^{31}$ g		

Table 7.2 indicates that the heavy elements in the solar system are predominantly > 98 % located in the sun. The required gas and dust mass for the formation of the planets was about 1 % of the mass of the sun. This quantity if often called the "minimum mass solar nebula". Thus, the required reservoir of mass for the planet formation was small when compared to the mass required for the sun. A disk mass of about 1 % of the mass of the central star is in rough agreement with the mass estimates from observations of disks around young stars.

For our model we distribute the 0.01 $M_\odot$ minimum mass for the solar proto-planetary disk in rotationally symmetric uniformly annuli. In radial direction we adopt the distribution of the heavy element mass of the current planetary system times the gas to dust ratio m_g/m_d. This yields a disk with a surface density distribution of gas according to

$$\Sigma = 2 \cdot 10^3 \Big(\frac{r}{\mathrm{AU}}\Big)^{-3/2} \mathrm{g\,cm^{-2}}$$

extending roughly from the orbit of Venus to the orbit of Neptune. At a separation of 1 AU this disk would consist in vertical direction of about 20 g of dust per cm^2.

Disk description. For an initial discussion the description of circumstellar disks can be simplified with several reasonable assumptions:

- the disk is considered to be rotationally symmetric and symmetric with respect to the disk plane and can be described in an r-z-coordinate system,
- the vertical structure is given by the hydrostatic equilibrium

$$\frac{dP}{dz} = -\rho(r,z) g_z \tag{7.8}$$

– the gas in the disk is assumed to move in quasi-Keplerian orbits and the specific angular momentum of the disk gas is (using $\omega = \sqrt{GM_S/r^3}$)

$$\ell(r) = r^2\omega = \sqrt{GM_S r}\,.$$

– for **thin, low mass disks** the g_z gravity component is defined by the central star and there is ($\tan\theta = z/r \approx \sin\theta$)

$$g_z = \frac{GM_S}{r^2}\sin\theta = \frac{GM_S}{r^3}z = \omega^2 z$$

– if the vertical density structure for the thin disk is isothermal $P \propto \rho kT/\mu = \rho c_s^2$, where c_s is the sound speed then the hydrostatic equilibrium becomes:

$$\frac{d\rho}{dz} = -\rho\frac{\omega^2}{c_s^2}\,z$$

with the solution

$$\rho(r,z) = \rho_0 e^{-z^2/2h^2(r)}\,, \tag{7.9}$$

where $h(r) = c_s(r)/\omega(r)$ is the vertical scale height and $\rho_0 = \rho(r, z = 0)$.

– The surface density

$$\Sigma(r) = \int_{-\infty}^{+\infty}\rho(r,z)\,dz \tag{7.10}$$

is used to describe the radial structure of the disk.

7.2.2 Accretion disks

A disk is active, if it is transferring mass onto the star. This means that mass must flow inwards in radial direction. This is only possible if this gas loses potential energy and angular momentum by some processes. Potential energy can be converted into heat and radiated away. The angular momentum is either redistributed to the outer regions of the disk or lost by a rotational (magnetically confined) mass outflow.

Energy budget. The general energy budget for accretion disks is defined by the following quantities.

– the energy source for the accretion disk is the potential energy of the in-flowing gas:

$$\Delta E_{\rm pot} = \frac{GM_S\dot{M}}{R_S}\,. \tag{7.11}$$

– a stationary accretion disk is in a gravitational equilibrium state and the virial theorem is applicable

$$2\,E_{\rm kin} + E_{\rm pot} = 0\,.$$

– The virial theorem requires that half of the potential energy released by the in-flowing gas is transformed into kinetic energy $E_{\rm kin}$ which consists mainly of orbital motion but includes also the heating of the gas

$$\Delta E_{\rm kin} = \frac{1}{2}\Delta E_{\rm pot}\,,$$

while the other half of the potential energy of the in-flowing gas must be expelled from the system by radiation or energetic gas outflows. If radiation dominates then the luminosity of the disk is

$$L_{\rm disk} = \frac{1}{2}\Delta E_{\rm pot} = \frac{1}{2}\frac{GM_S\dot{M}}{R_S}\,. \tag{7.12}$$

– In the boundary layer between disk and star the gas in Keplerian motion must be decelerated to the surface velocity (rotation) of the star. In this process most of the orbital energy of the gas is released in complicated energetic processes. We may lump everything together into a term which stands for the boundary layer luminosity:

$$L_* = \Delta E_{\rm kin} = \frac{1}{2}\Delta E_{\rm pot} = \frac{1}{2}\frac{GM_S\dot{M}}{R_S}\,.$$

– Most of the energy of the accretion disk is released at the inner boundary, 75 % of the energy is released within $r < 2\ R_S$

With the accretion formula 7.11 the following accretion luminosities are obtained when considering both, the radiated energy from the disk and the energy released in the boundary layer.

$$L_{\rm acc}[L_\odot] = 0.35\frac{M_S[M_\odot]\,\dot{M}[10^{-8}M_\odot/{\rm yr}]}{R_S[R_\odot]}$$

An accretion disk may be called active if $L_{\rm acc} \gtrsim L_S$ and passive if $L_{\rm acc} \ll L_S$. In the first case the disk structure is defined by the accretion flow, while in the second case the disk structure depends strongly on the irradiation from the star.

Geometric structure of accretion disks. The continuity equation of an accretion disk can be written as:

$$r\frac{\partial\Sigma}{\partial t} + \frac{\partial}{\partial r}(r\Sigma v_r) = 0\,, \tag{7.13}$$

where v_r is the radial drift velocity with $v_r < 0$ describing inflows.

Similarly one can write the equation for the conservation of angular momentum:

$$r\frac{\partial}{\partial t}(\Sigma\cdot r^2\omega) + \frac{\partial}{\partial r}(r\Sigma v_r\cdot r^2\omega) = \frac{1}{2\pi}\frac{\partial G}{\partial r} \tag{7.14}$$

where G describes the torques (angular momentum transport) for example due to drag or viscosity effects described by ν. G is

$$G = 2\pi r\cdot\nu\Sigma r\,\frac{d\omega}{dr}\cdot r\,,$$

where $2\pi r$ is the circumference, $\nu\Sigma r\,d\omega/dr$ are the viscous forces and r the length of the lever arm.

Stationary accretion disk. For a stationary accretion disk the $\partial/\partial t$-terms in the continuity equation and the angular momentum equation are zero.

There will be a constant mass flow through the disk:

$$\dot{M} = \frac{dM}{dr} = -2\pi r \Sigma(r) v_r(r)\,. \tag{7.15}$$

Similarly the angular momentum transport will be time-independent

$$\Sigma r^3 \omega v_r = \nu \Sigma r^3 \frac{d\omega}{dr} + \text{const.}$$

The constant can be evaluated for a good guess of the boundary conditions. It seems reasonable to assume that the disk gas will be decelerated at the inner boundary $r_* \gtrsim R_S$ (see Fig. 7.6), so that there is a radius without differential rotation $d\omega/dr = 0$ (and shear forces).

Figure 7.6: Schematic illustration of the angular velocity of the accreting gas from the star to the outer disk region.

The resulting integration constant is (for $\nu\Sigma r^3\, d\omega/dr|_{r_*} = 0$)

$$\text{const.} \propto \Sigma r_*^3 \omega_* v_{r_*} = -\frac{\dot{M}}{2\pi} r_*^2 \omega_* = -\frac{\dot{M}}{2\pi} r_*^2 \sqrt{\frac{GM_S}{r_*^3}}\,.$$

With some algebraic transformation one obtains the radial surface density distribution

$$\nu\Sigma = \frac{\dot{M}}{3\pi}\Big(1 - \sqrt{\frac{r_*}{r}}\Big)\,. \tag{7.16}$$

Thus, the surface density distribution away from the boundary layer is defined by the microscopic "viscosity" parameter $\nu(r)$

$$\Sigma(r) \propto \frac{1}{\nu(r)}\,.$$

This indicates that the surface density for a given mass accretion rate is high for radii with low viscosity and the other way round. Thus, the disk-accretion process is self-regulating, because a higher density will for most conditions enhance the viscosity and as a result reduce the surface density.

Temperature and brightness structure of accretion disks. On can derive a disk temperature profile using the simple assumption that the loss in potential energy of the in-falling gas is converted locally (at a given r) into heat and radiated way according to

$$D(r) = \frac{1}{4\pi r}\, G\frac{d\omega}{dr} = \frac{9}{4}\nu\Sigma\omega^2$$

For a disk radiating like a blackbody $D(r) = \sigma T^4(r)$ and considering that the disk radiates on both sides gives the disk surface temperature distribution as function of radius:

$$T^4_{\rm disk}(r) = \frac{3GM_S}{8\pi\sigma}\frac{\dot{M}}{r^3}\Big(1 - \sqrt{\frac{r_*}{r}}\Big)\,. \tag{7.17}$$

It is important to note that for $r \gg r_*$ away from the boundary layer

- the surface temperature drops with radius like $T_{\rm disk}(r) \propto r^{-3/4}$ and the temperature profile does not depend on the viscosity ν,
- the disk brightness and surface brightness is proportional to the mass accretion rate $\dot{M}$.

Figure 7.7: Spectral energy distribution of an accretion disk.

Spectral energy distribution for accretion disks. The spectral energy distribution SED) of an accretion disk follows from the integration of the emission from the inner radius to the outer radius

$$F_\lambda \propto \int_{r_{\rm in}}^{r_{\rm out}} 2\pi r\, B_\lambda(T(r))\, dr\,. \tag{7.18}$$

At long wavelength $\lambda \gg hc/kT(r_{\rm out})$ the SED has the Rayleigh-Jeans slope of

$$\lambda F_\lambda \propto \lambda^{-3}\,.$$

At short wavelength there is the exponential drop defined by the hottest disk region at the inner boundary

$$\lambda F_\lambda \propto \frac{e^{-hc/\lambda kT(r_{\rm in})}}{\lambda^{-4}}\,.$$

The middle wavelength region can be evaluated by numerical integration of Eq. (7.18). There is also an elegant analytical solution which yields

$$\lambda F_\lambda \propto \lambda^{-4/3}\,.$$

The shape of the resulting SED is shown schematically in Fig. 7.7. The SED looks like a stretched black-body curve. Accreting circumstellar disks produce an infrared excess but with a declining SED in the IR spectral region.

Energetic processes in accretion disks. One should not forget that the SED of an accretion disk comes together with an equally strong emission $L_* \approx L_{\rm disk}$ from the boundary layer. Because the involved process are very energetic, the boundary layer emission is strong in the UV wavelength range, extending up to the X-ray regime, while the contribution to the IR is relatively small (small emission region). The following processes may play an important role (see also Fig. 7.8):

- magnetic fields from the central star are so strong that they disrupt the accretion disk at a distance of about 0.1 AU,
- this boundary layer is approximately at the distance, where the Keplerian angular velocity of the disk is equal to the stellar angular velocity,
- the gas from the disk will follow the magnetic field lines and falls through accretion columns onto the star,
- hot spots, emitting far-UV and X-ray radiation, are produced where the gas hits the stellar photosphere.
- The gas in the vicinity of the star and the innermost part of the accretion disk will be ionized by the energetic processes. Strong emission lines like Hα, Ca II and others are produced in this region.
- The gas is accelerated along the open field lines from the star and the disk in a rotating wind, which carries away a lot of angular momentum. The outflow will be collimated by the magnetic field at a radius where the velocity of the rotating magnetic field line approaches the speed of light. This produces the strong jets observed in many systems.

Figure 7.8: Schematic illustration of energetic processes taking place in accretion disks.

7.2.3 Passive circumstellar accretion disks

A disk is passive if its accretion luminosity is smaller than the irradiated energy from the star. This condition is valid for accretion rates

$$\dot{M} \lesssim 10^{-8} {\rm M}_\odot {\rm yr}^{-1}$$

Classical T Tauri stars have accretion rates which are about 0.1 to 10 times this value. Thus, the irradiation becomes dominant for the T Tauri stars with low accretion rates. Weak-line T Tauri stars are all in the regime were irradiation dominates.
The structure of an irradiated disk depends a lot on details as illustrated for a so-called transition disk (disk with inner hole) in Fig. 7.9. Important aspects are:

- The radiation from the star decreases like $\propto L_S/r^2$ due to geometric dilution,
- the irradiation at a given radius of the disk depends on the disk flaring, the disk height to radius factor, which can be written as:

$$\frac{h(r)}{r} \propto r^{\beta} .$$

  For $\beta = -1$ the disk is flat, for $\beta = 0$ it is an annulus with a fixed surface slope and for $\beta > 0$, the disk is flared.
- The disk opacity is dominated by the dust particles at the surface of the disk, which shield the mid-plane from radiation which is therefore colder than the surface,
- the dust in the innermost region of the disk (small r) may be evaporated by the energetic radiation from the star,
- regions of enhanced irradiation, e.g. the inner disk rim, have enhanced temperatures and expand vertically, creating shaded disk regions.

Figure 7.9: Schematic illustration of the disk irradiation in passive disks.

Imaging of proto-planetary disk is currently a research field in very rapid evolution. First images of the scattered light of the central star, scattered by the dust in the surface layer of the disk have been taken by our ETH group for separations as small as 0.1 arcsec from the star. This corresponds for objects in nearby star-forming regions to physical separations of about 15 AU, or close enough to see the expected planet-forming region. Indeed we see interesting structures, like inner holes, gaps, and spiral structures which could be due to young planets or planets in formation (see Slide 7.8). We have high expectations for the SPHERE/VLT instrument which should be the ideal apparatus for even sharper and deeper disk images.

Also new is the ALMA mm/sum-mm telescope which can trace the gas and the thermal emission of the cold dust for the same disks with similar resolution as the images shown in Slide 7.8. It will take only a few year until we have much improved knowledge about planet-forming disks.

7.3 Planet formation

The formation of planets occurs in the circumstellar disks around new-born stars. The formation of terrestrial planets, probably also for giant planets, requires growth of particles in a stepwise process from micron-sized dust grains to large planets with diameters up to 100'000 km. The mutual interaction between these particles and between the particles and the gas in the disk depends strongly on the particle size:

- **dust particles** with $r \ll$ cm are strongly coupled to the gas,
- **"rocks"** are objects on the meter scale and their dynamics is determined by gas drag and Keplerian orbits,
- **"planetesimals"** are bodies $\gtrsim$ 1 km up to $<$ 1000 km in radius which are essentially decoupled from the gas. Planetesimals are large enough to grow from smaller entities via gravitational attraction.
- **"terrestrial planets"** are big enough to collect all objects in there vicinity by gravitational attraction and they may become the dominant object in their orbital region.
- **"giant planets"** which may form if a planet core becomes large enough to accrete gas from the disk.

7.3.1 The formation of planetesimals

Aerodynamic drag forces. Aerodynamic drag forces are important for the description of the dust motion in a gaseous disk. The drag forces are given by

$$F_D = -C_D \cdot \pi a^2 \cdot \frac{1}{2}\rho v^2 \,, \tag{7.19}$$

where C_D is the drag coefficient, πa^2 the dust particle cross section, and $\rho v^2/2$ the kinetic energy (describing the transfer of momentum from the gas to the particle per unit time) of the gas moving with a velocity v relative to the particle.

Small particles: The drag coefficient C_D for small particles ($\lesssim$ 1 mm), which are smaller than the mean free path of the gas molecules, is defined by the mean thermal velocity $v_{\rm th} = (8/\pi)^{1/2}\, c_s$ of the gas particles:

$$C_D = \frac{8}{3}\,\frac{v_{\rm th}}{v}$$

and the resulting drag forces are:

$$F_D = -\frac{4\pi}{3}\, v_{\rm th} \cdot a^2\, \rho v = -v_{\rm th} \cdot \rho v \cdot \frac{m}{a\rho_d} \tag{7.20}$$

where we used the relation $m = (4/3)\pi a^3 \rho_d$ for the dust particle.

The friction time scale $t_{\rm fric}$ is a key quantity describing the interaction of a solid particles with the gas. The friction time scale indicates the time scale on which the aerodynamic drag leads to a change of order unity in the relative motion between a solid object and the gas

$$t_{\rm fric} = \frac{mv}{|F_D|} \,. \tag{7.21}$$

For small particles the friction time scale is

$$t_{\rm fric} = a\rho_d \cdot \frac{1}{v_{\rm th}\rho}\,, \tag{7.22}$$

where the first term is the radius and density of the dust particles which is divided by the kinetic velocity of the gas particles and the gas density.

Two quantities are not expected to change much from disk to disk: the dust particle density is of the order $\rho \approx {\rm g/cm^3}$ and the kinetic velocity of gas particles in the disk is of the order $v_{\rm th}^- \approx {\rm km/s}$ (≈ 100 K). Thus the friction time scale for small particles behaves like:

$$t_{\rm fric}[s] = \frac{a\,[\mu{\rm m}]}{\rho\,[10^{-9}{\rm g/cm^{-3}}]}\,\frac{\rho_d\,[1\ {\rm g/cm^{-3}}]}{v_{\rm th}\,[{\rm km/s}]}\,.$$

This formula indicates that the dust particle moves with the gas

- if the dust particles are small ($<$ mm),
- if the gas density is high $> 10^{-11}{\rm g/cm^{-3}}$].

A gas density of the order $10^{-9}{\rm g/cm^{-3}}$ is a reasonable value for the mid-plane of a propto-planetary disk. The friction time scale becomes long, $t_{\rm fric} \gtrsim$ years, for densities of about $10^{-12}{\rm g/cm^{-3}}$ and cm-sized particles.

Large particles: The drag coefficient C_D for larger particles is defined by the Stokes drag or the (molecular) viscosity ν in the gas. The Reynolds number

$${\rm Re} = \frac{2av}{\nu}$$

is a dimensionless parameter which describes the flow characteristics (e.g. ${\rm Re} < 1$ laminar flow) and $\nu \approx 10^{-5}{\rm cm^2\,s^{-1}}$ for gases. The drag coefficient can be expressed as a piecewise function:

- $C_D = 24\,{\rm Re}^{-1}$ for ${\rm Re} < 1$,
- $C_D = 24\,{\rm Re}^{-0.6}$ for $1 < {\rm Re} < 800$,
- $C_D = 0.44$ for ${\rm Re} > 800$.

Example: The Reynolds number of a rock with a radius > 10 cm will therefore always be larger than ${\rm Re} > 10^6$ for velocities larger than 1 m/s. Thus the drag force is given by Equation 7.19 with $C_D = 0.44$. The resulting friction time scale is:

$$t_{\rm fricti} = \frac{m}{0.44 \cdot \pi a^2 \cdot \frac{1}{2}\rho\, v} = \frac{18\rho_d a}{\rho v}$$

where $\rho_d \approx 3{\rm g/cm^3}$. The friction time scale for large particles written in convenient units is:

$$t_{\rm fric}[s] = 18 \cdot 10^6 \frac{a\,[{\rm m}]}{\rho\,[10^{-9}{\rm g/cm^{-3}}]}\,\frac{\rho_d\,[1\ {\rm g/cm^{-3}}]}{v\,[{\rm km/s}]}\,.$$

Thus, we obtain that the friction time scale of a rock moving with $v = 1$ km/s is of the order years. This rough calculation clearly shows that rocks do not couple well to the gas.

Dust settling. All particles above or below the disk mid-plane will be accelerated gravitationally towards $z = 0$ according to

$$|F_{\rm grav}| = mg_z = m\frac{GM_S}{r^3}z = m\omega^2 z\,.$$

The drag forces counteract, for small particles according to Equation 7.20

$$|F_D| = v_{\rm th}\cdot\rho v\cdot\frac{m}{a\rho_d}\,.$$

The settling speed follows from $|F_{\rm grav}| = |F_D|$

$$v_{\rm settle} = \Big(\frac{\omega^2}{v_{\rm th}}\Big)\frac{\rho_d}{\rho}a\,z\,.$$

For $a = 10$ μm size particles the settling speed is of the order cm/s only and the settling time from a height $z = 1$ AU becomes

$$t_{\rm settle} = \frac{|z|}{v_{\rm settle}} \approx 10^4{\rm yr}\,.$$

These equations and numbers indicate the following basic aspects of dust settling:

- micron-sized and smaller particles cannot settle in a gaseous disk during the expected disk lifetime,
- if there are turbulent motions then also particles in the range $r \approx 1-100\,\mu$m do not settle,
- mm and cm - sized particles are the ideal particles for settling in the mid-plane. They will move with a speed of about 1 m/s towards the mid-plane within about 100 years.
- m-sized rocks do not couple to the gas and they will oscillate around the mid-plane for quite some time.

Current planet formation models indicate that the settling of the mm-sized particles to the disk mid-plane is an important first step for the efficient particle growths towards planetesimals.

Coagulation. Small particles must grow by coagulation. It is clear that this process can only be efficient if the particle density is high. The best place for coagulation is the disk mid-plane when some dust has settled. Larger bodies may move faster than the gas in the disk (see below). In this case fast growth is possible because faster moving small grains may stick to larger bodies like for hailstones in a ice/water cloud in the Earths atmosphere. This is just speculation, as we have almost no information about this process in proto-planetary disks. According to current ideas it seems:

- that small icy dust grain can stick together easily if they collide with small relative velocities forming in this way particles larger than 1 cm.
- object may grow fast to sizes beyond 1 m if they are immersed in a flow of gas full of small icy particles, which may build up a large "ice/snow" ball.

It is clear that coagulation is an important step for the formation of planetesimals. However we have no secure data about how this process works in proto-planetary disks.

Planetesimal growth by collisions. The collisions of two solid bodies can be divided into two main categories:

- **Collisional growth** of the larger body through accretion of a small body where most of the mass of the impactor becomes part of the final body. One may distinguish between impacts which leave the solid structure of the large body essentially unchanged and impacts which break up (shatter) the target object, which then reassembles gravitationally like a rubble pile. In both cases small fragments may be lost but overall there is a net growth.
- **Collisional erosion** of a body by the impact of smaller body occurs if the impactor is just deflected and some fragments of the initial bodies are also lost, so that there is a net reduction of the mass. One may speak of a **catastrophic destruction** of the target into many small objects if the largest fragment is smaller than half the mass of the initial target.

It is clear that impacts of relatively small objects with small velocities are ideal for the growth of planetesimals, while impacts of fast and large objects tend to lead to the erosion or even destruction of a body.

For small objects ($\lesssim 10$ m) growth or destruction depends a lot on the material strength. The material strength for rocks are about 10 times higher than for icy objects. Thus icy objects can be destroyed more easily in a catastrophic event. On the other side, they are the better targets for "absorbing" low energy impactors. Rocks may just undergo a collision which deflects the impactor while losing some fragments and some mass.

For larger objects $\gtrsim 1$ km the growth or destruction depends on the gravitation. An object can only be destroyed if the kinetic energy of the impactor (with velocity v and mass m) is of the order or larger than the gravitational energy of the target with mass M and radius R:

$$\frac{1}{2}mv^2 \gtrsim \frac{GM^2}{R}\,.$$

This relation indicates:

- it is much harder to destroy a high mass objects for a given mass ratio m/M and relative velocity v,
- for high mass objects the typical mass ratio between impactor and target is smaller, and therefore the probability for destruction decreases rapidly with mass.

Radial drifts in a gas disk. In the discussion on accretion disks we made the simplification that the gas is moving essentially with Keplerian speed v_K on Keplerian orbits: This is not fully the case for a gas disk, because the gas particles feels a positive gas pressure gradient with radius $dP/dr > 0$. This is equivalent to a reduced gravitational acceleration and therefore the azimuthal gas velocity is slower than the Keplerian velocity.

$$v_{\rm gas} = v_K(1-\epsilon)\,.$$

Depending on the particle size this effect has the following consequences:

- Dust particles < 1 cm are strongly coupled to the gas and they move therefore together with the gas and there is no differential radial drift between dust and gas.

- objects in the sizes equal or larger than 1 m "do not feel" the gas pressure gradient and they move therefore with Keplerian velocity v_K which is faster than the gas. Therefore they encounter a head-wind. This produces for ≈ 1 m sized objects a substantial radial drift towards the star. The effect on the tangential velocity seems to be small $\epsilon < 0.01$, but it is strong enough to induced an inward drift of > 10 m/s. Therefore the typical accretion time scale for meter-sized objects is very short, of the order 1000 years or less.
- Larger bodies > 100 m will have a slower radial drift because the friction time scales is proportional to the size of the objects $t_{\rm frict} \propto a$. Therefore large bodies of > 100 m are not rapidly removed from the disk but they may still migrate slowly through the disk.

According to this the radial drift is largest and most critical for meter-sized bodies. The growth of bodies from cm-size to km-size must be fast and efficient to overcome the removal of solid bodies from the disk.

7.3.2 Formation of terrestrial planets

Once planetesimals have formed in large numbers there must be a strong evolution towards a few large planet-sized objects which dominate the system. For example, if the Earth formed from 5 km sized objects then it is composed of $\approx 10^9$ of these. This means, that the transformation from planetesimals to proto-planets must be an efficient process.

We can calculate the collision rate, the number of collisions per unit time, of a large body with radius R passing through a region with a high density n of small objects moving with typical velocity v.

$$\frac{1}{t_{\rm coll}} = \pi R^2 \, nv \, .$$

The collision rate increases with density (of course) and with the relative velocity v. This formula does not take into account the gravitational focussing effect which can enhance the collision rate for small v.

Figure 7.10: Schematic illustration of the gravitational focussing in the rest frame of the more massive object.

Gravitational focussing. Gravitational focussing is a process which helps to increase the accumulation of mass onto a large object. We consider the angular momentum conservation and the energy conservation of the potential impactor with respect to the large

matter collecting object (see Fig. 7.10). The initial angular momentum and energy of a potential impactor are:

$$L = bv \quad \text{and} \quad E = \frac{1}{2}mv^2\,,$$

where b is the impact parameters and m the mass of the small object. This must be equal to the angular momentum and energy at closest approach or minimum separation r_0 with velocity v_0:

$$L_0 = r_0 v_0 \quad \text{and} \quad E_0 = \frac{1}{2}mv_0^2 - \frac{GMm}{r_0}\,.$$

Because the angular momentum is conserved the minimum separation is $r_0 = bv/v_0$ (or $v_0 = bv/r_0$) and with the energy conservation $E = E_0$ we obtain:

$$\frac{1}{2}mv^2 = \frac{1}{2}\frac{b^2}{r_0^2}mv^2 - \frac{GMm}{r_0}\,.$$

We can now replace the minimum separation with the radius R of the accreting large object. A potential impactor will hit the planetesimal if the impact parameter is smaller than

$$b^2 = R^2\left(1 + \frac{2GM}{Rv^2}\right) = R^2\left(1 + \frac{v_{\rm esc}^2}{v^2}\right). \tag{7.23}$$

This indicates that the cross section for collecting smaller objects by gravitational attraction is significantly enhanced with respect to the object size πR^2, if the relative velocities of the potential impactors are small, say smaller than the escape velocity $v_{\rm esc}$.

For example the escape velocity for Earth is 11 km/s. Thus Earth is capable to sweep up all objects with a relative velocity less than 100 m/s out to an impact parameter of $b \approx R_E v_{\rm esc}/v \approx 100 R_E$ which is beyond the orbit of the moon. For Pluto, which has a mass of only 0.2 M_E, the escape velocity is less, about 1 km/s. However, one can expect a smaller velocity dispersion so far out in the solar system and therefore the average impact parameter has a similar value as for Earth.

Oligarchic growth. Due to the gravitational focussing the more massive object grow faster than small objects. After some time only a few object, the so-called "oligarchs", dominate because they have swept up a large fraction of objects along their orbits. The time scale of this process depends on many details. Important aspects are:

- If orbital eccentricities of the proto-planet and the planetesimals are small then the relative velocities are small and the gravitational collection of bodies by a proto-planet is very efficient. However, the proptoplanet can then only collect bodies along a relatively narrow circular strip and its growth is limited by this.
- If orbital eccentricities of the proto-planet or the planetesimals are larger then the relative velocities will be higher and the gravitational accumulation will be less efficient. On the other side, the proto-planet can sweep up much more planetesimals in a wider radial range in the disk and grow to a bigger planet in the end.
- A particularly favorable case can occur, if there is still some gas in the disk which introduces a migration of the proto-planet and the planetesimals. Because the radial drift speed depends on mass (smaller planetesimals move faster in), the collection of planetesimals by a planet may be particularly efficient.

The growth of proto-planets is expected to be fast in the beginning and after 10^5 to 10^6 years the initial many billions of small km-sized planetesimals will be reduced to the order of millions with about 100 dominating proto-planets. It may then take 10^7 or 10^8 years to reduce further the population of planetesimals and the stabilization into a planetary system with 5 to 15 planets. Slide 7.10 shows some snap-shots of a simulation for the formation of proto-planets from planetesimals.

The proto-planet growth is much faster for short periods objects because it takes certainly more than 10 to 100 orbits to collect the planetesimals along a circular orbit. Since the orbital time scales for large separation planets $a > 30$ AU is about 2 orders of magnitude longer the growths "out there" is expected to be correspondingly slower.

Of course, if a strong gravitational instability event between two giant planets takes place then the outcome will not be a system with 5-15 planets on quasi-circular or low eccentricity orbits but only an outer and inner giant planet on eccentric orbits.

If the mass in the planetesimals becomes smaller than the mass in the proto-planets then the oligarchic growth will stop because there is simply no significant mass left for further growth.

7.3.3 Gas giant formation

Two theoretical models are put forward for the giant planet formation, the core accretion model and the disk instability model.

The core accretion model assumes that the core of a giant planet forms like a terrestrial planet as outlined in the previous section. The key requirements for growth beyond a planet core are:

- the core must become massive enough to hold a bound atmosphere of hydrogen and helium gas,
- the core must form fast, before all gas in the disk has been accreted or is dispersed.

These requirement imply that the formation of giant planets via core accretion is most effective at disk radii where a lot of material can be collected and where the time scale for accretion is short. An attractive region are disks regions with a separation of about 10 AU around solar mass stars due to the following reasons:

- the ice-line, which is at a separation where water-ice particles in the disk start to evaporate, is around $r_{\rm ice} \approx 3 - 5$ AU for solar type stars. Inside $r_{\rm ice}$ there will be much less solid material available and the particles are probably not "sticky". This is not ideal for the fast formation of planetary cores.
- The available mass for planet formation decreases for smaller radii in the disk. This is true despite the fact that the surface density is expected to increase with smaller radius $\propto 1/r^{3/2}$ while the mass per unit annulus width behaves like $\Delta r \propto 1/r^{1/2}$. But, at small separation the size of the gravitational potential well of a proto-planet with given mass is smaller $R_{\rm Hill} \propto r$ so that less mass is available for the planet formation.
- At large disk radii $r > 30$ AU the time scale for the accumulation of planetesimals in a proto-planet becomes too long because this time scale is proportional to the orbital period.

A strong argument in favor of the core accretion model is the strong positive correlation between planet occurrence rate and host star metallicity obtained by the radial velocity planet surveys. A high metallicity in the disk means a lot of solid material and therefore a fast planet core formation by the accretion of planetesimals.

The disk instability model is based on the assumption that the disk can collapse locally under its own gravitation and form a compact gas sphere which evolves into a giant planet. The process is not much different from the formation of stars in a gas cloud. The condition that a gravitational instability occurs in a disk is often described by the Toomre Q parameter (introduced for galactic disks) which must be small:

$$Q = \frac{c_s \omega}{\pi G \Sigma} < Q_{\rm crit} \approx 1 \, .$$

One should note that the sound speed c_s gives essentially the gas temperature $c_s = (kT/\mu)^{1/2}$. A small Q parameter, a necessary condition for collapse is possible for the following cases:

- if the surface density Σ of the disk is high (= high disk self-gravity),
- if the temperature (sound-speed) and therefore the gas pressure is small, what can be achieved if the disk is optically thin for the emission of thermal radiation,
- if the angular velocity is small, because this is equivalent to a small specific angular momentum with respect to the center of the collapsing disk region.

Models for giant planet formation via disk instabilities indicate that this process could happen in massive disks at large separations > 30 AU from the central star. The recent detection of the 4 giant planets in the system HR 8799 at separations of 15, 24, 38 and 68 AU strongly supports the disk instability model. Core accretion is expected to be not efficient enough for the formation of at least the outermost two planets.