Detection and Characterization of Extrasolar Planets

A. Quirrenbach

1 Methods of Planet Detection

1.1 The Quest for Planets Around Other Stars

The realization that our Sun is just one "average" star amongst billions and billions in the Sky naturally brings with it the question whether some – or perhaps most – of the other stars may also harbor planetary systems like our own. We live in a remarkable epoch, being the first generation that has obtained an affirmative answer to this question, that is undertaking programs to characterize the physical properties of planets outside the Solar System, and that is developing the tools to search for twins of the Earth. For the first time in human history, we are on the verge of being able to address the questions whether there are other habitable worlds, and to search for life elsewhere in the Universe with scientific methods.

The search for extrasolar planets has a long and checkered history (see e.g., Boss 1998a for an easily readable overview). Because of the enormous brightness contrast between planets and their parent stars, the direct detection of planets by taking images of the vicinity of nearby stars would be extremely difficult. Early searches for planets were therefore mostly carried out with the astrometric method, which seeks to detect the motion of the star around the center of mass of the star–planet system (see Sect. 9). First reports on the detection of massive planets ($\sim 10\,M_{\rm jup}$) were published during World War II (Strand 1943; Reuyl and Holmberg 1943), but remained controversial, both with regards to the reality of the results and to the question whether the detected bodies should be called "planets". Much painstaking work over the next few decades lead to the realization that these "detections" were spurious. Continued improvements in the astrometric accuracy finally culminated in the announcement of a planet 1.6 times as massive as Jupiter in a 24-year orbit around Barnard's Star (van de Kamp 1963). A decade earlier Otto Struve had written a remarkable paper, in which he noted the possibility that Jupiter-like planets might exist in orbits as small as 0.02 AU, proposed to search for these

Quirrenbach A (2006), Detection and characterization of extrasolar planets. In: Mayor M, Queloz D, Udry S and Benz W (eds) Extrasolar planets. Saas-Fee Adv Courses vol 31, pp 1–242

DOI 10.1007/3-540-29216-0_1

objects with high-precision radial-velocity measurements, and pointed out the feasibility of photometric searches for planets eclipsing their parent stars – all on little more than one journal page (Struve 1952).

By the mid-sixties, the search for extrasolar planets thus appeared to be a thriving field, with eight planetary companions known from astrometric observations (two of them classified as "existence not completely established"), and a number of potentially promising alternative search methods under consideration (O'Leary 1966). By the same time it had also been recognized that brown dwarfs (termed "black dwarfs" at the time) would form a class of their own, with properties intermediate between those of stars and planets. Both astrometric searches for brown-dwarf companions to low-luminosity stars, and attempts at finding them directly with high-resolution imaging techniques, seemed to be successful (Harrington et al. 1983; McCarthy et al. 1985).

Sadly, none of these early claims for detections of planets and brown dwarfs withstood the test of time. It turned out that systematic instrumental errors had been mistaken for the "planetary companion" of Barnard's Star (Gatewood and Eichhorn 1973). What appeared to be the most convincing detection of a brown dwarf, a companion to the star VB 8, could never be confirmed (Perrier and Mariotti 1987; Skrutskie et al. 1987). Other putative planets and brown dwarfs did not fare better. By the mid-nineties, all that remained was a candidate brown dwarf companion of HD 114762, detected with the radial-velocity method (Latham et al. 1989).[1]

This situation changed completely and abruptly with the discovery of 51 Peg b, a Jupiter-like planet in a 4-day orbit (Mayor and Queloz 1995), which has opened a completely new field of astronomy: the study of extrasolar planetary systems. About 150 planets outside our own Solar System are known to date, and new discoveries are announced almost every month. These developments have revolutionized our view of our own place in the Universe. We know now that other planetary systems can have a structure that is completely different from that of the Solar System, and we have set out to explore their properties and diversity.

The following chapters introduce the most important methods that have been employed (or proposed) for the detection of extrasolar planets, and for studies of their physical characteristics. Emphasis is given to observational techniques, their foundations, limitations, and their practical implementation. As far as possible, published results are mentioned in the context of the respective observing techniques, and some outstanding implications for the astrophysics of planets and planetary systems are discussed. This will hopefully elucidate the capabilities, strengths, and weaknesses of the many

[1] The radial-velocity technique does not allow measuring the companion mass, but only $m \sin i$, where i is the unknown inclination of the orbit (see Sect. 4). It could therefore not be excluded that HD 114762 B is a low-mass star in a nearly face-on orbit.

complementary observational approaches. It should be kept in mind, however, that the study of extrasolar planets is a rapidly expanding field, in which new and unanticipated results appear almost every month. Technical developments in fields such as adaptive optics, coronography, and interferometry are also occurring at a staggering pace. Nevertheless, the systematic introduction of the fundamental principles and methods attempted in this article will hopefully remain a useful guide for a while to come.

1.2 What is a Planet?

The Definition of "Planet"

Before we can begin to answer the question how planets outside the Solar System can be detected and characterized, we must first agree on an operational definition of the term "planet". The Greek root of the word literally means "unsteady" or "transient"; it was historically applied to the five known "wandering stars" Mercury, Venus, Mars, Jupiter, and Saturn. The Copernican Revolution added the Earth to the list, and the discoveries of Neptune, Uranus, and Pluto completed the census of the large bodies in the Solar System as we know it. The example of Pluto clearly demonstrates the need for a clean definition of the term "planet". With the discovery of a large number of bodies belonging to the Kuiper Belt (Jewitt and Luu 1993; Luu and Jewitt 2002) it has become clear that Pluto is but the largest member of the class of Trans-Neptunian Objects (TNOs). It has therefore be argued that Pluto should be demoted from its rank among the planets. I would side with the majority view, however, that the use of the term "planet" in the Solar System is based on historical developments and should not be changed retroactively.

The history in our own Solar System thus shows that the use of the term "planet" has been expanded from the original five members of this class, to newly discovered objects that shared the most important properties of the established examples. Two of these additions (Neptune and Uranus) were rather undramatic, one was based on the realization that the Earth shared important properties with the planets (it orbits the Sun between Venus and Mars), and one added a physically distinct and different body to the list (Pluto). Progress in our knowledge about the planets has also taught us that our list includes bodies encompassing wide ranges in mass, composition, and other physical characteristics.

When we look outside our Solar System, we should certainly expect to find a variety of objects that share many characteristics with our planets, but that may be different in one or more important ways. It is thus a matter of definition what we call a "planet" and where we draw the boundaries to other classes of objects. From a practical point of view, this definition should be based on properties that are easily verifiable observationally; this favors

a definition based on mass over a definition based on the formation history. Nonetheless, we should not expect that we can easily come up with a set of criteria that will in each case allow an unambiguous classification of a newly discovered as a "planet" (or not). For example, if a maximum mass is included among the defining properties, all objects discovered with the radial-velocity technique – and thus with known $m \sin i$, see Sect. 4.1 – could strictly speaking only be called "planet candidates" before additional information on their orbital inclination is secured.

For the purposes of this article, I take a "planet" to be an object that fulfills the following criteria:

- *A planet is an object in orbit around a star or a multiple star system.* This excludes free-floating planet-mass objects. A number of such objects have been detected with direct-imaging surveys in young clusters (e.g., Zapatero Osorio et al. 2000, 2002; Béjar et al. 2001; Lucas et al. 2001).[2] Free-floating objects are not considered further here, although it is possible that some of them originally formed in a circumstellar disk, and were ejected by a collision with another planet (e.g., Bryden 2001).
- *A planet is not in orbit around another planet.* This requirement excludes moons, but one should point out that the distinction between moons and planets is also somewhat fuzzy. For example, the Pluto–Charon systems could be called a double planet rather than a planet with a moon.
- *A planet has a minimum mass of* 10^{22} kg. This distinguishes planets from planetesimals, asteroids, and comets.[3]
- *A planet has a maximum mass of* $13\,M_{\rm jup}$. This sets the boundary between planets and brown dwarfs. The value of $13\,M_{\rm jup}$ has been chosen to roughly coincide with the Deuterium burning limit (e.g., Burrows et al. 1997a). This criterion will often be applied fairly loosely, as objects with $m \sin i < 13\,M_{\rm jup}$ will also be called "planets" even if there is no additional information on i.

As with the word "planet", we will use other terms established in the Solar System and and apply them to analogous bodies and material around other stars; we can thus speak of "moons" and "rings" around extrasolar planets, about "exo-planetesimals", "exo-comets" and "exo-zodiacal dust".

The Thermal Evolution of Giant Planets

A basic understanding of the evolution of planets is an important prerequisite for a discussion of detection methods. The fundamental principle is rather

[2] Note that a tentative detection of free-floating planet-mass objects in the cluster M 22 (Sahu et al. 2001) has been retracted (Sahu et al. 2002).

[3] The value of 10^{22} kg is quite arbitrary, of course. I have chosen it because it separates Pluto from the "minor bodies" in the Solar System.

simple: planets are born hot and are initially self-luminous; they cool during their evolution until they reach radiative equilibrium with their parent stars. The age of a planet is therefore an important parameter that determines how difficult it is to detect its thermal emission.

The luminosity evolution of giant planets, alongside with that of brown dwarfs and low-mass stars, is shown quantitatively in Fig. 1; it can be seen from this figure that an old planet is about four orders of magnitude(!) fainter than it was at an age of 1 Myr. Another important conclusion is that luminosity alone is an extremely poor indicator of the mass of substellar objects – information about the age is crucial to distinguish between low-mass stars, brown dwarfs, and planets. (Dynamically determined masses are even better, of course.) We see, however, that at the same age the luminosity of gaseous planets increases strongly with mass. Figure 1 thus provides a very useful first orientation for general considerations about the detectability of giant planets. For more detailed calculations, one has to take into account the planet's temperature, which determines the spectral energy distribution, and modifications to the luminosity evolution due to irradiation by the host star (e.g., Burrows et al. 2003).

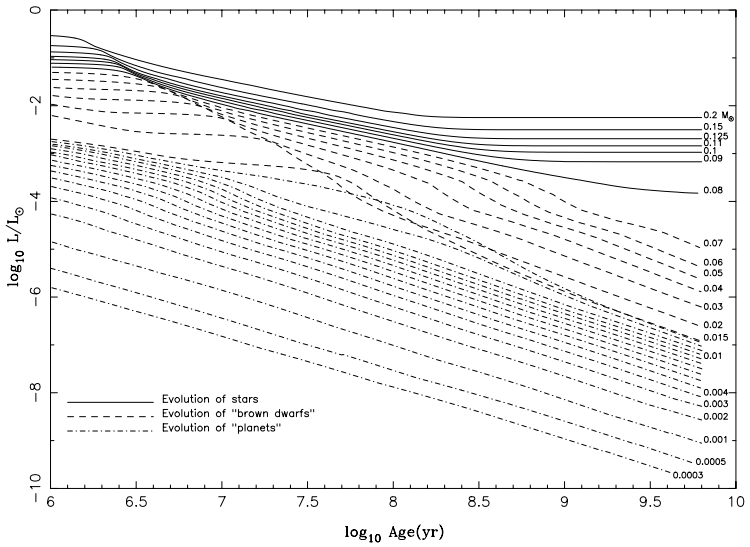

Fig. 1. Evolution of the luminosity (in $L_\odot$) of solar-metallicity M dwarfs and substellar objects vs. time (in yr) after formation. The stars, "brown dwarfs" and "planets" are shown as solid, dashed, and dot-dashed curves, respectively. In this figure, we arbitrarily designate as "brown dwarfs" those objects that burn deuterium, while we designate those that do not as "planets." The masses (in $M_\odot$) label most of the curves, with the lowest three corresponding to the mass of Saturn, half the mass of Jupiter, and the mass of Jupiter. From Burrows et al. (1997a)

1.3 Pulsar Planets

The First Extrasolar Planets

While the considerations of the preceding sections appear to give a solid framework for planet searches, the first firm discovery of objects that fulfill the above definition of a planet came totally unexpected and from a completely different line of research. The extremely stable rotation of pulsars provides a high-precision clock, which can be used for the indirect detection of planets, in a way that is quite similar to the radial-velocity method that will be discussed in detail below (Sect. 4). High-precision monitoring of the time-of-arrival (TOA) of the radio pulses can reveal subtle motions of the pulsar, such as its reflex motion due to the presence of a planetary companion. For a planet with mass m_p in a circular orbit with period P and inclination i, and a "canonical" neutron star mass of $1.35\,M_\odot$, the amplitude of the timing residuals τ is

$$\tau = 1.2\,\mathrm{ms}\left(\frac{m_p}{M_\oplus}\right)\left(\frac{P}{1\,\mathrm{yr}}\right)^{2/3}\sin i\ . \tag{1}$$

For millisecond pulsars, TOA measurements are possible with a long-term precision of a few μs (e.g., Wolszczan 1994). This implies that planets down to $\sim 0.01\,M_\oplus$ are detectable around pulsars; this limit is far lower than that of any other search method currently contemplated.

After a few false starts (e.g., Bailes et al. 1991; Lyne and Bailes 1992), two planets just a factor of ~ 3 more massive than the Earth were found orbiting the pulsar PSR B1257+12 (Wolszczan and Frail 1992). The two planets are in a 3:2 orbital resonance, which leads to accurately predictable periodic perturbations of the two orbits. The detection of this mutual gravitational attraction between the planets provided the final proof of the reality of the first pulsar planets; the same data set also revealed the presence of a third planet with even lower mass in the same system (Wolszczan 1994). The properties of the planets orbiting PSR B1257+12 are listed in Table 1.

Table 1. Parameters of the PSR B1257+12 planetary system

planet	A	B	C
semi-major axis [light-ms]	0.0035	1.3106	1.4121
eccentricity e	0.0	0.0182	0.0264
orbital period [days]	25.34	66.54	98.22
longitude of periastron	–	249°	106°
planet mass $[M_\oplus]$	$0.015/\sin i_1$	$3.4/\sin i_2$	$2.8/\sin i_3$
distance from pulsar [AU]	0.19	0.36	0.47

After Wolszczan (1999)

The Keplerian timing residuals (1) depend on $m_p \sin i$; this means that the mass of the planet and its orbital inclination cannot be determined independently. In contrast, the strength of the mutual interaction between the planets depends directly on their masses. It has thus become possible to infer the masses and inclinations of planets B and C from modeling of a long series of timing data, which now covers more than a decade (Konacki and Wolszczan 2003). The derived masses are $4.3 \pm 0.2\, M_\oplus$ and $3.9 \pm 0.2\, M_\oplus$, respectively, and the orbital inclinations $53^\circ \pm 4^\circ$ and $47^\circ \pm 3^\circ$ (or 127° and 133°), indicating that the two orbits are nearly co-planar.

Even after taking the three planets and the interaction between planets B and C into account, there remains a long-term systematic variation of the TOA residuals (see the lower two panels of Fig. 2). These residuals could be indicative of the presence of a fourth planet with longer orbital period

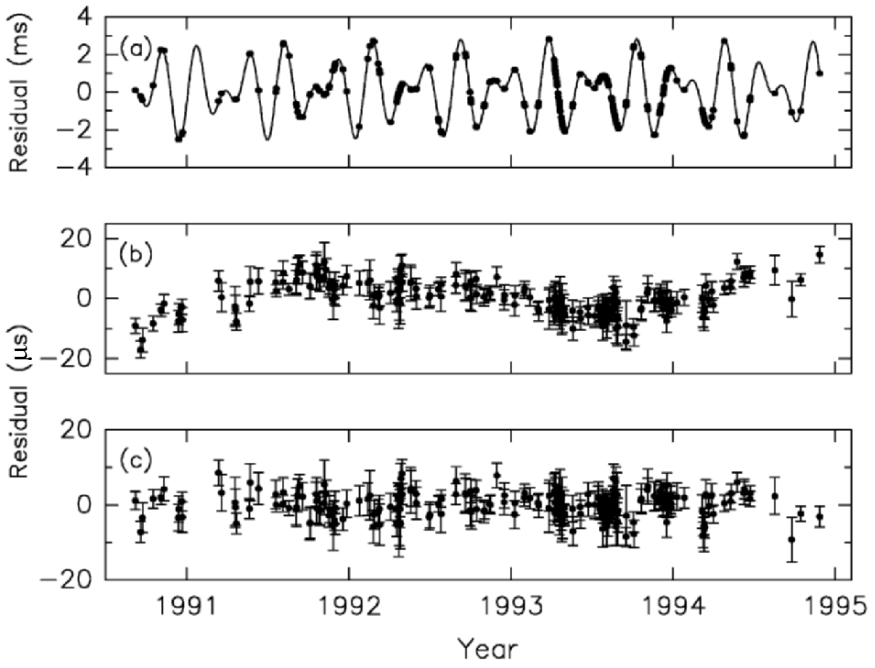

Fig. 2. Timing residuals for PSR B1257+12 at 430 MHz, for three increasingly detailed models. (**a**) Residuals after the fit of the standard timing model without planets. The time-of-arrival variations are dominated by the Keplerian orbital effects from planets B and C. (**b**) Residuals for the model including the Keplerian orbits of planets A, B, and C. Residual variations are determined by gravitational perturbations between planets B and C. (**c**) Residuals for the model including all the standard pulsar parameters, and the Keplerian and non-Keplerian orbital effects. From Konacki and Wolszczan (2003)

(Wolszczan et al. 2000). If the apparent three-year periodicity of the residual signal can be confirmed, this would point to an origin of the disturbance within the pulsar planetary system itself. It will probably be difficult to ascertain the nature of this ionized material – a "coma" ablated from a fourth body or a warped disk are among the possibilities.

Pulsar planets appear to be rare. Only one other pulsar, PSR B1620−26 near the core of the globular cluster M4, has a confirmed planet (Arzoumanian et al. 1996; Joshi and Rasio 1997). The B1620−26 system is rather interesting, too. The planet orbits an inner binary system, which consists of a millisecond pulsar and a white dwarf companion in a half-year orbit. The most likely mass of the planet is $m_p \sin i_p \approx 7\,M_{\rm jup}$, and the semi-major axis and eccentricity of its orbit are $a \approx 60\,{\rm AU}$ and $e \approx 0.45$ (Thorsett et al. 1999; Ford et al. 2000).

The Formation of Pulsar Planets

The theories for the formation of pulsar planets can be broadly divided into two classes: (a) scenarios, in which the planets were formed together with a "normal" star, and survived its evolution from the main-sequence to become a red giant and later a rapidly spinning neutron star; and (b) scenarios in which the formation of the neutron star precedes the formation or acquisition of its planets (Podsiadlowski 1993). The first category implies that the planets must be able to survive the formation of the pulsar, which involves a violent transformation in a supernova explosion, and the supernova recoil. This possibility is generally regarded as unlikely, and scenarios of type (b) are favored.

Consequently, most theories of planet formation around millisecond pulsars concern themselves with possible ways to disrupt, evaporate, ablate, or otherwise dismember the companion star, and thus to transform a fraction of the companion's mass into a gaseous disk around the neutron star (Phinney and Hansen 1993). Such a disk could be formed, for example, by an asymmetric supernova explosion in a binary system, which kicks the neutron star into its companion. In this picture, a high-velocity single neutron star with a planetary system is created from the remains of the former binary companion. One could thus speculate that the presence of planets around PSR B1257+12 is related to its unusually high proper motion.

Neutron star disks are clearly very different from those commonly found around pre-main-sequence stars. The disk is exposed to intense radiation and particle flux, close to or even above the Eddington luminosity of the neutron star ($\sim 10^{38}\,{\rm erg\,s^{-1}}$). The metallicity is very high, but initially there are no grains, and the temperature is well above the sublimation temperature of even the most refractory materials. The disk must therefore expand and cool before planets can be formed. Calculations of the evolution of such disks indicate that the formation of "terrestrial" planets such as those of the PSR B1257+12 system may indeed by possible, but the more massive and distant planet

around PSR B1620−26 must have a different origin (Phinney and Hansen 1993).

The location of PSR B1620−26 near the core of the globular cluster M4 suggests that the pulsar acquired its planetary companion through an exchange interaction with a cluster star (Sigurdsson 1992, 1995). One plausible formation scenario begins with an old neutron star in a binary system, which interacts with a main-sequence star–planet system (Sigurdsson 1993). The original companion of the neutron star is ejected, while the main-sequence star and its planet are captured. The planet ends up in a wide orbit around the inner binary comprised of the neutron star and the main-sequence star. When the main-sequence star evolves to become a red giant, it transfers mass to the neutron star, spinning it up to become a millisecond pulsar. The chief difficulty of this scenario is the requirement that the age of the millisecond pulsar must be smaller than that of the triple system. However, the expected lifetime of the triple in the dense cluster core is of order $3 \cdot 10^7$ yr, while the estimated age of the binary pulsar is $\gtrsim 10^9$ yr (Ford et al. 2000). This scenario would thus require that the system, currently observed in projection near the edge of the cluster core, is in fact on an orbit that allows it to spend most of its lifetime in the far less dense cluster halo, and thus to escape disruption for a sufficiently long time.

An alternative formation scenario involves a dynamical exchange interaction between a pre-existing binary millisecond pulsar and a wide main-sequence star–planet system, in which the main-sequence star is ejected and the planet left in a wide orbit around the binary pulsar (Ford et al. 2000). Numerical simulations show that the probability of retaining the planet in the encounter is smaller than that of retaining the main-sequence star, but could still be as high as 10%...30%. It is interesting to note that this scenario postulates the formation of a giant planet in a wide orbit around a "normal" star in a globular cluster environment; this is to be contrasted with the apparent absence of "hot Jupiters" in 47 Tucanae (see Sect. 6.4).

1.4 Overview of Planet Detection Methods

The Most Important Detection Techniques

Turning our attention to "normal" stars again, we will now look at the question how we might be able to find planets around them. Many different techniques have been proposed, in spite (or perhaps: because) of the difficulty of the task. The most promising strategies that are used in current detection efforts, or under development for use in the near future, are:

- Direct imaging of the star–planet system.
- Interferometric imaging of the star–planet system.
- Detection of the planetary spectrum in a composite spectrum of star and planet.

- Interferometric detection of the planetary spectrum through the wavelength dependence of the position of the photocenter of the star–planet system ("differential phase method").
- Photometry of planetary transits in front of the star.
- Spectroscopic detection of planetary transits.
- Photometric detection of the light reflected by a planet through its periodic variation with phase angle.
- Astrometric detection of the stellar motion around the star–planet center of mass.
- Radial-velocity measurement of the stellar motion around the star–planet center of mass.
- Imaging of circumstellar disks, which may show signatures of disk–planet interaction.
- Gravitational microlensing.
- Eclipse timing in binaries.

Each one of these techniques has unique strengths and weaknesses, and they vary widely in the information they can provide about the properties of the detected planets. Turning the question the other way around, we can take a list of characteristics that we would like to know about extrasolar planets, and ask which techniques can provide the requested information. Table 2 gives an overview of the most important planetary properties, and how they may be determined. More detailed discussions about the strengths and limitations of individual methods will be given in the subsequent sections. For the moment, the most important observation is that no single approach can give all the desired information; many complementary techniques will be needed to study the different aspects of extrasolar planets and planetary systems.

Typical Order-of-Magnitude Estimates

In order to understand the instrumental and observational requirements for the different planet detection techniques, we need to consider typical values for the potential observables. The large range in the properties of planets obviously implies a large difference in the difficulty to detect them. The Earth and Jupiter provide useful benchmarks (see Table 3), but one should also keep in mind that there are additional classes of planets, e.g., Uranus and Neptune in the Solar System, or the "hot Jupiters" orbiting their central stars at very small orbital radii. The properties of the host star, and the distance from the observer play important roles, too.

The chief difficulty of direct detection methods is the large contrast between the planet and its parent star at a very small angular separation. The reflex motion of the parent star due to the gravitational pull of the planet is very small, so that astrometry and the radial-velocity technique must reach extremely high precision to detect this effect. The photometric signature of transiting planets seems somewhat more easily accessible, at least for giant

Table 2. Important properties of planets, and techniques that can be used to determine them

property	technique	applicability
orbit	astrometry	++
	radial velocity	+
	direct imaging	∘
mass	astrometry	++
	radial velocity	+
	microlensing	∘
radius	transit photometry	++
radius, albedo	photometry of reflected light	++
radius, temperature	direct detection in mid-IR	++
surface features	photometry of reflected light	+
atmospheric composition	IR or visible spectroscopy	++
	transit spectroscopy	∘
presence of moons	transit timing	+
system multiplicity	astrometry	+
	radial velocity	+

The symbols ++, +, and ∘ denote how well the different methods can provide the required information

planets. However, in this case additional complications arise from the small probability that the orientation is such that transits actually occur. It is thus clear from the values listed in Table 3 that there is no "easy" technique for planet detection – this is the reason, of course, why it was not before 1995 that the first planet around a main-sequence star was discovered.

1.5 "Exotic" Concepts for Planet Detection

The subsequent chapters of this review will be devoted to introductions to some of the most promising planet detection techniques. Many more interesting approaches have been proposed, which deserve at least a brief description. It is entirely possible, of course, that one or the other of these "exotic" concepts will turn out to be more fruitful than some of the techniques that are considered "mainstream" today. The variety of physical effects that could in principle be observable should illustrate the diverse opportunities for the immediate and more distant future, and stimulate further ideas about possible ways to obtain more detailed information about planets during various phases of their life cycles, and about their interaction with the host stars.

Table 3. Typical values of observables for Jupiter-like and Earth-like planets

observable	Jupiter	Earth
angular separation	$0\farcs5$	$0\farcs1$
brightness contrast at visible $\lambda\lambda$	6×10^{-7}	1.5×10^{-10}
brightness contrast at 10 µm	1.5×10^{-7}	1.2×10^{-7}
astrometric amplitude	500 µas	0.3 µas
radial-velocity amplitude	$13\,\mathrm{m\,s^{-1}}$	$0.1\,\mathrm{m\,s^{-1}}$
transit probability	10^{-3}	5×10^{-3}
transit depth	1%	10^{-4}
transit duration	30 h	13 h
timing residuals	2.5 s	1.5 ms

The host star is assumed to be a Sun twin at a distance of 10 pc

Radio Emission from Extrasolar Planets

Five of the planets in the Solar System (Earth, Jupiter, Saturn, Uranus, and Neptune) produce non-thermal cyclotron radio emission, in a process that is thought to be driven by the Solar wind interacting with the planetary magnetospheres. The emission frequency is typically near the electron gyrofrequency in the magnetic field, i.e., of order 30 kHz . . . 30 MHz. The emission is very intense (at times, Jupiter is brighter than the Sun at frequencies below 20 MHz), and there exist fairly simple scaling laws that relate the observed radio power to the ram pressure of the Solar wind on the cross-sectional area of the magnetosphere (Zarka et al. 2001). There also exist scaling laws for the magnetic dipole moment of giant planets (Farrell et al. 1999); these scaling laws together can be used to predict the emitted radio power and peak frequency. In a few favorable cases the emission should be observable with current instruments, but no detections have been made so far (e.g., Bastian et al. 2000). This may either be due to the intermittent nature of the cyclotron emission, or to a smaller velocity or density of the stellar wind compared to the Solar wind, or to a smaller magnetic moment of the planet, due perhaps to tidal synchronization. In any case, future low-frequency arrays such as LOFAR or a Square Kilometer Array (Strom et al. 2001) should be able to observe the radio emission from magnetized giant planets.

Interaction-Induced Stellar Activity

Tidal or magnetic interaction between a giant planet in a short-period orbit and its host star might also increase the stellar activity, which could lead to variations in the shape of chromospheric lines in phase with the orbital period. Hints of systematic modulations of the Ca II H and K lines have been

found in a few such systems, but they need further confirmation (Cuntz and Shkolnik 2002). It has also been speculated that very strong flares observed in some Solar-type stars could be due to magnetic reconnection between fields of the primary star and a close-in Jupiter-like planet (Rubenstein and Schaefer 2000). A systematic search for unusual flaring stars might thus be a new way of looking for planets, but a better understanding of the planet–star interaction would clearly be needed.

Young Planets Heated by Giant Impacts

The formation of planets proceeds through a phase of giant impacts (e.g., Wetherill 1990). The largest of these impacts may melt all or most of the surface of an Earth-size planetary embryo, and heat it to a temperature of about $1,500 \ldots 2,500\,\mathrm{K}$. This would make its thermal emission detectable with a large ground-based interferometer (Stern 1994). Giant impacts on giant planets (such as the event that may have tipped the rotation axis of Uranus) will likely heat them to similar temperatures, making them even more easily detectable. The cooling times are of order a few hundred to several thousand years; this should make the number of impact-heated objects at any given time large enough to expect one detection per every few hundred pre-main-sequence stars surveyed. One would then still have to establish the planetary nature of the detected object, of course, and distinguish it from more massive companions or other possible interlopers.

Planets Swallowed by Giant Stars

As a Solar-mass star evolves off the main sequence, it expands and ascends the red giant branch of the Hertzsprung–Russell diagram. During that evolutionary phase, the star develops a large convective envelope with a radius of up to $\sim 100\,R_\odot$. Planets within that radius will be accreted by the star, and thus deposit energy, angular momentum, and elements such as Lithium in the stellar envelope. It has therefore been argued that the infrared excess (due to a substantial expansion of the star and ejection of a shell) and high Li abundance observed in $\approx 5\%$ of the G and K giants could be caused by the accretion of a giant planet or a brown dwarf (Gratton and D'Antona 1989; Siess and Livio 1999).

In an even later evolutionary stage, when the star becomes an asymptotic giant branch (AGB) star, it swells to an even larger size and develops a strong wind. Planets with even larger orbital radii can then get engulfed in the extended atmosphere, or interact with the wind flow. Episodic accretion of wind material on the planet may give rise to optical flashes and affect SiO maser emission (Struck et al. 2002). The details of these interactions are quite complex and poorly understood at the moment; this limits their potential use as a diagnostic tool and indicator for the presence of planets.

Planets Around White Dwarfs

It should be clear from the preceding paragraph that the orbits of planets can change drastically during the star's post-main-sequence evolution. Low-mass companions will spiral into the star due to the viscous and tidal forces exerted by the bloated atmosphere during the giant phase, but it might also be possible that some of them may be left in an orbit of radius $a \lesssim 1\,\mathrm{AU}$ around the ensuing white dwarf. This would be a very favorable situation for detecting the planet, because its radius would be ~ 10 times larger than that of the parent star! In the Rayleigh–Jeans portion of the combined spectrum (i.e., for observations at wavelengths much longer than that corresponding to the peak of the Planck function), the ratio of the total emission to that of just the white dwarf is given by

$$\frac{I_{\mathrm{tot}}}{I_{\mathrm{WD}}} = 1 + \frac{R_p^2 T_p}{R_{\mathrm{WD}}^2 T_{\mathrm{WD}}} \approx 1 + 100\,\frac{T_p}{T_{\mathrm{WD}}}\,. \tag{2}$$

For example, a planet with $T_p = 200\,\mathrm{K}$ orbiting a white dwarf with $T_{\mathrm{WD}} = 10,000\,\mathrm{K}$ would dominate the total emission of the system at long wavelengths.

Several groups have conducted near-infrared searches for substellar companions of white dwarfs, and some low-mass companions have been reported, but no planet has been discovered (e.g., Zuckerman and Becklin 1992). The above argument suggests, however, that searches should be conducted in the mid-infrared, where the planet can produce a strong excess over the white dwarf spectrum (Ignace 2001). The Spitzer (formerly SIRTF) infrared mission should have sufficient sensitivity to detect such planets out to a distance of $\sim 10\,\mathrm{pc}$.

Occultations by the Moon or Artificial Satellites

The planet detection schemes discussed in the previous few paragraphs intend to take advantage of special situations in which the the signature of the planet is not swamped by the nearby bright host star. In the general case, one may try to address the contrast problem by blocking the light from the star. This could either be achieved by using the dark limb of the Moon as an occulting edge (Elliot 1978), or by building a spacecraft carrying an occulting screen (Schultz et al. 1999, 2000; Copi and Starkman 2000). In either case, the observations would be carried out with a space telescope, which has to maintain alignment with the occulter to a precision of a fraction of an arcsecond (the typical angular separation of the planet from its parent star). The main obstacles for Lunar occultations are the rather large brightness of even the dark side of the Moon, and the difficulties of maneuvering the telescope. While it is possible to find orbits that give rather long ($\sim 1\,\mathrm{h}$) occultations of arbitrary stars, an enormous amount of propellant would have to be used to change targets.

Artificial occulters face similar problems. The diameter of the occulting screen clearly has to be larger than the telescope aperture, i.e., at least $\sim 10\,\mathrm{m}$.

To subtend an angle of no more than $0\farcs1$, it must therefore be placed at a separation of at least 20,000 km from the telescope. Furthermore, the intensity of the starlight in the shadow of the occulter is not zero; it must be computed with Fresnel's diffraction theory (e.g., Born and Wolf 1997).[4] Application of Babinet's Principle gives the approximation (Schultz et al. 1999)

$$\frac{I}{I_0} \approx \frac{16}{\pi^2} \cdot \frac{\lambda a}{D^2} = \frac{16}{\pi^2} \cdot \frac{\lambda}{\varphi D} \ , \tag{3}$$

where I and I_0 are the intensity in the presence and in absence of the occulting disk, λ the observing wavelength, a the distance between the occulter and the telescope, D the diameter of the occulter, and φ the angle subtended by the occulter as seen by the telescope. Equation (3) is valid for $D^2 \gg \lambda a$. For the above numbers ($D = 10\,\mathrm{m}$, $a = 20,000\,\mathrm{km}$) and $\lambda = 500\,\mathrm{nm}$, the on-axis intensity is still 16% of the value in the absence of an occulter. This shows that diffraction at the edge is a serious problem. With λ and φ fixed, and clear limits on the potential to increase D and a, the only viable way of obtaining a better starlight suppression is the use of a tapered occulter, i.e., a screen which is not completely opaque but has a transmission that continuously increases from 0 at the center to 1 at the edge (Copi and Starkman 2000). Manufacturing such a screen with precisely prescribed transmission function is a considerable technological challenge. This, together with the requirement of maneuvering the screen and telescope very precisely, has so far prevented serious consideration of this approach for a planet-detection mission.

A variation of the occultation idea is the use of a coronograph, which includes an occulting spot in the focal plane of the telescope. Compared to an external occulter, a coronograph has the disadvantage that the starlight is blocked only after passage through the telescope optics. The telescope therefore has to be built to very stringent specifications on wavefront quality and light scattering level. Nevertheless, this approach is currently regarded more promising than that of an external occulting screen.

1.6 The Search for Extraterrestrial Intelligence

The Drake Equation

Speculations about the possibility of life, of conscious beings, and of civilizations elsewhere in the Universe have a long history (Dick 1982, 1998). The search for extraterrestrial intelligence (SETI) as a scientific endeavor was born with the realization that our own technology had advanced to the point that radio signals could be transmitted and detected over interstellar distances

[4] One may recall Poisson's famous bright spot. Using Fresnel's theory, Poisson – who was very critical of that theory – predicted the seemingly absurd appearance of a bright spot behind a circular obstruction. This spot was almost immediately found experimentally by Arago, a great triumph of the wave theory of light.

(Cocconi and Morrison 1959). Soon the question was raised how many civilizations in the Galaxy might be engaged in attempts at communicating with each other, leading to the formulation of the famous *Drake Equation* (Drake 1962)

$$N = R_* \cdot f_p \cdot n_h \cdot f_l \cdot f_i \cdot f_c \cdot L\,. \tag{4}$$

The individual factors in this equation have the following meanings:

- N: the number of communicating civilizations in Galaxy;
- R_*: the rate of star formation in the Galaxy (expressed in stars per year);
- f_p: the fraction of stars that harbor planetary systems;
- n_h: the average number of planets (or moons) with conditions that are suitable for the genesis of life;
- f_l: the fraction of habitable planets on which life actually develops;
- f_i: the probability that evolution produces intelligent life;
- f_c: the fraction of intelligent civilizations that try to communicate over interstellar distances;
- L: the length of the communication phase (in years).

Unlike the other equations in this book, which (hopefully!) quantify our insights and knowledge, it is the main purpose of the Drake equation to organize our ignorance. We know that $R_* \approx 1\,\mathrm{yr}^{-1}$ (Trimble 1999), and we can now state fairly confidently that $f_p \geq 0.01$.[5] The determination of n_h is one of the great observational challenges for the coming ten to twenty years, as described extensively in this overview. With some luck, we might even be able to obtain an estimate for f_l from astronomical observations. This factor may also be amenable to biochemical experimentation in the tradition of the famous Miller–Urey experiments (Miller 1953) and modern attempts to generate synthetic life forms (Szostak et al. 2001). At present, in the absence of any evidence for extraterrestrial life, we have to admit that f_l could be anywhere between 10^{-9} and 1.

The next factor, f_i, is equally uncertain. Biologists are deeply divided about the question whether life necessarily evolves towards intelligence once it gets going. On the one hand, one may point out that a staggeringly improbable series of events has lead to the emergence of intelligent life on Earth (Gould 1989); on the other hand, one can argue that convergence is a ubiquitous property of life, which makes it likely that particular biological properties and features will sooner or later manifest themselves as part of the evolutionary process (Conway Morris 1998). In addition, we do not understand the biological basis of intelligence at all. What is the "quantum leap" that separates *homo sapiens* from *pan troglodytes*, the chimpanzee? Would *homo neanderthalensis* have become capable of constructing radio telescopes, if he

[5] This estimate is based on the number of planets detected in the Solar neighborhood (Sect. 3.1), with a "safety factor" applied for the possibility that the efficiency of planet formation may vary with the Galactic environment.

hadn't been displaced by a more advanced species? Finding an answer to these questions seems to be a key step towards a better estimate of f_i.

The factors f_c and L fall into the realm of sociology. It is tempting to speculate that $f_c \approx 1$, given the human drive for exploration, but we do not know with certainty that this extrapolation from our anthropocentric view of the world is really justified. The value of L depends on external factors such as global epidemics and giant impacts by comets and asteroids, and on internal factors that could lead to a quick end of a "semi-intelligent" civilization – wars or the exhaustion of natural resources. It appears possible that our own species and our offspring may populate the Earth at least for the remainder of the main-sequence lifetime of the Sun ($L \approx 5 \cdot 10^9\,\mathrm{yr}$), but if we are not careful, we may not live to see $L_{\mathrm{homo}} = 100\,\mathrm{yr}$.

We may thus characterize the emergent fields of exo-planetary astronomy and astrobiology as attempts to systematically explore the individual factors of the Drake Equation, from the left to the right. In contrast, SETI (which should perhaps better be called "Search for Extraterrestrial Technology" or "Search for Interstellar Communication") is an attempt to bypass this painstaking process by going directly for the grand prize. The chances of success are very uncertain, as the above arguments are consistent with estimates that range from an average distance between "neighbors" of $\approx 30\,\mathrm{pc}$, to a Galaxy that is void of life save that on a lonely, solitary Earth.

The Fermi Paradox

If civilizations are common in the Galaxy, one may ask the question why we have not found any incontrovertible evidence for their existence yet. More to the point, it has been argued that the absence of extraterrestrials from the Solar System implies that we are alone in the Galaxy, and that any searches for extraterrestrial civilizations are futile. The chain of arguments, which is known as *Fermi's Paradox*, goes as follows:

1. Lets's assume that our civilization is not the only one in our Galaxy that has developed technology.
2. Then our civilization must be "typical". This means that it is not the most advanced of all, and that other civilizations share our desire to explore.
3. Space travel is not too difficult for civilizations "slightly" more advanced than ours.
4. The time scale to colonize the whole Galaxy is $\lesssim 10^8\,\mathrm{yrs}$, i.e., small compared to the age of the Galaxy.
5. Then one must conclude that the Solar System should have been colonized a long time ago. But this is not the case.

So it appears that we have encountered a logical difficulty if we believe in the ubiquity of life. However, each step in this chain has potential loopholes, some of them more severe, others less.

The assumption in step (1.) can certainly be questioned. As explained in the previous section, we know very little about f_i, the likelihood for intelligence to emerge through evolution. If this factor is small, we may indeed be alone in the Galaxy.

Step (2.) seems to be quite plausible. The development of intelligence on Earth may have been a singular event, but if we assume that it has occurred in *one* other place, it very likely occurred in *many* other places. Then it is very unlikely that we are the most advanced civilization, given that there are many Solar-type stars (i.e., stars with comparable mass and metallicity) that are several Gyrs older. Life on a terrestrial planet around any of these stars would have a several-Gyr head start compared to the Earth. And assuming that *none* of these earlier civilizations would be interested in exploring the Galaxy (or that *all* of them would refrain from doing so for ethical reasons) seems extremely unlikely, too.

To justify step (3.), we can invoke some physical considerations. Several methods of attaining speeds necessary for interstellar travel ($v \gtrsim 0.1c$) have been suggested, including pulsed fusion and antimatter-powered rockets, light sails pushed by lasers, and interstellar ram jets (Crawford 1990). The biggest hurdle to overcome for interstellar travel are the enormous energy requirements; accelerating a spaceship to a substantial fraction of the speed of light in a reasonable time would require a few times the current global power production. This is a staggering power requirement, but it is plausible that it could be met by humanity very soon. If our power production grows at an average rate of only $\sim 1\%\,\mathrm{yr}^{-1}$, it will take less than 1,000 years, and the power requirement for interstellar travel will only be a fraction of a per cent of the global power consumption.[6] For comparison, a Saturn V rocket during lift-off consumed $\sim 0.5\%$ of the global power production. It thus seems likely that a civilization that is only slightly more advanced than ours will have the technical means to travel through interstellar distances.

The argument in step (4.) is based on the assumption that civilizations will establish colonies, and that each colony will again establish sub-colonies once it gets firmly established. With reasonable assumptions about the mean distance between these colonies, and about the time it takes for a colony to establish itself and to spawn a new settlement, one can estimate that it takes only a few Myrs to reach every habitable planet in the Galaxy (Crawford 2000).

Step (5.) also contains an assertion that can be questioned. While we have no scientifically valid evidence for the presence of other life forms in the Solar System, we do not have strong evidence for their absence, either. Observational limits on artificial probes within the Solar System are very weak, and small probes may even be hiding among us (Tough 2000).

[6] One should be somewhat careful with arguments based on sustained exponential growth, of course. At the same growth rate, humanity would need to generate more than $1\,L_\odot$ in less than 10,000 years.

In summary, all possibilities are still open. Other intelligent forms of life may be here (within the Solar System), they may be there (within the Galaxy), or they may be nowhere.

SETI Strategies

The fundamental task of SETI is finding artificially generated signals beamed towards the Earth, and distinguishing them from the astrophysical and instrumental backgrounds. A good strategy is looking for signals whose time-bandwidth product approaches the limiting value $B\tau \approx 1$ set by the uncertainty principle (Tarter 2001). Most SETI experiments have been searches for narrow-band signals in the radio regime, where stars are comparatively weak emitters. (FM and TV transmitters are factors 10^6 to 10^9 brighter than the quiet Sun.) The main difficulty of this approach is the enormous search space that needs to be covered (position, frequency, frequency drift due to relative acceleration of emitter and receiver, pulse format). One either has to concentrate on "magic frequencies", assuming that the transmitter will choose e.g., a frequency close to the 1,420 MHz hydrogen line, or construct a spectrometer with many millions of frequency channels, and provide sufficient computer power for the data processing.

An alternative approach are optical searches for very short laser pulses. The flux from a Solar-type star at a distance of 300 pc is $\sim 4 \cdot 10^5$ phot $\mathrm{m}^{-2}\,\mathrm{s}^{-1}$ in a broad-band optical filter. The arrival of two or more photons with in sub-microsecond time window is thus exponentially suppressed by Poisson statistics; such events could therefore be attributed to a laser beacon. The operational optical SETI experiments use two or more fast avalanche photodiodes in coincidence to detect nanosecond pulses.

A further choice has to be made regarding the targets of the search. One can either point the (optical or radio) telescope at selected nearby stars, thereby maximizing the sensitivity and therefore the chances to detect a nearby, relatively weak transmitter, or conduct a sky survey, which optimizes detectability of more distant, very strong emitters. Sometimes an intermediate solution is chosen, by piggy-backing a SETI receiver to a telescope during observations taken for a different purpose. This circumvents the difficulty of obtaining a large amount of dedicated telescope time, but makes the search somewhat less efficient, depending on the nature of the primary observing program.

Instead of searching for signals beamed explicitly towards us, one might also envisage looking for electromagnetic signals generated for the internal communication needs of extraterrestrial civilizations, leaking more or less isotropically from their home planet into space. Since we have been using radio transmitters for the better part of a century, intelligent life on Earth is now detectable out to a distance of ~ 25 pc, but the signal is relatively weak – current targeted SETI programs can detect the equivalent power of strong TV transmitters out to ~ 0.3 pc, not a very useful distance.

The Scientific Impact of SETI

SETI projects have always been justified on the basis of the large impact that a positive result would have. But a well-designed scientific experiment should also satisfy the criterion that even a null-result constitutes significant progress, and provides new insight. So far, the failure of all SETI efforts to pick up a signal does not tell us very much, because the search space that has been covered is still quite small. To quantify this statement, we can use the SETI figure of merit (Dreher and Cullers 1997)

$$\mathcal{S}(P_T) \equiv N_{\rm stars}(P_T) \cdot \ln(\nu_h/\nu_l) \cdot \eta_{\rm pol} \cdot N_{\rm looks}\,. \tag{5}$$

Here P_T is the effective isotropically radiated power (EIRP) of the transmitter[7], $N_{\rm stars}(P_T)$ the number of stars observed by a SETI project for which a transmitter of strength P_T is detectable, ν_h and ν_l the upper and lower limits of the frequency range covered, $\eta_{\rm pol} \in \{0.5, 1\}$ a polarization efficiency factor, and $N_{\rm looks}$ the number of observations for each target star. In Fig. 3, $\mathcal{S}(P_T)$ is plotted as a function of P_T for the most important ongoing surveys. Taking for simplicity all factors on the right-hand side of (5) except the first to be of order unity,[8] we can draw some simple conclusions from this figure. First of all, there are few (if any) civilizations in the Galaxy transmitting in excess of 10^{23} W equivalent power to us at the frequencies covered by the radio surveys (mostly near the 1,420 MHz line). Second, we can define a "reasonable" equivalent transmitter power P_T, which we think an advanced civilization could afford, by multiplying the actual power with the antenna gain. The curves in Fig. 3 then tell us how many stars have been surveyed to that depth. Conversely, we can wage a guess the rate of incidence of civilizations (one per 10^x stars, where x is your guess), look up that number on the y-axis, and infer the equivalent power that these civilizations would have to use to be detected in our surveys.

In the coming decades, new radio arrays with large collecting area (the Allen Telescope Array, and perhaps later a Square Kilometer Array) will enable much more sensitive all-sky surveys than currently possible. The detection of artificial signals in these surveys is largely a computational problem; the extrapolation of Moore's Law therefore predicts a dramatic increase in the search capabilities in a relatively short time. By the middle of the 21st century, it should be possible to search a significant fraction of all stars in the Galaxy to an interesting limit, comparable to the strongest current man-made signals. A null result from such a survey would indeed place strong constraints

[7] The actual power needed for a directed transmission is much smaller, of course. Using an optical telescope with a resolving power of $0\farcs1$ to send a beacon, for example, reduces the power requirement by a factor 10^{14} compared to an isotropic emitter. This factor is called *antenna gain*.

[8] The logarithmic frequency factor is actually as small as 10^{-3} for some of the surveys shown (Tarter 2001).

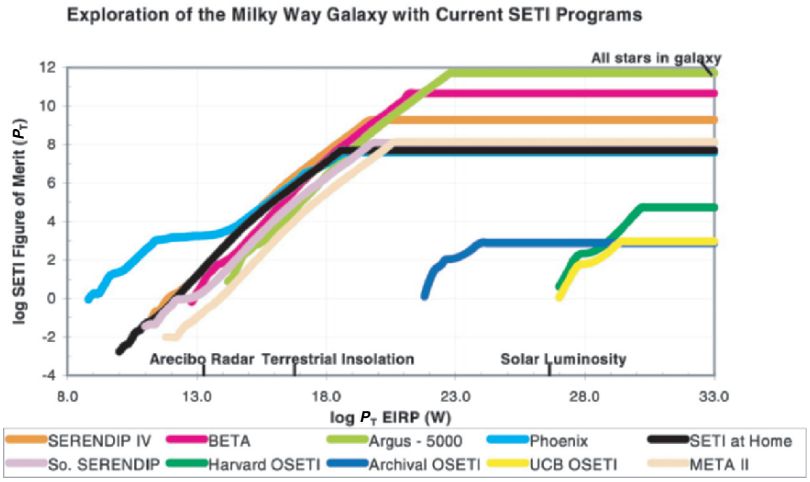

Fig. 3. SETI Figure of Merit as defined in (5) for current searches. The power axis represents the average effective isotropically radiated power for narrow-band continuous wave searches, but is the peak transmitter power for short optical pulses. From Tarter (2001)

on the number of civilizations in the Galaxy. On the other hand, we might be lucky and receive transmissions from intelligent beings even before we can launch the planned space missions TPF/DARWIN which will be capable of identifying the chemical signature of primitive life through spectroscopy of the atmospheres of nearby habitable planets.

2 Planet-Forming Disks

The origin of planetary systems is intimately linked to the formation of their parent stars. The theory of star formation and observations of young stellar objects are large fields of astronomy; we will only summarize briefly those aspects that are relevant for planet detection methods. For excellent introductions into star formation see e.g. Shu et al. (1987), van Dishoeck and Blake (1998), and Dutrey (1999). It is now generally accepted that circumstellar disks play an important role during the pre-main-sequence phase, and that planets are formed in these relatively massive disks. The disk hypothesis for the formation of the Solar System actually dates back to Kant (1755) and Laplace (1796); it provides a framework in which the following salient features of the Solar System can be explained:

- The planetary orbits are nearly circular and coplanar.
- The orbital motions and rotations of the Sun, planets, and moons are predominantly in one sense.

- The Solar System is differentiated, with systematic trends in the planetary properties with distance from the Sun.
- The Solar System contains a large number of small bodies (asteroids and comets), again with properties that vary systematically with distance from the Sun.

These properties are not prescribed by Kepler's Laws or other fundamental laws of physics, but they are a direct consequence of the way the Solar System formed (e.g., Encrenaz 2001). We may therefore expect that extrasolar planetary systems may share these general characteristics with ours because they were formed by the same processes and in a similar environment. It should be pointed out, however, that mechanisms such as orbital migration and two-body scattering might in many cases have played a more prominent role than in the Solar System. The detections of massive planets in orbits with small radii, and of planets in highly eccentric orbits (see Sect. 3) are certainly an indication that a large range of outcomes is possible from the processes that shape nascent planetary systems.

2.1 Star Formation: the General Framework

"Bimodal" Star Formation

Stars form in molecular clouds, the coldest (typically $10 \ldots 50\,\mathrm{K}$) and densest ($n \approx 10^3\,\mathrm{cm}^{-3}$) phase of the interstellar medium. Because of the low temperatures and extremely high opacity, the first phases of star formation can only be observed at radio, millimeter, and mid-infrared wavelengths. These wavelength ranges contain rich spectral information, which can be used to diagnose the physical state of the gas, to trace large-scale motions, and to pinpoint the sites at which stellar embryos are forming. Rotational and vibrational transitions of molecules such as H_2, CO, CS, NH_3, H_2O, CH, OH, and many others, provide information on gas temperature, density and kinematics, the far-infrared continuum emission probes the distribution and temperature of dust, near-infrared images reveal young embedded stars, and infrared fine structure lines and radio free-free emission trace hot gas that has been ionized by newborn massive stars (e.g. Genzel 1992).

One can broadly distinguish between two distinct star-forming environments: cold dark clouds, which form only low-mass stars, and giant molecular complexes, which are associated with high-mass star formation (Evans 1999). In giant complexes, of which the Orion A cloud and NGC 3603 are prominent examples, high-mass and low-mass stars form in close proximity to each other (e.g., Eisenhauer et al. 1998); the stellar density in the Orion region is $\sim 2 \cdot 10^4\,\mathrm{pc}^{-3}$. In contrast, low-mass star forming regions are much less dense, and form only stars with masses $\lesssim 2\,M_\odot$. There are also differences between individual regions; whereas the Taurus–Auriga clouds have a filamentary structure with isolated star formation, the ρ Ophiuchi cloud is a cluster with a higher density of newly born stars.

The Solar neighborhood is a mix of stars born in different environments. Roughly 80% of all Solar-mass stars may have originated in dense regions like Orion, but for any individual old star it is practically impossible to ascertain whether it formed in this way or in relative isolation. Consequently, we are looking at planetary systems (or stars without planetary companions!) that were exposed to widely different environmental influences during their youth. It is quite possible that the ionizing flux of young massive stars modifies or even destroys the circumstellar disks in the surrounding cluster before they can form planets (Armitage 2000); this may indeed be happening near the Orion Trapezium (Johnstone et al. 1998). In this context one should keep in mind that all current information on extrasolar planets comes from observations of low-mass stars (up to $\approx 1.5\,M_\odot$). It is certainly possible that all known planets are associated with stars that were born in relative isolation. Massive stars, which form exclusively in dense clusters, may therefore conceivably have planetary systems with vastly different properties, or no planets at all. This interesting question can for example be addressed with astrometric surveys of high-mass stars and of pre-main-sequence stars (see Sect. 9.7).

Molecular Cloud Collapse, Fragmentation, and Low-Mass Star Formation

Molecular clouds typically have a clumpy structure; the densest clumps are associated with star formation. According to the famous *Jeans criterion*, a parcel of gas with temperature T, density ρ, and mean molecular mass $\overline{\mu}$ will collapse under its own gravitational attraction if its mass is above the critical value

$$m_J \equiv \left(\frac{\pi k T}{4 G \overline{\mu} m_H}\right)^{3/2} \rho^{-1/2}$$

$$\approx 5 \cdot 10^4 \left(\frac{T}{100\,\mathrm{K}}\right)^{3/2} \left(\frac{\rho}{\mathrm{cm}^{-3}}\right)^{-1/2} M_\odot\,. \qquad (6)$$

This expression shows that only clumps with masses large compared to individual stars can start collapsing. (Even for gas as cold as 10 K, a density of $2.5 \cdot 10^6\,\mathrm{cm}^{-3}$, far in excess of typical values for quiescent molecular clouds, would be required for a $1\,M_\odot$ clump to collapse.) After the collapse has been initiated, fragmentation will produce sub-clumps, which will then proceed to form individual stars. Conservation of angular momentum naturally leads to an increasingly flattened morphology, and finally to the formation of a disk.

Starting with these processes, the formation of a low-mass star can be described by four main phases (e.g. Shu et al. 1987):

1. A slowly rotating pre-stellar core forms by losing its magnetic and turbulent support.

2. The core collapses "from the inside out". A central protostar and disk are formed, which are deeply embedded within an infalling envelope of dust and gas. The luminosity of the protostar is derived from accretion.
3. Deuterium ignites in the central region, which becomes convective. A stellar wind develops, primarily perpendicular to the disk. This is the bipolar outflow phase.
4. Accreting material falls on the disk rather than the star. The opening angle of the wind widens; it blows away the surrounding gas. This is the T Tauri phase, in which the pre-main-sequence star is surrounded by a remnant circumstellar disk.

The later two of these phases correspond to four observationally distinct classes of objects, described in Fig. 4 (André 1994). During the pre-main-sequence evolution, there is a general trend for the peak of the spectral energy distribution to move towards shorter wavelengths, partly because of increasing temperatures, partly because of decreasing opacity towards the central source. During the same time, the mass of the envelope/disk decreases from nearly a Solar mass to a small fraction of that value. The formation of planetary systems must coincide with the phases that are accompanied by a sufficiently massive disk; the time available for planet formation is therefore limited by the time scale for disk dispersal.

Observational data on the lifetime of pre-main-sequence disks can most easily be obtained from a census of the circumstellar disk fraction in young clusters, spanning a significant range in age (Haisch et al. 2001b). In the youngest clusters the disk fraction is very high ($\gtrsim 80\%$); it decreases rapidly with cluster age. About half of the stars lose their disks within $\sim 3\,\mathrm{Myr}$; at and age of $\sim 6\,\mathrm{Myr}$ essentially all disks have disappeared. Strictly speaking these data refer to the hot inner disk that gives rise to the near-infrared excess, but the lifetime of the outer disk, in which planets can form, seems to be closely related (Haisch et al. 2001a). We can thus infer that planets must form within a few Myrs.

2.2 Observations of Dusty Disks

Pre-Main-Sequence Disks

As explained in Sect. 2.1, T Tauri stars are new-born Solar-type stars with ages of 10^5 to 10^7 yrs, which have emerged relatively unobscured from their natal molecular clouds (Bertout 1989). The detection of an infrared excess around many of these stars, and the subsequent modeling of their spectral energy distribution, led to the wide acceptance of a circumstellar disk interpretation (Beckwith and Sargent 1993). A key argument in favor of this model was the realization that a spherical geometry is prohibited, since the dust in the system would then completely obscure the central star; an axial symmetry

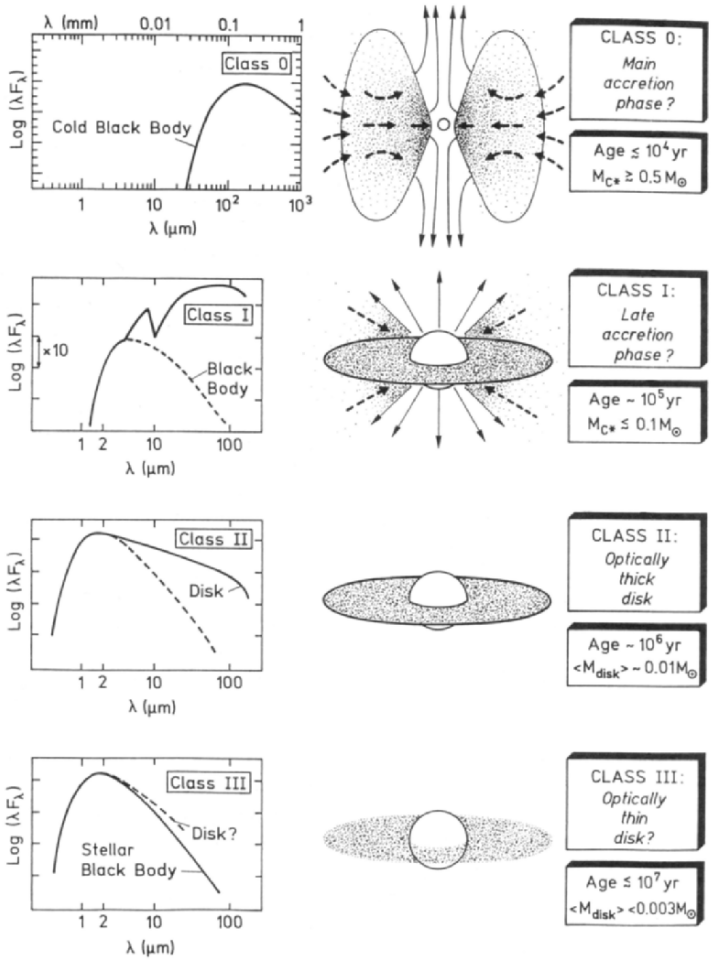

Fig. 4. The main stages of star formation (from *top* to *bottom*). The *left* column shows schematic spectral energy distributions. The overall geometry of the protostar and its immediate environment is sketched in the *middle* column. The *right* column gives approximate ages and disk masses. For details see the text. From André (1994)

is also supported by observations of bipolar optical jets and molecular outflows.

Observations of a number of pre-main-sequence disks in the millimeter/submillimeter continuum can be fitted with a λ^{-1} emissivity law for the dust, which implies that grain growth is occurring in them (Dutrey 1999). This is in general agreement with the ideas about grain growth discussed above. Since the emission is mainly optically thin, it allows measuring the disk mass;

values in the range $10^{-3} \ldots 10^{-1}\,M_\odot$ have been derived.[9] The star HL Tau is a good example for a very young (~ 0.1 Myr) object of mass $\sim 1\,M_\odot$. Its spectral energy distribution can be fitted by a toroidal circumstellar envelope of mass $\sim 0.11\,M_\odot$ (Men'shchikov et al. 1999). This model requires the presence of very large ($\gtrsim 100\,\mu$m) dust grains in the inner 100 AU, and much smaller grains ($\lesssim 1\,\mu$m) in the outer regions, again in agreement with expectations.

The first resolved images of circumstellar dust disks were obtained in the mid-eighties with interferometric observations of the millimeter continuum emission, and a few years later with optical and infrared high-resolution imaging (Koerner 1997). Famous examples are the compact objects seen silhouetted against the bright H II region associated with the Orion Trapezium Cluster (O'Dell et al. 1993; McCaughrean and O'Dell 1996; Bally et al. 2000). The objects closest to the OB stars have a cometary morphology, with a tail pointing away from the hot stars, and an ionization front towards them. Objects further from the OB stars have a more elliptical appearance with a star visible in the center, as expected for a sample of pre-main-sequence objects surrounded by disks with diameters of a few hundred AU.

Many other pre-main-sequence disks have been imaged in scattered light at infrared and visible wavelengths over the past few years, both from the ground and from space (e.g., Roddier et al. 1996; Potter et al. 2000; Stapelfeldt et al. 1998; Krist et al. 2000, 2002; Schneider et al. 2003; McCaughrean et al. 2000). The object HH-30 is a particularly interesting case (Burrows et al. 2006). The dust disk is seen nearly edge-on, and appears as a dark lane with a diameter of ~ 500 AU. The central star is completely obscured, so that the surface of the flaring disk can clearly be seen in scattered light. A highly collimated jet is aligned with the rotation axis of the disk; clumps of gas are ejected along the jet axis at a speed of $\sim 200\,\mathrm{km\,s^{-1}}$. The fortuitous orientation of HH-30 thus allows an unusually detailed look at the geometry of the disk and jet, demonstrating that the collimation must take place within ~ 30 AU of the star.

The steeply descending radial density and temperature profiles make the detection of the outer disks (beyond ~ 200 AU) in the mm continuum very difficult. This region can be probed through its CO emission, however. Disks around T Tauri stars such as GM Aur are found to have outer radii of $100 \ldots 800$ AU; velocity gradients along the major axes of the disks clearly demonstrate Keplerian rotation, as expected (Koerner et al. 1993; Guilloteau and Dutrey 1998; Dutrey et al. 1998). Many other molecules have been detected in pre-main-sequence disks and provide detailed diagnostics of the physics and chemistry of the gas. Some molecules (HCN, H_2CO) are depleted by factors ~ 100 due to condensation onto grains (Aikawa and Herbst 1999; Dutrey et al. 1997). While large CO disks are common around stars at an age

[9] Note that gas masses derived from observations of dust or CO are based on assumptions about the dust/gas and CO/H_2 ratios, respectively. In circumstellar disks these ratios can be substantially different from the "standard" ISM values. For details see Dutrey (1999).

of $\sim 10^6$ yrs, they have not been found around older stars, at $10^7 \ldots 10^8$ yrs (Liseau and Artymowicz 1998; Greaves et al. 2000).

Debris Disks

Dusty circumstellar disks around main-sequence stars are quite different in nature from pre-main-sequence disks, as their dust mass is at least 10 times higher than their gas mass. The Infra-Red Astronomical Satellite (IRAS) discovered that some bright nearby stars (among them Vega and Fomalhaut) emit much more strongly at wavelengths between 25 µm and 100 µm than expected (Aumann et al. 1984; Backman and Paresce 1993; Lagrange et al. 2000). Coronographic observations of β Pictoris, the star with the strongest infrared excess, soon showed that this "Vega phenomenon" is due to a circumstellar dust disk (Smith and Terrile 1984). A disk- or ring-like morphology has also been found in several other cases (including Vega, Fomalhaut, ε Eri, and HR 4796 A, Zuckerman 2001), but some more distant Vega-excess stars are associated with reflection nebulosities reminiscent of the Pleiades, indicating that their far-IR excess may also be caused by local heating of interstellar dust (Kalas et al. 2002).

The properties of a few prominent and well-studied main-sequence dust disks are listed in Table 4. They are all much more luminous than the dust disk of the Solar System, and they are associated with relatively young stars. Indeed, there is a clear relation between the fractional dust luminosity $f_d \equiv L_{\rm dust}/L_*$ with stellar age τ_*; it can be represented by a power law $f_d \propto \tau_*^{-1.76}$ (Spangler et al. 2001). There is a substantial gap in our knowledge of dust

Table 4. Properties of debris disks

star	d [pc]	r_c [AU]	$T_{\rm dust}$ [K]	$s_{\rm dust}$ [µm]	$L_{\rm dust}/L_*$	$\tau_{\rm PR}$ [yr]	$\tau_{\rm coll}$ [yr]	τ_* [yr]
α Lyr	8	150	85	$10 \ldots 100$	$2 \cdot 10^{-5}$	10^7	10^7	$4 \cdot 10^8$
α PsA	7	150	60	≈ 10	$8 \cdot 10^{-5}$	10^7	10^6	$2 \cdot 10^8$
β Pic	19	75	110	≈ 1	$2 \cdot 10^{-3}$	10^6	10^4	$1.2 \cdot 10^7$
ε Eri	3	39	50	≈ 10	$7 \cdot 10^{-5}$	10^8	10^6	$8 \cdot 10^8$
Zodiacal dust			280	$1 \ldots 100$	$1 \cdot 10^{-7}$			
Kuiper Belt			< 40	> 1	$< 10^{-7}$			

Listed are the distance d, characteristic radii r_c (derived from the color temperature), dust color temperature $T_{\rm dust}$, estimated grain sizes $s_{\rm dust}$, disk luminosity compared to the star $L_{\rm dust}/L_*$, Poynting–Robertson lifetime of typical grains $\tau_{\rm PR}$, collisional time scale $\tau_{\rm coll}$ at r_c, and estimated age of the star τ_*. Properties of Solar System dust (Zodiacal dust and dust in the Kuiper Belt) are included for comparison. Adopted from Mann (2001), with the age of β Pic taken from Zuckerman et al. (2001)

disks between relatively young stars and the Sun, due to the limited sensitivity of the IRAS and Infrared Space Observatory (ISO) missions. This gap will hopefully be closed soon by the Space Infra-Red Telescope Facility (SIRTF), which should be sensitive down to the mass in small grains inferred for our present-day Kuiper Belt ($6 \cdot 10^{22}$ g) surrounding a Solar-type star at 30 pc (Meyer et al. 2001).

Dust grains spiral into their parent stars due to Poynting–Robertson drag, and they are destroyed by mutual collisions (Dermott et al. 2001). Typical time scales for these processes are listed in Table 4 together with the ages of the stars.[10] We see that in all Vega-excess disks the life time of dust particles is much smaller than the age of the star. This leads to the important conclusion that the dust disk cannot be a remnant from the star formation process, but must be replenished by erosion of planetesimals or cometary bodies; hence the expression *debris disk* for these objects. The general trend of decreasing disk mass with age is then consistent with the evolution of the Solar System; the prominent Vega-excess disks correspond to the time of heavy bombardment in the early Solar System. If the dust disk of β Pic is generated by collisional destruction of 1-km planetesimals, a very large mass ($\sim 125\, M_\oplus$) must reside in planetesimals (Artymowicz 1997). An alternative scenario, based on the evaporation of cometary bodies, will be discussed in Sect. 2.3.

The amount and nature of gas in the disks of Vega-excess stars has been investigated with emission and absorption spectroscopy; the largest body of data is (not surprisingly) available for β Pic. Radio observations give an upper limit of $\sim\ 1.5\, M_\oplus$ to the atomic hydrogen content from a non-detection of the 1,420 MHz line (Freundling et al. 1995). The data on H_2 seem to be contradictory. Tentative detections of the S(0) and S(1) rotational transitions with ISO imply an H_2 mass of $\sim 0.2\, M_{\rm Jup}$ (Thi et al. 2001), whereas the lack of H_2 absorption against the stellar disk in the ultraviolet places an upper limit of $\sim 3 \cdot 10^{-4}\, M_{\rm Jup}$ on that same quantity (Lecavelier des Etangs et al. 2001). This discrepancy could in principle be resolved by a clumpy structure of the molecular hydrogen – the clumps would show up in emission, but the line-of-sight towards the star may not be covered. It seems difficult, however, to hide such a large mass of H_2 in this way; more data are clearly required to resolve this issue. In contrast to H_2, many ions and neutral species, as well as CO, have been found in absorption against β Pic (Vidal-Madjar et al. 1994; Lagrange et al. 1998; Roberge et al. 2000); the relative abundances of the refractory elements appear to be close to Solar. The amount of neural gas can be estimated from Na I emission (Olofsson et al. 2001) and from a tentative detection of the 157.7 µm [C II] fine structure transition with ISO

[10] Note that the Poynting–Robertson and collisional time scales depend strongly on the particle size and position in the disk. The tabulated values therefore give only rough indications on the correct orders-of-magnitude. For more details see Artymowicz (1997) and Dermott et al. (2001).

(Kamp et al. 2003); the latter authors conclude that a gas mass in the range $0.2 \ldots 4\,M_{\oplus}$ and a gas-to-dust ratio of $0.5 \ldots 9$ is consistent with all observations.

Evidence for Planets in Disks?

As we have seen above, the existence of debris disks in itself provides a strong argument for the existence of solid bodies in these systems, as there must be a mechanism to replenish the dust such as the erosion of km-sized planetesimals. This does not tell us anything about the existence of planets, however; for some reason the later stages of their formation process may simply have failed. It might be possible, however, to detect planets indirectly by their influence on the morphology of circumstellar disks. It has been pointed out earlier that sufficiently massive planets open gaps in disks, which may be observable directly or indirectly (cf. Fig. 5). After most of the disk material has been

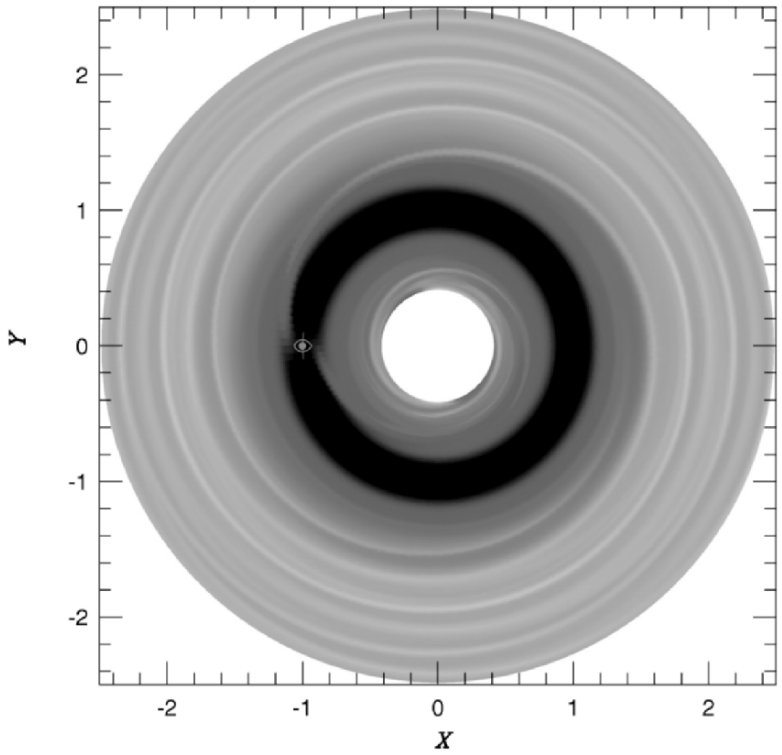

Fig. 5. Numerical simulation of a planet forming in a circumstellar disk. From Kley et al. (2001)

cleared, dust rings may remain an provide signposts of recent planet formation (Kenyon and Bromley 2002).

The temperature of grains at any given radius in the disk is determined by radiative equilibrium with the star; the spectral energy density (SED) of the disk is therefore the superposition of the Planck functions appropriate for grains at the proper temperatures (neglecting spectral features). If a gap is opened in the disk, there are no grains in the corresponding temperature range, which should result in a dip in the SED. Unfortunately, this dip is not very deep and extremely broad, extending over three orders of magnitude in wavelength. Without further information, it is impossible to identify such a feature, and to distinguish it from variations in the overall disk and dust parameters (Steinacker and Henning 2003). Interferometric imaging in the mid-infrared or at mm/sub-mm wavelengths is a better technique for the detection of gaps in disks (Wolf et al. 2002). An obvious caveat is that for observations at $\lambda \sim 10\,\mu$m the emission is dominated by dust at ~ 1 AU, where the temperature is ~ 300 K. This means that gaps at larger radii have to be observed at longer wavelengths. The Atacama Large Millimeter Array (ALMA) will be very well suited for this task.

Most observational data on gaps in disks come from somewhat older objects, already close to or in the debris disk phase. A good example is HD 4796 A, which is surrounded by a prominent ring-like structure that has been observed at near- and mid-infrared wavelengths (Schneider et al. 1999; Telesco et al. 2000). It is tempting to attribute the relatively sharp truncation at the inner and outer edges of the annulus to the dynamical action of one or more "shepherd planets". The inner edge of the bright ring appears to be slightly asymmetric, which has been interpreted as an indication of a planet in an eccentric orbit (Wyatt et al. 1999). A similar disk with and apparent gap at a radius of ~ 250 AU has been found around HD 141569 A, a ~ 5 Myr-old intermediate-mass star (Weinberger et al. 1999, 2000). However, improved coronographic imaging of this object with the ACS instrument on HST shows that the previously observed structure in the disk is not a ring but rather tightly wound spiral, which could be due to tidal interaction with the nearby binary HD 141569 BC (Clampin et al. 2003). As this example shows, it is quite difficult to proof beyond doubt that an observed structure in a disk cannot be caused by anything but a planet.

Millimeter and sub-millimeter images of the most prominent Vega-excess stars (α Lyr, α PsA, β Pic, and ε Eri), which probe relatively cool dust on scales $\gtrsim 50$ AU, have consistently revealed surprising morphologies with strong clumping, quite different from simple smooth disks (Greaves et al. 1998; Holland et al. 1998, 2003; Liseau et al. 2003). Perhaps most intriguing is the dust disk of ε Eri, whose morphology has been modeled with a resonant patter due to a mean-motion resonance with a planet (Ozernoy et al. 2000; Quillen and Thorndike 2002). These models predict changes of the observed structure with time. It should simply revolve around the star if the planetary orbit is circular, or change its shape with orbital phase if the orbit is eccentric. It will

thus be possible to distinguish between different models on the orbital time scale of the presumed planet(≈ 140 yrs).

The disk of β Pic is a special case again, because its brightness has enabled so much high-quality information to be gathered, and because its edge-on orientation allows observations of out-of-plane distortions. Large-scale asymmetries and warps have been reported on all scales accessible to observations ($10 \ldots 400$ AU), both in reflected light and in thermal infrared emission (Heap et al. 2000; Pantin et al. 1997). The warps could be caused by a planet in an orbit that is slightly inclined with respect to the disk plane; a fairly large range of planetary masses and orbital radii would be compatible with the observations (Mouillet et al. 1997). Substructure in the outer disk, somewhat reminiscent of Saturn's rings, has been detected by comparing details in space-based and ground-based images (Kalas et al. 2000). They might have been generated by a close stellar encounter in the last $\sim 100,000$ yrs, which could then also have triggered the outer warps. The planetary hypothesis remains the best explanation of the recently detected small-scale warp within the inner ~ 20 AU, which is even stronger than the outer distortions (Weinberger et al. 2003; Wahhaj et al. 2003). It is also possible that two planets in orbits that are inclined with respect to each other and with respect to the disk cause the complicated warped structure. An unambiguous resolution of this question will probably have to await observations with even higher spatial resolution.

While imaging observations probe the structure of circumstellar disks on scales of many AU, photometry is a better tool to diagnose inhomogeneities in edge-on disks at much smaller radii. Deep periodic occultations of the T Tauri star KH 15D have indeed been attributed to eclipses by dust, indicative of either a clump or warp in the circumstellar disk at a radius of ~ 0.2 AU (Hamilton et al. 2001; Herbst et al. 2002). It is tempting to speculate that this distortion is caused by the gravitational action of a planet, but many alternative explanations are still viable. In any case, continued monitoring of this star will yield interesting information on the structure of circumstellar material close to a young star.

2.3 Observations of Infalling Material and Evidence for Extrasolar Planetesimals

The scenario of planetary system formation outlined above implies that a large number of planetesimals remain in the circumstellar disk during the phase when large planet-mass bodies are already there. Scattering of these small planetesimals by the planets leads to drastic changes of their orbits. Some of them are thrown out of the system or into bound orbits with very long periods, others impact one of the planets, and yet others get so close to the star that they evaporate. These processes correspond to the formation of the Oort Cloud in the Solar System, and the epoch of "heavy bombardment" of the terrestrial planets that is still evident in the cratering record of the Moon.

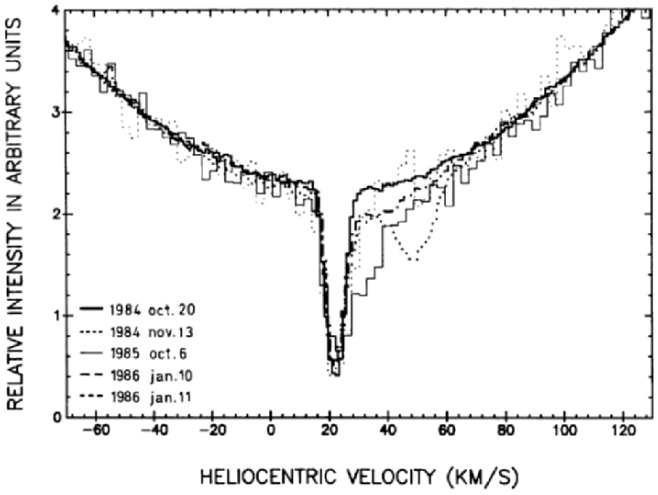

Fig. 6. Spectra of β Pic in the region of the Ca II K line. They have been normalized such that the "continua" (i.e., the stellar line broadened by rotation) coincide. Remarkably, the variations of the line profile all arise redshifted relative to the stellar radial velocity. From Ferlet et al. (1987)

There are strong indications that this phase has now been observed spectroscopically in several young planetary systems (Grinin 1999; Grady et al. 2000b). Intermittent narrow absorption features have been detected in the red wings of the UV and visible lines of β Pictoris (Ferlet et al. 1987, see Fig. 6). These absorption events have time scales from a few hours to many days. The only explanation consistent with most of the observational material is that of swarms of evaporating planetesimals; the dense absorbing cloud is essentially a large cometary coma. The required evaporated mass per event (if abundances appropriate for planetesimals are adopted) usually corresponds to km-size planetesimals. The infall episodes occur up to 200 times per year, but the event rate varies substantially from year to year. This, together with the large predominance of redshifted over blueshifted features, suggests that the observed infalling bodies belong to an orbital family (perhaps fragments of a disintegrated large comet)[11] with the mean orbit oriented in such a way that they approach the star when they are seen against it (Artymowicz 1997, see Fig. 7). The evaporation of comets could also be the dominant process responsible for replenishing the gas disk of β Pictoris (Lecavelier des Etangs 1998).

Redshifted absorption features have also been detected in the spectra of Herbig Ae/Be stars, which are pre-main-sequence stars of intermediate mass and thus the progenitors of objects like β Pictoris. In fact, variations of the

[11] Analogous groups of Sun-grazing comets exist also in the Solar System (Sekanina 2002 and references therein).

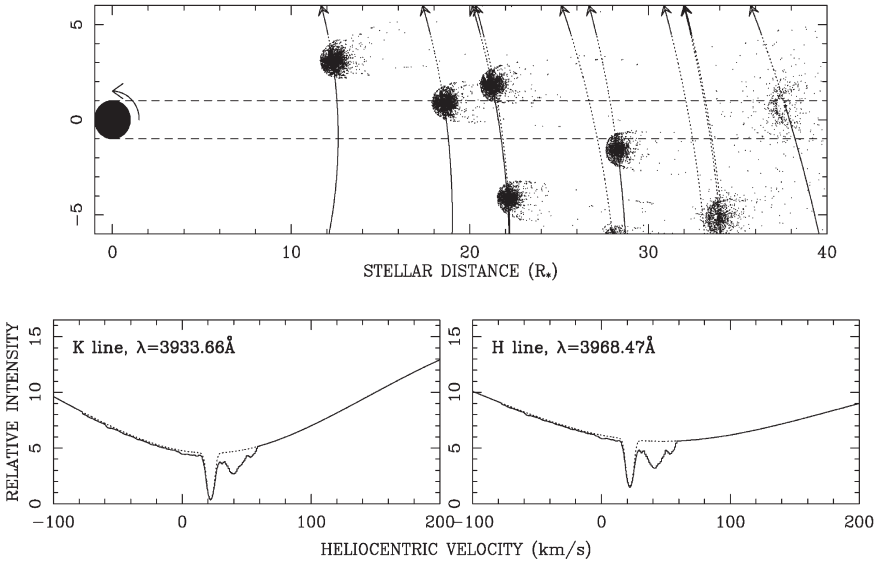

Fig. 7. Top panel: Simulated view of falling evaporating bodies (planetesimals or comets) passing in front of the star (*black circle* at the origin; the observer is to the right). Each comet evaporates creating a coma filled with dense gas containing singly ionized calcium, occulting a fraction of the stellar disk. *Bottom panels*: The cometary comae from the top panel cause multiple, redshifted, variable absorption features superimposed on both the stable narrow circumstellar and the broad stellar K and H lines of Ca II. From Artymowicz (1997)

line profiles appear to be ubiquitous, in particular in stars that are believed to be surrounded by an edge-on disk (Grady et al. 1996). The interpretation of the spectral variability is more difficult for these objects, however. They are still surrounded by a fairly massive gaseous disk, which makes it more difficult to distinguish between accretion of relatively unprocessed disk material and cometary events. Detailed multi-line analyses of UX Orionis have concluded that for this star the infalling (and outflowing) gas cannot be heavily hydrogen-depleted, as would be expected if it originated from the evaporation of solid bodies (Natta et al. 2000; Mora et al. 2002). Accretion of material with Solar-like chemical composition appears to be the dominant process for most very young stars, giving way to more episodic infall of metal-rich gas at an age of ~ 10 Myr (Grady et al. 1997, 2000a; Mora et al. 2003). The best cases of comet evaporation associated with younger stars are 51 Oph (Roberge et al. 2002) and WW Vul (Mora et al. 2003).

Our current understanding of the relation between the occurrence of planetesimals/comets and stellar age is limited by the relatively small number of observed stars, and by the low duty cycle with which they have been monitored. High-precision photometry from space (see Sect. 6.5) is a much

more efficient observing technique than high-resolution spectroscopy, and may lead to the detection of a large number of extrasolar comets (Lecavelier des Etangs et al. 1999). It will be necessary, however, to discriminate between eclipses caused by comets, by planets, or by dust clumps.

The information from infalling comets, combined with mid-infrared spectra of Herbig Ae/Be stars, contains important clues about the ways dust is processed in pre-main-sequence disks. Silicates contained in interstellar dust are predominantly amorphous, but observations from the ground and with the Infrared Space Observatory (ISO) have shown that the disks surrounding β Pic and the Herbig star HD 100546 contain significant amounts of crystalline silicates (Knacke et al. 1993; Malfait et al. 1998). The production of these crystals requires processing at temperatures near 1,000 K, which are only reached in the innermost disk near the star (Hill et al. 2001). Since the disks of these stars are most likely being replenished by evaporating comets that formed at radii beyond the snow line, the crystalline material must have been incorporated in these comets. One is thus compelled to conclude that there must have been a large-scale transport process connecting the different regions of the disk at the epoch of comet formation (see Fig. 8). Similar indicators of crystalline silicates have also been found in some Solar System comets, and lead to the suggestions that large-scale mixing has been important here, too (Hanner et al. 1994). Observations of extrasolar comets can thus help us understand the chemical and mineralogical evolution of grains and particles from which planets can form.

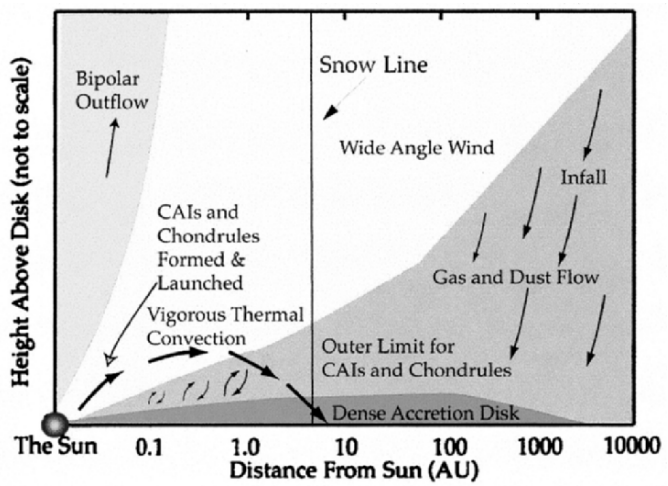

Fig. 8. Schematic drawing of a protostellar nebula (distance on a logarithmic scale) centered on the proto-Sun. The region of dust processing is shown (< 1 AU together with the snow line (≈ 5 AU), infalling envelope, bipolar outflow, and less collimated wind. Comet formation is expected to occur from ≈ 5 AU to beyond 40 AU. From Hill et al. (2001)

The evaporation of comets may be observable not only at a very young age, but also near the end of the stellar life time. When the star evolves off the main sequence, it expands and becomes much more luminous; this leads to the evaporation of icy bodies at increasing radii, up to several hundred AU (Stern et al. 1990). During the late phase of stellar evolution, convection of material from the core can enhance the outer layers with carbon. The oxygen in the winds from such carbon stars will be almost completely locked in CO, with a predicted H_2O abundance $\lesssim 10^{-11}$ (Willacy and Cherchneff 1998). The discovery of water vapor around the carbon star IRC+10216 with an implied abundance $\sim 10^{-7}$ can therefore only be explained by an external addition of H_2O to the wind, most plausibly through the evaporation of the Kuiper Belt around this star (Melnick et al. 2001). The observed rate of H_2O mass loss requires the evaporation of $10\,M_\oplus$ of ice over the life time of the strong stellar wind; this is comparable to the original mass of water ice in the Solar System's Kuiper Belt. More sensitive future observations with ESA's Herschel telescope will be able to probe the presence and mass of Kuiper Belts around a much larger sample of carbon-rich stars.

3 The Currently Known Extrasolar Planets

The discovery of a planet orbiting the star 51 Peg (Mayor and Queloz 1995, see Fig. 9) has opened a new field of observational astrophysics: the systematic study of planetary systems, their dynamical properties, formation, and evolution. About 150 extrasolar planets have now been detected, giving us a first glimpse at the diversity of these objects, and allowing the first statistical inferences. Unexpected discoveries have stimulated interesting new theoretical developments. Not the least of the big surprises in this field was the discovery of 51 Peg b itself, a giant planet with an orbital period of only 4 days! It probably formed at a much larger distance from its parent star, and migrated subsequently to its current position. In this chapter we will take a look at this and other phenomena, including orbital eccentricities and dynamical interaction between planets in multiple systems.

3.1 The First Hundred Planets Around Solar-Type Stars

The Presently Known Planets

The radial-velocity surveys (see in Sect. 4.4 for more details) have discovered about one hundred planets over the past seven years. One can thus rightfully speak about extrasolar planets as a new branch of observational astrophysics. The first discoveries showed that our Sun is not unique in having planetary companions. The field has progressed very quickly from this exploratory stage to a phase where it has become possible to perform the first meaningful statistical analyses of the properties of giant planets, of the distribution of their

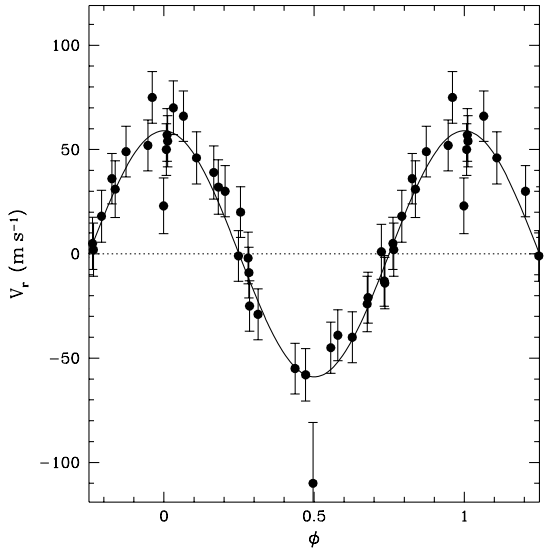

Fig. 9. Discovery plot of 51 Peg b, showing radial velocity (corrected for a slow variation of γ) as a function of orbital phase. The solid line represent an orbital fit in which the eccentricity was fixed at $e = 0$. From Mayor and Queloz (1995)

orbits, and of the characteristics of their parent stars. It is certainly necessary to keep selection effects related to the precision of the Doppler surveys and the limited time covered so far by systematic monitoring campaigns in mind, but most main-sequence stars of spectral type F or later within 50 pc and brighter than $V \approx 8$ have now been observed for a few years. This puts statistical inferences on a fairly firm footing, with the cautionary remark that our knowledge about planets around M dwarfs is still quite incomplete because of the intrinsic faintness of these stars.

The number of publications announcing new planet detections (and sometimes calling previous announcements into question) is growing rapidly; it is therefore quite a challenge to keep an authoritative list of the known extrasolar planets. Several groups maintain web pages with such lists, among them Geoff Marcy and co-workers, (`http://exoplanets.org`), Michel Mayor and colleagues (`http://obswww.unige.ch/exoplanets`), and Jean Schneider (`http://www.obspm.fr/planets`). The inclusion of a specific claimed planet detection in these lists is sometimes a matter of personal judgment, and the underlying philosophies are somewhat different. The International Astronomical Union has established a Working Group on Extrasolar Planets (Boss et al. 2003), which is about to publish a list with fairly conservative selection criteria (including publication in a refereed journal) at `http://www.ciw.edu/boss/IAU/div3/wgesp`. These compilations can serve as useful starting points for synoptic studies, and as guides to the original literature on planet detections.

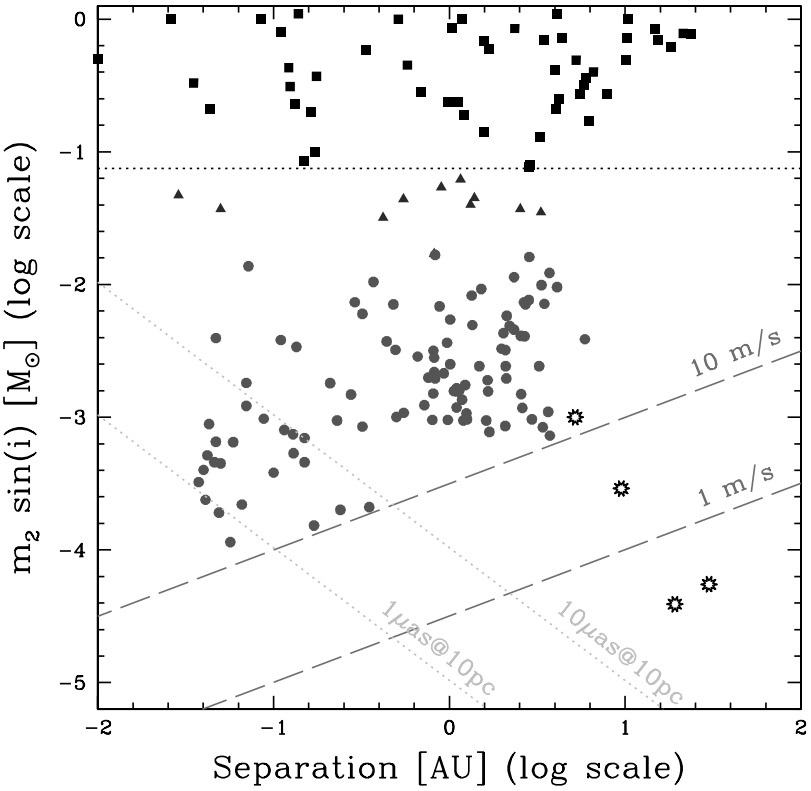

Fig. 10. Minimum mass $m_p \sin i$ versus orbital semi-major axis for the known extrasolar planets (*circles*) and low-mass stellar companions (*squares*). Hipparcos data have shown that most of the objects plotted as triangles are low-mass stars in nearly face-on orbits, i.e., their true mass is above the hydrogen burning limit ($0.075\,M_\odot$), shown as a dashed line (Halbwachs et al. 2000). The four giant planets in the Solar System are shown as stars. The detection limits of radial-velocity and astrometric surveys are also shown. Courtesy Stéphane Udry

The Parameter Space Covered by Planet Surveys

The Doppler technique has a strong detection bias in favor of massive planets in short-period orbits, because the velocity amplitude scales with $K_* \propto m_p \cdot P^{-1/3}$ (19), or equivalently $K_* \propto m_p \cdot a^{-1/2}$. With a single-measurement precision of $3\,\mathrm{m\,s^{-1}}$, planets giving rise to wobbles with $K_* \gtrsim 10\,\mathrm{m\,s^{-1}}$ can reliably be detected (Marcy et al. 2000a); this limit is shown in the mass-separation plane in Fig. 10.[12] The lowest-mass planet known, HD 49674 B.[13]

[12] With a sufficient number of observations it should also be possible to get orbits for planets with K_* close to the measurement precision (Cumming et al. 2002).

[13] Editor note added in proof: The lowest mass planet know at this time is HD160691 (Santos et al. 2004)

During the revision of the text (Butler et al. 2003), does indeed occupy the lower left corner of this diagram, just above the $10\,\mathrm{m\,s^{-1}}$ line. Its parameters are $m_p \sin i = 0.12\,M_{\mathrm{jup}}$, $P = 4.948$ days, $a = 0.057$ AU, and $K_* = 14\,\mathrm{m\,s^{-1}}$. The best sensitivity can of course only be reached for stars that are photospherically quiet, as discussed in Sect. 4.2. Planets have also been found around young, more active stars, but more data points are then required for a reliable orbital solution even for relatively large K_* (Kürster et al. 2000).

Another limiting factor in the parameter space covered so far is the limited time over which high-accuracy radial-velocity monitoring has been performed. It is obvious that the time needed from the first measurement to a secure detection is of the order of the orbital period; at high signal-to-noise perhaps half or a quarter of that time may be sufficient (Eisner and Kulkarni 2001a). Among all currently known planets, the one with the longest period, $P = 5,360$ days, and the largest orbital semi-major axis, $a = 5.9$ AU is 55 Cancri d (Marcy et al. 2002). It occupies an orbit with moderate ellipticity, $e = 0.16$, and weighs in at $m_p \sin i = 4.05\,M_{\mathrm{jup}}$, thus giving rise to a stellar reflex motion with amplitude $K_* = 49.3\,\mathrm{m\,s^{-1}}$. As its designation implies, 55 Cnc d is a member of a multiple system (see Sect. 3.5). It is apparent from Fig. 10 that 55 Cnc d is the only known extrasolar planet with an orbital semi-major axis larger than that of Jupiter; this is consistent with the ~ 15 years over which observations with $\lesssim 10\,\mathrm{m\,s^{-1}}$ have been performed.

The cutoffs at the bottom and to the right in Fig. 10 can thus easily be understood as selection effects due to the precision and time coverage limitations of the present radial-velocity data. It is important to point out that no such limitations exist towards the left and top in this diagram. The paucity of planets in these areas is a true astrophysical phenomenon, not an observational artifact; it will be further discussed in Sect. 3.2. One should further add that there is no very strong detection bias with regard to the other orbital elements, with the exception of the inclination. (Nearly face-on orbits are strongly disfavored, of course, because $K_* \propto \sin i$.) It is therefore a reasonable approximation to assume that the observed eccentricity distribution is representative of the intrinsic one. In fact, a large range of eccentricities have been observed, from circular orbits up to $e = 0.927$ (Naef et al. 2001a, see also Sect. 3.3).

It is certainly noteworthy that a large fraction of the accessible mass – separation – eccentricity parameter space is actually populated with giant planets (Figs. 10 and 13). This was not at all anticipated ten years ago, when it was thought that the orbits of gas giants should be similar to those of Jupiter and Saturn, with periods of several years and low eccentricities. The great diversity of giant extrasolar planets is the first big surprise in this field, which has stimulated a wealth of new ideas about the formation and evolution of planetary systems.

On the other hand, it is also clear from Figs. 10 and 13 that one cannot argue that our Solar System is special in any way. A few planets have actually been detected that are somewhat similar to Jupiter, which is located near

the corner of the currently accessible parameter space; Saturn analogs are simply out of reach at the present Doppler accuracy and monitoring time. It is therefore quite possible that many Jupiter/Saturn analogs will be discovered by future higher-accuracy Doppler surveys or other detection methods.

The Frequency of Giant Planets Around Solar-Type Stars

About 150 planet detections in a little less than 90 distinct systems, among $\sim 2,000$ stars surveyed, gives a lower limit of $\sim$5% of all nearby Solar-type stars that have planets in the detectable parameter range, i.e., with radial velocity variations $\gtrsim 10\,\mathrm{m\,s^{-1}}$ and separations $\lesssim 3\,\mathrm{AU}$. The best-studied sample comprises 51 stars that have been monitored at Lick Observatory over the past 15 years. Eight planetary systems (two of them triple, one double, and five single) have been detected among these 51 stars, corresponding to a $\sim 15\%$ detection rate (Fischer et al. 2003a). This number will probably rise somewhat over the next few years, as more planets close to the detection limits will be found. On the other hand, we also now that the majority of stars do not have companions in the currently observable range; about 60% of the stars observed by the Lick/Keck/AAT team have a radial velocity r.m.s. of $5\,\mathrm{m\,s^{-1}}$ or less (G. Marcy, priv. comm.).

Somewhat stronger statements can be made in those parts of the parameter space where the observational data give an essentially complete picture. First of all, brown dwarf companions in orbits of less than 3 AU are very rare, with an incidence well below 1% (Fischer et al. 2002b; Vogt et al. 2002). Second, it is relatively easy to identify "hot Jupiters", i.e., planets with orbital periods $\lesssim 5\,\mathrm{days}$. 11 such planets are known to date, which means that $\sim$0.5% of all stars in the Solar neighborhood are orbited by these objects.

3.2 Distribution of Masses and Orbital Radii

Planetary Masses

Figure 11 shows the histogram of the minimum mass $m \sin i$ for all currently known companions to Solar-type stars. Perhaps the most important conclusion that can immediately be drawn from this figure is that this distribution is bimodal, i.e., that there are two distinct populations of companion objects. This establishes "planets" as a physically distinct class of their own; they are not just the low-mass tail of the stellar binary population. The bimodal nature of the minimum mass distribution provides supporting evidence for the view that planets and stellar companions form through different mechanisms.

It has already been pointed out that the regimes of planets and stellar companions are separated by the "brown dwarf desert" $10\,M_{\mathrm{jup}} \lesssim m \sin i \lesssim 80\,M_{\mathrm{jup}}$. Only very few objects have been detected in this mass range, although they could easily be found by the radial-velocity surveys. In fact, most of the objects with $m \sin i$ in this mass range are actually stellar companions in nearly

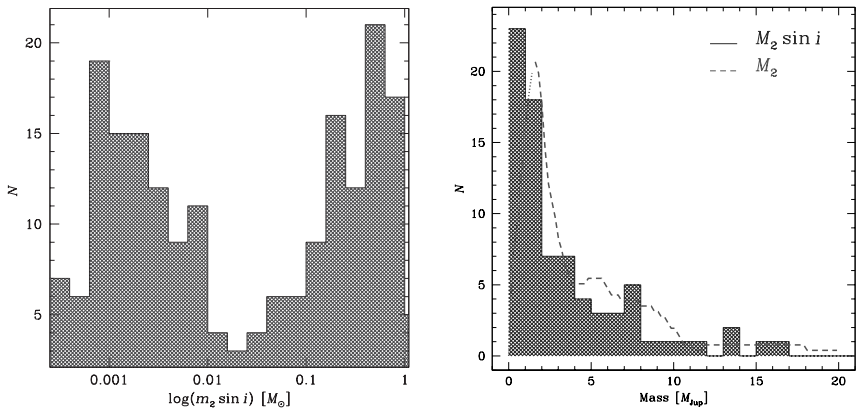

Fig. 11. *Left*: Distribution of minimum masses from the currently known low-mass companions to Solar-type stars. Although the radial-velocity method has a higher sensitivity to higher-mass companions, the observed distribution rises very steeply towards the low-mass domain. From $\sim 0.01\,M_\odot$ up to the stellar regime, only a few objects have been detected; this region is frequently called the "brown dwarf desert". This gap in the mass distribution of low-mass companions to Solar-type stars supports the view that there are two distinct populations (*planets and stars*), with different formation mechanisms. From Santos et al. (2002). *Right*: Same figure but with linear mass scale. The dashed line indicates a statistical estimate of the "true" planetary mass distribution. Updated from Jorissen et al. (2001)

face-on orbits; they produce an astrometric wobble sufficiently large to be detected with the Hipparcos satellite (Halbwachs et al. 2000). These objects are marked with triangles in Fig. 10, because their actual mass lies above the hydrogen burning limit. Conversely, the planet candidates with $m \sin i \lesssim 10\,M_{\rm jup}$ have usually not been detected with Hipparcos, which confirms that they are not stellar companions seen face-on (Zucker and Mazeh 2001a). The precision of Hipparcos is insufficient to rule out brown dwarf companions in these cases, but the a priori distribution of $\sin i$ (see Sect. 4.1) implies that the vast majority of the planet candidates really have masses below $10\,M_{\rm jup}$. The bottom line is that the brown dwarf desert is even less populated than it might appear from an $m \sin i$ histogram.

The small number of planets in the two lowest mass bins in Fig. 11 is clearly due to incompleteness close to the detection limit. Aside from this incompleteness, the number of planets rises somewhat towards lower masses when a logarithmic x-axis is chosen; this corresponds to a steep increase towards lower masses on a linear scale.

With a sufficiently large sample of detected planets, it is possible to disentangle the $\sin i$ projection effects from the observed $m \sin i$ histogram on a statistical basis, and to derive the underlying true mass distribution (Jorissen et al. 2001; Zucker and Mazeh 2001b). This is shown as a dashed line in Fig. 11. The sharp drop-off around $10\,M_{\rm jup}$, the low-level tail extending to

$\sim 20\, M_{\rm jup}$, and the steep rise towards lower masses are again immediately apparent. Zucker and Mazeh (2001b) derive a flat distribution for $dN/d\log m_p$, which corresponds to $dN/dm_p \propto m_p^{-1}$; Marcy et al. (2003a) give a similar scaling, $dN/dm_p \propto m_p^{-0.7}$, for $m_p < 8\, M_{\rm jup}$. It appears tempting to extrapolate these power laws to lower masses, in order to predict the numbers of Neptune-mass or even smaller planets. This could be misleading, however, because the formation of massive gas planets may be governed by physical processes that are different from those determining the frequency of lower-mass planet formation. Any extrapolation should therefore await a better understanding of the physics of planet formation.

Orbital Semi-Major Axes

With the first few detections of extrasolar planets it became rapidly apparent that gas giants occur with a large diversity of orbital separations, ranging from 0.05 AU (51 Peg b, Mayor and Queloz 1995), over 0.23 AU (ρ CrB B, Noyes et al. 1997), 0.43 AU (70 Vir B, Marcy and Butler 1996), 2.1 AU (47 UMa b, Butler and Marcy 1996) to 5.2 AU (Jupiter) and 9.6 AU (Saturn). The current statistical basis of ~ 150 extrasolar planets allows a much more detailed view of the distribution of orbits up to $a = 3$ AU. In Fig. 12 the minimum mass of the known planets is plotted as a function of their orbital semi-major axes, with a linear mass scale, which shows some trends and features more clearly than the logarithmic scale of Fig. 10:

- There is a remarkable "pile-up" of planets in orbits with $a \approx 0.05$ AU, i.e., with periods $P \approx 4$ days.

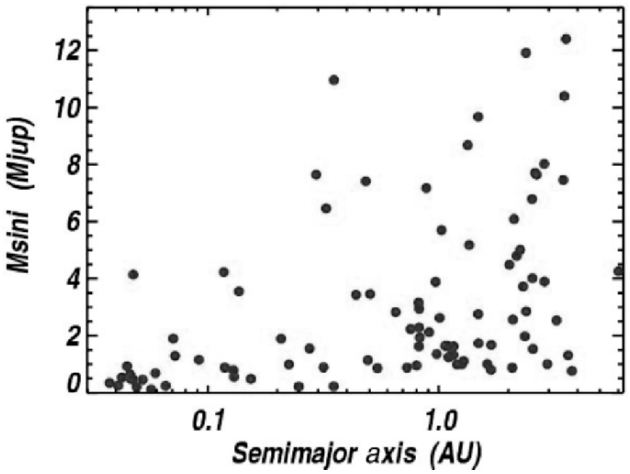

Fig. 12. Minimum mass $m_p \sin i$ versus semi-major axis for the known extrasolar planets, with a linear mass scale. From Marcy et al. (2003b)

- Orbits with $a \approx 0.3\,\mathrm{AU}$ are slightly less common than smaller and larger orbits.
- There are no planets with $m \sin i \geq 4.1\,M_{\mathrm{jup}}$ at $a \leq 0.3\,\mathrm{AU}$, whereas one third of all known planets at $a > 1\,\mathrm{AU}$ has a mass above this value.

As pointed out in Sect. 3.1, these findings cannot be due to an observational bias; they clearly tell us something about the formation and/or orbital evolution of giant planets.

Orbital Migration

Perhaps the most striking result from the Doppler surveys is the discovery that there are giant planets in orbits with $a < 3\,\mathrm{AU}$ at all. In fact, because our understanding of the Solar System implies that the gas giants were formed at $a > 5\,\mathrm{AU}$ and stayed there over the past 5 Gyrs, it was generally expected that this would also be the case in extrasolar systems. So could planets like 51 Peg b have formed much closer to their parent stars, at the location where they are found now? Given our general knowledge about star and planet formation (Sect. 2), there are several arguments why this appears exceedingly unlikely (Lin et al. 1996):

- At 0.05 AU the temperature in the pre-planetary disk is about 2,000 K, too hot for refractory materials to condense. Therefore planetary cores cannot form there.
- The surface density within $\sim 0.5\,\mathrm{AU}$ is too small for $\sim 10\,M_{\oplus}$ planetary cores to form.
- Even if a core is present (or if a core is not needed to form a gas giant), there is likely not enough gas to form a $\sim 1\,M_{\mathrm{jup}}$ planet.
- During their formation phase, young planets have radii up to ~ 10 times larger than their present values (Bodenheimer and Pollack 1986). Combined with the high temperature implied by small orbital distances, this gives very low escape speeds. The planet will thus be susceptible to evaporation, and to ablation by the stellar wind.

These arguments do not complete rule out the possibility of in situ formation at small radii (Bodenheimer et al. 2001), but the much more plausible conclusion is that the planets now found at small orbital radii must have formed at much larger distances, and subsequently migrated to their present locations.

Planet Survival and "Hot Jupiters"

The realization that orbital migration is an important mechanism for shaping planetary systems also poses a new question: why are the observed planets found in their present locations, i.e., why did their migration stop at the observed semi-major axes? Why did they not spiral all the way into the stellar photosphere? One possible mechanism involves the transfer of angular

momentum from the spinning star to the planet through tidal interaction. Young stars rotate fairly rapidly, so that the Keplerian period at $a \gtrsim 0.05\,\mathrm{AU}$ is longer than the stellar rotation period. In that case the tides raised on the star exert an outward torque on the planet; the time scale of the consequent orbital evolution is (Goldreich and Soter 1966; Lin et al. 1996)

$$\tau_a \equiv -\frac{a}{\dot{a}} = \frac{P}{9\pi}\left(\frac{a}{R_*}\right)^5 \frac{m_*}{m_p} Q_* \ , \tag{7}$$

where $Q_* \approx 1.5{\cdot}10^5$ is a parameter, which describes how efficiently the tidal energy is dissipated within the star. It is apparent from (7) that the tidal torque depends very sensitively on a. The tidal interaction therefore sets in very suddenly during the inward migration; this could be a possible reason for the "pile-up" of the observed orbits at $\sim$0.05 AU. An alternative explanation might be the truncation of the disk by the stellar magnetosphere, which could also occur at a similar radius (Shu et al. 1994).

Determining the fate of close-in planets is further complicated by the concurrent evolution of the star and the disk. Due to the spin-down of the star through torquing by the wind, the rotation period will eventually become longer than the orbital period of hot Jupiters. This reverses the sign of the orbital evolution, and may cause the planet to spiral into the star. The time scale for this process is $\propto m_p^{-1}$ (7), which could explain the absence of massive planets in short-period orbits (Pätzold and Rauer 2002; Jiang et al. 2003). In a computation of the orbital and structural evolution of planets that took into account Roche lobe overflow and consequent mass loss of the planet, three classes of objects could be identified: the planets with the lowest initial masses were completely destroyed, intermediate-mass planets lost some of their mass and ended up in stable orbits at $\sim$0.04 AU, and the most massive planets did not migrate very far (Trilling et al. 1998). A combination of migration, mass loss, and the opening of gaps in the disk is thus likely needed to explain the observed minimum mass – semi-major axis distribution (Zucker and Mazeh 2002; Udry et al. 2003b).

Finally, one can pose the question what consequence the migration of giant planets has for the formation of terrestrial planets. A migrating gas giant would certainly destroy already existing small planets, and it would probably also suppress the subsequent formation of terrestrial planets at $\sim$1 AU (Armitage 2003). This suppression occurs because the gas that flows into the inner disk behind the migrating planet is depleted of dust as a result of having already formed planetesimals at larger radii. It is thus likely that the systems with hot Jupiters were not able to form terrestrial planets, even if there are dynamically stable orbits for habitable planets in those systems.

Planets with Long Periods

Although the radial-velocity technique selects against planets with a long orbital period ($K_* \propto P^{-1/3}$), the sensitivity of the current surveys is clearly

good enough to detect Jupiter analogs (see Fig. 10). Nonetheless, no such object has been found with certainty yet, and it is hardly possible to make statistical inferences about the occurrence of planets with periods $\gtrsim 5\,\mathrm{yr}$. This situation will likely change dramatically within a few years, when the surveys covering a large number of stars with high precision, which were initiated or enlarged after the detection of 51 Peg b, will reach a sufficient time baseline. The typical orbital period of the newly announced extrasolar planets is already shifting towards longer periods, reflecting the increasing time baseline of the ongoing Doppler surveys. But most of the confirmed long-period planets that are known today still have relatively high masses (Santos et al. 2001b; Fischer et al. 2002b; Marcy et al. 2002), with the exception of 47 UMa c, the outer member of a system with two planets (Fischer et al. 2002a).

Planets with long periods first tend to show up as linear trends in radial-velocity data; deviations from linearity become apparent with continued monitoring, until finally the survey covers a full Keplerian period (e.g., Naef et al. 2001b). Most planet search teams will wait until roughly this time before announcing a planet detection and publishing an orbital solution. Important statistical information can be gleaned earlier, however. For example, radial-velocity variations indicative of distant planets appear to be significantly more common for stars that are already known to harbor an inner planet than for single stars (Fischer et al. 2001). Again, only time will tell how the properties of inner and outer planets are related to each other.

3.3 Orbital Eccentricities

The second big surprise (after the discovery of the "hot Jupiter" 51 Peg b) was the highly eccentric orbit of 70 Vir b, with $e = 0.40$ (Marcy and Butler 1996). In contrast, Jupiter and Saturn have $e \approx 0.05$, and even Mercury's ($e = 0.21$) and Pluto's ($e = 0.25$) orbital eccentricities are modest in comparison. It had generally been expected that extrasolar planets would also be found in nearly circular orbits, because they form in a circumstellar disk, and dissipation in that disk should generally lead to the circularization of the orbits. Even more extreme examples have subsequently been found, such as HD 89744 b with $e = 0.7$ (Korzennik et al. 2000) and HD 80606 b with an astonishing $e = 0.927$ (Naef et al. 2001a).

The radial-velocity curves of eccentric orbits deviate strongly from the sinusoidal variations associated with circular orbits. This characteristic shape makes it easy to distinguish planetary companions from other sources of radial-velocity jitter, as in the case of ι Dra (Frink et al. 2002). This may introduce a slight bias towards higher eccentricities in planet samples, but the effect is probably rather insignificant.

A plot of orbital eccentricity versus period for the known extrasolar planets is shown in Fig. 13, together with the same quantities for five of the Solar-System planets and stellar binaries. A few properties of the distribution of orbital parameters are worth noting:

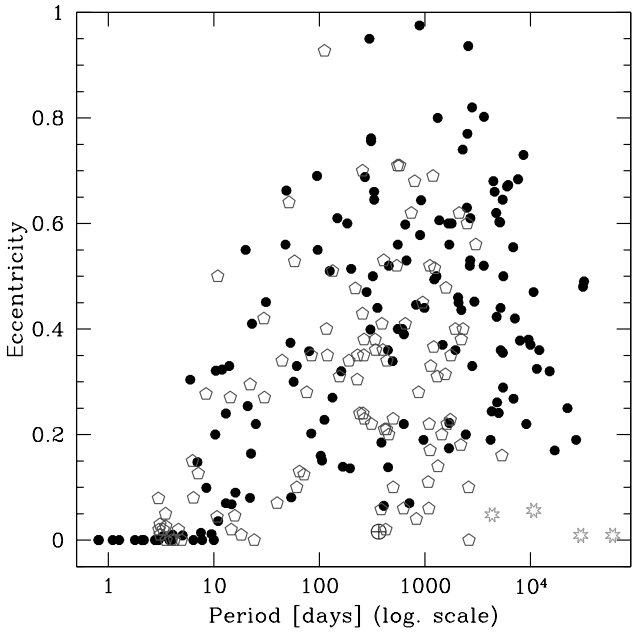

Fig. 13. Eccentricity versus orbital period for the known extrasolar planets (*pentagons*), stellar binaries (*filled dots*), giant planets in the Solar System (*stars*), and the Earth. From Santos et al. (2003b)

- All planets with short-period orbits have small eccentricities ($e \lesssim 0.1$ for $p \lesssim 10\,\mathrm{days}$).
- For longer periods, larger eccentricities are fairly common.
- The eccentricity–period diagrams for planets and stellar binaries look remarkably similar.
- There is also a class of long-period planets with nearly circular orbits (see also Vogt et al. 2002).
- The giant planets in the Solar System have small eccentricities, but they are not unusual.

To interpret these observations correctly, we have to consider not only the original orbit, but also the evolution of the eccentricities after the formation of the planet. This is a new insight that could only come from data on extrasolar planets and not from the Solar System, quite analogous to the case of the evolution of the major axes due to migration.

Tidal Circularization

The small eccentricities of the short-period orbits are generally attributed to *tidal circularization*, due to dissipation within the planet. The tidal bulges

raised on a planet in an eccentric orbit lead to a decay of the eccentricity on a time scale (Goldreich and Soter 1966; Bodenheimer et al. 2001)

$$\tau_e \equiv -\frac{e}{\dot{e}} = \frac{4Q_p}{63n}\,\frac{m_p}{m_*}\left(\frac{a}{R_p}\right)^5 = \frac{Q_p}{10^6}\,\frac{m_p}{M_{\rm jup}}\left(\frac{M_\odot}{m_*}\right)^{3/2}\left(\frac{a}{0.05\,{\rm AU}}\right)^{13/2}\left(\frac{R_{\rm jup}}{R_p}\right)^5 {\rm Gyr}\ , \qquad (8)$$

where Q_p is the tidal dissipation parameter for the planet (analogous to Q_* defined above), and $n = 2\pi/P$ the mean motion of the planet. Very little is known about plausible values for Q_p, but indirect arguments indicate that $10^5 \lesssim Q_p \lesssim 10^6$ for Jupiter (Goldreich and Soter 1966). Assuming that extrasolar gas giant planets possess similar Q_p values (which is a bit of a wild guess, see Marcy et al. 1997), the orbits of old hot Jupiters should be circularized according to (8). This equation also shows that the circularization time scale depends steeply on the orbital distance, both directly ($\tau_e \propto a^{13/2}$) and indirectly, because strongly irradiated planets are somewhat bloated, which leads to a further shortening of τ_e at small a. The tidal circularization scenario provides thus a fairly natural explanation of steep decrease in the upper envelope for e at $P \lesssim 10\,$days.

Origin of Eccentric Orbits

The poor efficiency of tidal circularization for longer-period orbits is a necessary, but clearly not a sufficient condition for the existence of planets with high-eccentricity orbits. A rather common mechanism (or even several mechanisms) that can induce large values of e either during the formation of planets or thereafter through interactions is obviously needed to explain the observations (Fig. 13). A special case are planets orbiting a component of a wide stellar binary, such as the companion of 16 Cyg B (Cochran et al. 1997). Such systems can oscillate between high- and low-eccentricity states, if the inclination of the orbital plane of the planet with respect to that of the stellar binary is appreciable (Kozai 1962).

For those planets that are not found in wide stellar pairs (which is the large majority), gravitational interaction of planets in multiple systems is the most plausible way to generate a large value of e. If several giant planets form in a massive disk, their mutual perturbations induces a gradual increase in their orbital eccentricities (Lin and Ida 1997). The orbits may eventually become unstable and cross each other, so that several planets can merge and form a very massive planet, which tends to end up in an orbit with high eccentricity ($0.2 \lesssim e \lesssim 0.9$) and relatively small semi-major axis ($0.5\,{\rm AU} \lesssim a \lesssim 1\,{\rm AU}$). Alternatively, if a large number of massive planets form nearly simultaneously through fragmentation in a disk or protostellar envelope, dynamical relaxation leads to the ejection of most of the planets, while the remaining ones end up in highly eccentric orbits (Papaloizou and Terquem 2001). The distribution of

eccentricities (and of the orbital inclination with respect to the stellar equator, and of the mutual inclinations of orbits in multiple systems) thus encapsulate information on the dynamical history of planetary systems, which are in turn directly related to the formation mechanism.

3.4 Properties of the Parent Stars

In addition to the properties of the known planets themselves, we may ask the questions whether there is a direct relation between these properties and the characteristics of the host stars, and whether the population of stars harboring planets is in any way different from their parent population of stars in the Solar neighborhood.

Stellar Metallicities

The most striking observation related to the properties of planet host stars is that they are more metal-rich on average than comparison stars without known planets in the Solar neighborhood (see Fig. 14). Another way of stating this is saying that the probability of finding a giant planet increases with metallicity, as seen clearly in the right panel of Fig. 14. This is an interesting physical relationship between planets and their parent stars.

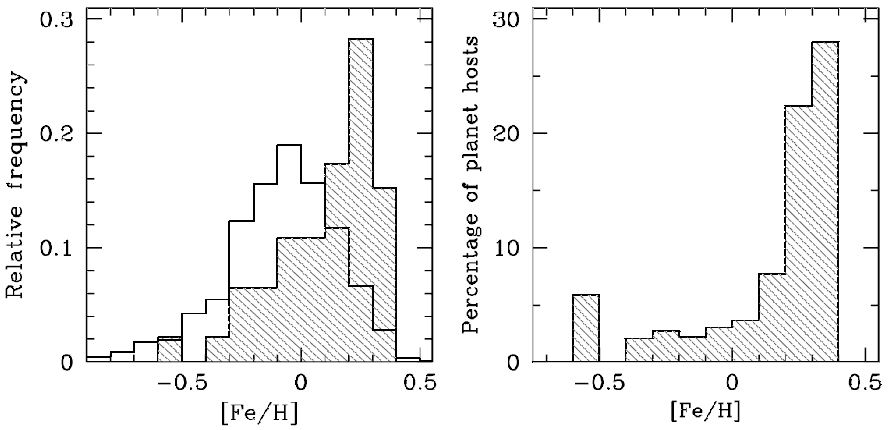

Fig. 14. *Left panel*: Metallicity (i.e., [Fe/H]) distributions for stars with planets (*shaded histogram*) compared with the same distribution for field dwarfs in the Solar neighborhood (*open histogram*). In this panel, both distributions are normalized by the respective number of data points. Most planet hosts are more metal rich than our Sun. *Right panel*: The percentage of stars that have been found to harbor a planet, for each metallicity bin. This plot shows clearly that the probability of finding a giant planet increases with metallicity. Updated from Santos et al. (2001a)

In establishing the metallicity–planet correlation, one has to be a bit careful about selection biases (Butler et al. 2000). In this context it is important that the targets of the large radial-velocity surveys have not been pre-selected on the basis of metallicity; stellar type (FGKM dwarf) and apparent brightness have indeed been the main selection criteria. Nonetheless, the rate of planet detections might be modified by some more subtle metallicity-dependent effects. Metal-rich stars have deeper absorption lines, which improves the attainable Doppler precision. Furthermore, metal-rich stars are brighter than metal-poor stars of the same spectral type, which leads to a Malmquist-type bias that could enrich magnitude-limited samples with metal-rich stars. However, while present to a certain degree, these selection effects are too small to explain the observed metallicity enhancement among planet host stars. A number of studies with somewhat different approaches therefore all come to the conclusion that the metallicity–planet correlation is a well-established physical fact (Gonzalez et al. 2001b; Santos et al. 2001a, 2003a; Reid 2002).

Why are Planets Found Preferentially around Metal-Rich Stars?

Two main hypotheses have been advanced to explain the enhanced metallicity of planet host stars: either planets form preferentially in the disks of metal-rich stars, or the atmospheres of planet-bearing stars have been polluted with high-Z material, perhaps from planets that have migrated all the way into the star, or from planetesimals scattered into star-impacting orbits by migrating planets (Quillen and Holman 2000). In the first scenario, the higher metallicity is the cause of the enhanced occurrence of planets, in the latter scenario, the presence of planets causes an enhanced metallicity. Several tests have been devised to distinguish between these two hypotheses, with somewhat mixed results.

One test consists of comparing the metallicity enhancement separately for stars with different kinematic properties (Barbieri and Gratton 2002). If high metallicity is the cause for the presence of planets, there should be no correlation between the occurrence of planets and galactocentric distance. It is only the overall metallicity that is important, and within each metallicity bin the distribution of stars with perigalactic distance should be the same for stars with and without planets. If on the other hand the presence of planets causes enhanced metallicities, one should expect that at any galactocentric distance planet host stars should be more metal-rich than average, and in each metallicity bin, stars with planets should tend to have smaller perigalactic distances. Barbieri and Gratton (2002) find precisely this effect from a reconstruction of Galactic orbits of planet hosts and comparison stars without planets, and conclude that scenarios in which the presence of planets is the cause of higher metallicities are strongly favored.

A different class of tests is based on more direct attempts to determine the effect of planet engulfment on the atmospheric composition of the parent star. Hydrodynamic simulations show that Jupiter-like planets spiraling into

stars with $1.0\,M_\odot \leq m_* \leq 1.3\,M_\odot$ are partially or totally dissolved within the convection zone, and can thus indeed enhance the metallicity significantly (Sandquist et al. 1998). This would in particular deposit the isotope ^{6}Li in the atmosphere, which is normally destroyed during the early phases of stellar evolution.[14] The detection of ^{6}Li in the planet host star HD 82943 has therefore been interpreted as evidence for planet engulfment (Israelian et al. 2001).

The strongest argument for the alternative explanation, i.e., that planet engulfment does not play a dominant role, comes from an analysis of metallicity with stellar temperature. More massive stars (within the range of interest) have much shallower convective envelopes (Murray et al. 2001); adding a given amount of planetary material should thus have a much stronger effect on their surface composition. No such trend of metallicity with mass of the convection zone is observed, which means that a "primordial" source of the metallicity excess is much more likely (Santos et al. 2001a, 2003a). The "standard" model of giant planet formation through planetesimal formation and runaway gas accretion actually predicts that low-metallicity systems are much less likely to form planets than their high-metallicity counterparts, because the surface density of solid material in the pre-planetary disk plays a critical role for planet formation (Youdin and Shu 2002). One could thus even argue that the enhanced metallicities of planet hosts provide an empirical argument for this planet formation scenario over the disk instability model, in which no strong dependence on metallicity is expected (Boss 2002).

With arguments for either interpretation, the question about the cause of the planet–metallicity correlation has not been completely settled yet. It is certainly possible, that accretion of iron-rich material, high primordial metallicity, and selection effects all play a certain role (Murray and Chaboyer 2002).

Stellar Rotation Rates

Among the effects of tidal interaction between the star and planet in a small orbit is a spin-up of the star, which ultimately leads to synchronization of the stellar rotation rate with the orbital motion of the planet. The time scale of this spin-up is given by (Goldreich and Soter 1966; Trilling 2000)

$$\begin{aligned}\tau_s \equiv -\frac{\omega}{\dot{\omega}} &= Q_*\,\omega \left(\frac{R_*^3}{G m_*}\right) \left(\frac{m_*}{m_p}\right)^2 \left(\frac{a}{R_*}\right)^6 \\ &= 13\,\frac{Q_*}{10^5} \cdot \frac{\omega}{2 \cdot 10^{-6}\,\mathrm{s}^{-1}} \cdot \frac{m_*}{1.05 M_\odot} \times \\ &\quad \times \left(\frac{1.2 R_\odot}{R_*}\right)^3 \left(\frac{0.45 M_{\mathrm{jup}}}{m_p}\right)^2 \left(\frac{a}{0.051\,\mathrm{AU}}\right)^6 \mathrm{Gyr}\ ,\end{aligned} \tag{9}$$

[14] ^{6}Li is destroyed at relatively low temperatures ($\sim 1.6 \cdot 10^6$ K) through (p,α) reactions. During the early evolutionary phases, the proto-star is completely convective, so that the cool surface material is mixed with the hot stellar interior, which leads to the destruction of all ^{6}Li.

where ω is the angular velocity of the stellar rotation, and Q_* the tidal dissipation parameter of the star defined already in the context of (7). The scaling parameters in the second line of (10) have been chosen to match estimates for the 51 Peg system; the time scale for rotational synchronization is in that case somewhat longer (by a factor of a few) than the age of the system. Among the known "hot Jupiters" τ Boo has by far the shortest synchronization time ($\tau_s \approx 0.8$ Gyr, Trilling 2000). This star is the also the only one in a sample of planet hosts for which the rotation period is almost identical to the orbital period of the planet (Barnes 2001). This strongly suggests that tidal spin-up has indeed occurred in the case of τ Boo.

Planets in Multiple Stellar Systems

To explore the factors that influence the formation and evolution of planetary systems, one should clearly look for planets in diverse environments, i.e., around as large a variety of parent stars as possible. In this respect, binaries and multiple stellar systems provide rich opportunities, because in such systems the stability of pre-planetary disks and of planet orbits is restricted to limited distance ranges. In a binary system with component separation a_B, stable regions exist at small radii around each component (up to $\lesssim 0.3 a_B$, depending on the mass ratio), and around the binary system at radii $\gtrsim 2a_B$. The statistics of planets in binaries could therefore hold important clues to issues related to the formation and migration process. Furthermore, the relative orientation of the stellar rotation axes, the binary orbit, and the planetary orbits could help us understand the processes that govern the distribution of angular momentum during the early phases of star and planet formation. While the mass of the Solar System is dominated by the Sun, most of its angular momentum resides in the orbital angular momenta of the planets – it would certainly be interesting to investigate the angular momentum distribution in more complicated systems.

The first planet orbiting a component of a wide binary was found around 16 Cyg B; it has already been mentioned that in this case the stellar companion might be responsible for pumping up the eccentricity of planetary orbit (Cochran et al. 1997). Meanwhile it has been discovered that 16 Cyg A is a binary with separation $3\farcs4$ itself (Patience et al. 2002), and another planet has been found in a similar hierarchical triple system (HD 178911, Zucker et al. 2002). The statistics have been improved both by searching for planets in known stellar binaries, and by searching for stellar companions to known planet hosts with speckle and adaptive optics techniques (Patience et al. 2002; Luhman and Jayawardhana 2002). Progress has also been made on the theoretical front; for example, numerical simulations indicate that terrestrial planets can actually form in systems like α Cen (Barbieri et al. 2002; Quintana et al. 2002). Much needs still to be done, however, before we can link the potential information content of planets in binary systems to the pressing current questions about the physical processes that shape planetary systems in general.

3.5 Systems with Multiple Planets

The Zoo of Planetary Systems

The story of extrasolar planetary systems – in the sense of multiple planets orbiting one and the same star – began with the discovery of a second and third planet around υ And (Butler et al. 1999). Because of the analogy with the Solar System we might expect that different types of planets (Earth-like, gas, and ice giants) might form and evolve together. We can study the general architecture of multiple systems and ask questions such as: What is the spread in the planetary masses? Do the masses increase from the inside out, or the other way round? Are the orbits (nearly) coplanar? In addition, the gravitational interaction between the planets can be so strong that orbital resonances play a dominant role for their dynamical behavior (see Sect. 3.6), providing a rich new field of investigation.

The most important parameters of the ten known multiple systems are summarized in Table 5. (Note that the star HD 83443 was also believed to harbor two planets, but recent measurements demonstrated that the existence of the outer planet is not firmly established (Butler et al. 2002) Two stars (υ And and 55 Cnc) are known to harbor three planets each; pairs of two planets have been detected in the other 8 systems. The notation (b, c, d, ...) follows the sequence in which the planets have been discovered; there is no specific ordering with respect to semi-major axis or mass.

Since all planets in Table 5 have been discovered with radial-velocity measurements, it is clear that they are drawn from the part of the parameter space accessible to this method (see Sect. 3.1). In addition to the general selection bias favoring massive planets and short-period orbits, this means that the present sample should be biased towards systems in which the masses increasing with the orbital semi-major axes. A good example is υ And, in which all three planets give rise to roughly equal radial-velocity amplitudes K, because the loss of sensitivity with a is compensated by the increase in m_p (see Table 5 and Fig. 15). A system with masses decreasing from the inside out would be much harder to detect, because the outermost planet would be hard to detect by virtue of its low mass combined with large a.

Another shortcoming of the radial-velocity method is that it does not provide any information about the inclinations of the planets' orbits. Knowing these would be important for several reasons:

- If the individual planets in a system have different inclinations, we may misinterpret the overall architecture of the system. (The planet with the largest $m_p \sin i$ is not necessarily the one with the largest mass.)
- The relative inclination of the orbits is an important diagnostic for the dynamical evolution of the system, see Sect. 3.3.
- The gravitational interaction between a pair of planets depends on their masses. Knowing these only modulo factors $\sin i$ limits the ability to model perturbations of the orbits due to these interactions.

Table 5. Parameters of multiple planetary systems, derived from Keplerian fits to the radial-velocity curves

star	m_* $[M_\odot]$	P [days]	K $[\mathrm{m\,s^{-1}}]$	e	$m_p \sin i$ $[M_{\mathrm{jup}}]$	a [AU]	comment
υ And b		4.617	70.15	0.01	0.64	0.058	
υ And c	1.30	241.16	53.93	0.27	1.79	0.805	apsidal lock
υ And d		1276.15	60.62	0.25	3.52	2.543	
55 Cnc b		14.653	71.5	0.03	0.83	0.115	3:1 resonance
55 Cnc c	1.03	44.3	11.2	0.40	0.18	0.241	
55 Cnc d		4400	50.2	0.34	3.69	5.2	
GJ 876 b	0.32	61.020	210.0	0.10	1.89	0.207	2:1 mean motion and secular resonance
GJ 876 C		30.120	81.0	0.27	0.56	0.130	
47 UMa b	1.03	1079.2	55.6	0.05	2.86	2.077	7:3 resonance?
47 UMa c		2845.0	15.7	0.00	1.09	3.968	
HD 168443 b	1.01	58.1	470.0	0.53	7.64	0.295	
HD 168443 c		1770.0	289.0	0.20	16.96	2.873	
HD 37124 b	0.91	153.3	35.0	0.10	0.86	0.543	
HD 37124 c		1942.0	19.0	0.60	1.00	2.952	
HD 12661 b	1.07	263.6	74.4	0.35	2.30	0.823	secular resonance
HD 12661 c		1444.5	27.6	0.20	1.57	2.557	
HD 38529 b	1.39	14.309	54.2	0.29	0.78	0.129	
HD 38529 c		2174.3	170.5	0.36	12.7	3.68	
HD 82943 b	1.05	444.6	46.0	0.41	1.63	1.159	2:1 resonance
HD 82943 c		221.6	34.0	0.54	0.88	0.728	
HD 74156 b	1.05	51.6	112.0	0.65	1.61	0.278	
HD 74156 c		> 2650	125.0	0.35	> 8.21	> 3.82	

From Marcy et al. (2003b)

It will thus be very important to develop techniques that can measure these inclinations, such as astrometry (see Sect. 9). For the moment, keeping the limitations of the Doppler method in mind, we can still learn a lot from the available data.

Individual Systems

In the following we will take a closer look at each of the systems from Table 5 in turn, highlighting some of the important features and peculiarities. This will provide the observational backdrop for the discussion of gravitational interactions and dynamical resonances in Sect. 3.6.

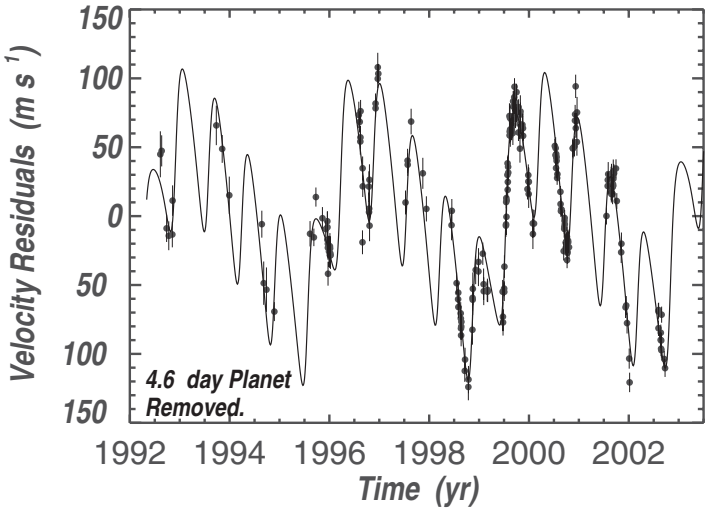

Fig. 15. Lick Observatory residual velocities for υ And after removal of the Keplerian wobble caused by the inner companion, using best-fit orbital parameters of $P = 4.6171$ days and $K = 75\,\mathrm{m\,s^{-1}}$. Two time scales are apparent in the residuals at 3 and 0.7 yr. The solid line shows the theoretical velocity curve caused by the outer two companions. Updated from Butler et al. (1999)

υ Andromedae.

Among the early results from the radial-velocity survey at Lick Observatory was the discovery of a planet orbiting the F8V star υ And with a 4.6-day period (Butler et al. 1997). The real claim to fame for this star came somewhat later with the identification of two additional companions with orbital periods of 241 and $\sim 1,280$ days (Butler et al. 1999, see Figs. 15 and 16). Since this original detection, the observed radial velocities have followed the predictions from a triple Keplerian fit without any indication for gravitational interaction between the planets, or for a fourth planet in the system (Marcy et al. 2003b). However, the apsidal lines of the outer two orbits are nearly aligned with each other (entries for ω in Table 6), hinting at a secular resonance involving these planets (see Sect. 3.6).

As demonstrated in Fig. 16, there are three massive planets in the υ And system in a volume that in the Solar System is populated only by the much smaller terrestrial planets. It is thus not surprising that gravitational interaction between the planets plays a much more significant role systems such as υ And than in the Solar System. Since these interactions depend directly on the masses of the planets, it would be interesting to get limits on their orbital inclinations. The astrometric signature expected from the outermost planet, υ And D, is just at the threshold of being detectable at the precision of the

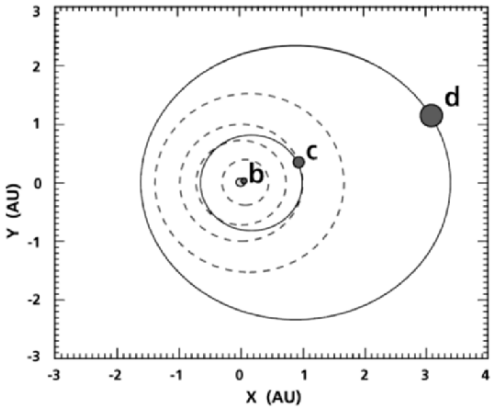

Fig. 16. Orbits of υ And b, c, and d, compared to the inner Solar System. The orbits of the three planets are shown co-planar and face-on; the actual relative orbital inclinations are not known, however. Because of the existence of multiple massive planets close to each other, dynamical interactions are much more important in the υ And system than in the Solar System

Table 6. Parameters of the planets of υ Andromedae, derived from Keplerian fits to the radial-velocity curve

parameter	planet B	planet C	planet D
P [days]	4.6170	241.5	1284
T [JD - 2,450,000]	2.093	160.5	64
e	0.012	0.28	0.27
ω [deg]	73	250	260
K_* [$\mathrm{m\,s^{-1}}$]	70.2	53.9	61.1
a_p [AU]	0.059	0.83	2.53
$m_p \sin i$ [M_{jup}]	0.69	1.89	3.75

Updated from Butler et al. (1999)

Hipparcos data (see also Sect. 9.2). The χ^2 of the Hipparcos measurements is indeed minimized by $m_p = 10.1\,M_{\mathrm{jup}}$, but a mass as low as $4.1\,M_{\mathrm{jup}}$ (which corresponds to $i = 90°$) or as high as $19.6\,M_{\mathrm{jup}}$ would also be allowed at the $2\,\sigma$ level (Mazeh et al. 1999).

55 Cancri.

The same paper that announced the discovery of the first planet around υ And also reported a planet orbiting the star 55 Cnc with a period of 14.65 days (Butler et al. 1997). The velocity residuals after subtracting the best-fit

Keplerian orbit showed a clear long-term trend, suggesting a possible second companion. Continued monitoring of 55 Cnc did indeed reveal two additional periodicities, with $P = 44.3$ d and $P = 12$ yr, respectively (Marcy et al. 2002). While the latter can clearly be attributed to another planet in the system, the former is close to the rotation period of the star, and might thus be caused by surface inhomogeneities. There are a number of arguments against the rotational modulation hypothesis (including the requirement that the surface structure would have to be stable over at least 14 years), and thus it seems likely that there are indeed three planets around 55 Cnc, with a 3:1 orbital resonance between the inner two.

Significant excess emission at 60 μm was observed towards 55 Cnc with the Infrared Space Observatory (ISO), suggesting that this star may be a Vega-excess object (Dominik et al. 1998). This interpretation gained support by the putative detection of a scattered-light disk in ground-based near-infrared observations (Trilling and Brown 1998). Observations with the NICMOS instrument on the Hubble Space Telescope failed to detect this disk, however, indicating that the ground-based result is probably spurious (Schneider et al. 2001). From a non-detection at $\lambda = 850$ μm, Jayawardhana et al. (2002) obtain an upper limit of less than $10^{-3}\,M_\oplus$ in small dust grains associated with 55 Cnc; they suggest that the 60 μm excess results from a nearby submillimeter source within the ISO beam.

Gliese 876.

A planet with $m_p \sin i = 2\,M_{\rm jup}$ in a 60-day orbit around the M4 dwarf star GJ 876 was independently discovered by the Swiss and Californian planet search teams (Delfosse et al. 1998; Marcy et al. 1998). This planet is remarkable because of the low mass of its host star, which offers interesting opportunities for follow-up observations. The astrometric wobble of GJ 876 caused by the gravitational pull of this planet has indeed been detected with the Fine Guidance Sensors on the Hubble Space Telescope; this marks the first secure astrometric detection of an extrasolar planet (Benedict et al. 2002, see also Sect. 9.2). The inferred inclination $i = 84°$ implies that the mass of the planet m_p is close to its minimum mass $m_p \sin i$.

Continued observations of GJ 876 soon revealed that the radial-velocity data could not be modeled with a single Keplerian orbit; a second planet with a period of 30 days is needed to obtain a satisfactory fit (Marcy et al. 2001a, see Fig. 17). The two orbits have a 2:1 period ratio, and their axes appear to be nearly aligned (see Table 7); this is strong evidence that the two planets are locked in an orbital resonance. Taking the planet–planet interaction into account actually improves the χ^2 of a fit to the radial-velocity data considerably compared to a fit with two Keplerians (Marcy et al. 2003b).

47 UMa.

The two planets orbiting 47 UMa (Butler et al. 1996; Fischer et al. 2002a) have nearly circular orbits, with periods that are not close to any small-integer

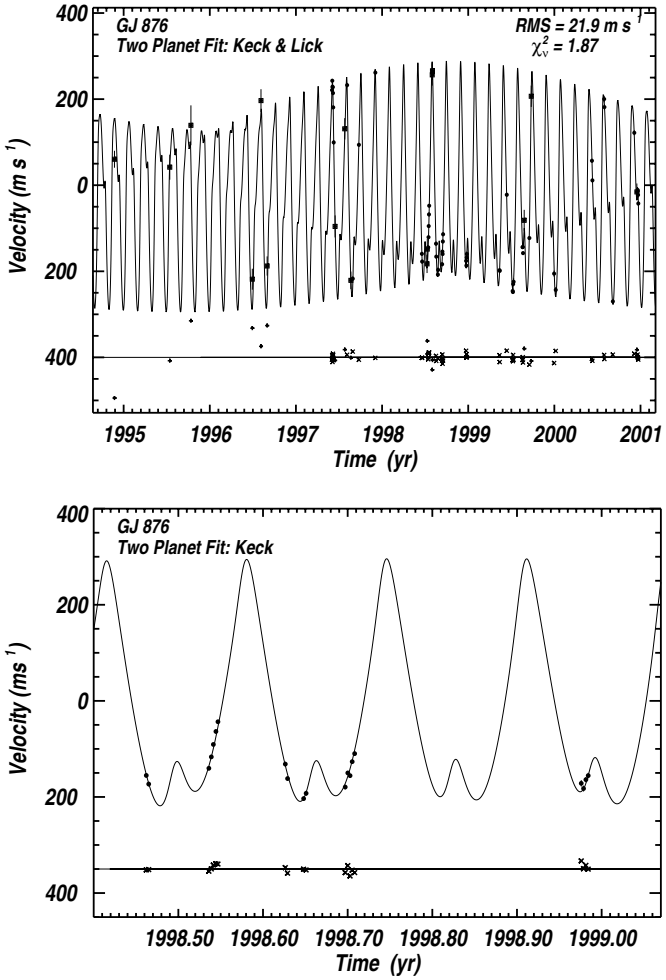

Fig. 17. *Top*: Combined Lick and Keck Observatory velocities for GJ 876, fitted with a model containing two planets in Keplerian orbits. The filled circles represent Keck velocities, and filled squares represent those from Lick. Residuals are shown below the radial-velocity curve. *Bottom*: Zoom on the time interval between 1998.4 and 1999.1, with velocities from Keck. The inflections in the velocities are due to the "beating" between the two planets. From Marcy et al. (2001a)

ratio. There is some resemblance to Jupiter and Saturn, which have similar period and mass ratios as the 47 UMa planets. Modeling the formation of the two planets in the 47 UMa system within the core accretion – gas capture model leads to the conclusion that they can both have formed through this mechanism within ~ 3 Myr at their present distances from the star (Kornet et al. 2002).

Table 7. Parameters of the two planets orbiting GJ 876, derived from Keplerian fits to the radial-velocity curve

parameter	inner planet	outer planet
P [days]	30.12 ± 0.02	61.02 ± 0.03
T [JD - 2,450,000]	31.4 ± 1.2	106.2 ± 1.9
e	0.27 ± 0.04	0.10 ± 0.02
ω [deg]	330 ± 12	333 ± 12
K_* [$\mathrm{m\,s^{-1}}$]	81 ± 5	210 ± 5
$a_* \sin i$ [AU]	0.00022	0.00117
a_p [AU]	0.130	0.208
$m_p \sin i$ [M_{jup}]	0.56 ± 0.09	1.89 ± 0.3

From Marcy et al. (2001a)

HD 37124.

Systems like HD 37124 (Vogt et al. 2000; Marcy et al. 2003b) have be called "hierarchical", because the two planets have widely spaced orbits, which makes them structurally and dynamically reminiscent of hierarchical stellar systems. A slightly puzzling aspect of the HD 37124 system is the large eccentricity of the outer planet, whose origin is unknown.

HD 12661.

The planets orbiting HD 12661 have periods of 260 and 1440 d (Fischer et al. 2001, 2003b). Although this means that this system is also hierarchical ($P_1/P_2 = 5.48$), it has been argued that it resides in a secular resonance (Lee and Peale 2003).

HD 168443.

The HD 168443 system is interesting because it contains two very massive companions, with $m_p \sin i = 7.7\, M_{\mathrm{jup}}$ and $m_p \sin i = 17\, M\mathrm{jup}$, respectively (Marcy et al. 2001b; Udry et al. 2002). Beyond the somewhat irrelevant question whether these objects should legitimately be called "planets" or something else, one may ask whether they likely formed in the same way as their lower-mass analogs. If one is willing to speculate that some planets form through core accretion and gas capture, while others form through gravitational instabilities in a gas disk, the planets around HD 168443 would be prime candidates for the second process.

HD 38529.

With periods of 14.3 d and 6.0 yr, the two planets of HD 38529 form an extremely hierarchical system (Fischer et al. 2001, 2003b). The host star has spectral type G4IV and $m_* = 1.39\,M_\odot$, making it the most massive planet-bearing star known. The main sequence progenitor of HD 38529 was probably of spectral type F5V; it would have been difficult to obtain the radial-velocity precision during that stage that can now be obtained for the subgiant.

HD 82943.

The two planets orbiting HD 82943 have periods of 220 and 440 d, respectively, indicating a likely 2:1 mean motion resonance (Santos et al. 2003b; Mayor et al. 2004).

HD 74156.

The parameters of the outer planet in the system HD 74156 are not known precisely yet; the period is at least 6 yr and $m_p \sin i \gtrsim 8\,M_{\rm jup}$ (Marcy et al. 2003b). Since the period of the inner planet is only 51.6 d, this is also a very hierarchical system (Naef et al. 2004).

3.6 Interactions Between the Planets in Multiple Systems

General Formalism

As a first approximation, it is normally assumed that planets orbit their parent stars in Keplerian orbits; gravitational interactions between the planets and tidal effects can be treated as small perturbations. It has been pointed out already that in some extrasolar systems these perturbations are much stronger than in the Solar System, because they contain massive planets in close proximity to the parent star and to each other. This situation is somewhat similar to the ring and moon systems of the giant planets in the Solar System, which also exhibit a wealth of phenomena due to tidal and mutual gravitational interaction (e.g., De Pater and Lissauer 2001; Murray and Dermott 1999).

It is well known that there is no analytic solution to the general three-body problem. Various simplifications have therefore been studied, including the *restricted three-body problem*, in which one of the bodies is assumed to have negligible mass, and *Hill's problem* in which the mass of one body is much larger than the other two. One usually stars by writing the potential as the sum of a part that describes the Keplerian motion of the bodies about the central star, plus a part called *disturbing function*, which contains the *direct terms* accounting for the pairwise interactions among the planets and the *indirect terms* associated with the back-reaction of the planets on the central star. The gradient of the disturbing function describes the additional forces on the planets. One can then proceed by expanding the disturbing function

in terms of small parameters; depending on the system under consideration one can consider expansions in the eccentricities, inclinations, or ratios of the planets' masses to the mass of the star. This expansion can now be inserted into *Lagrange's planetary equations*, which form a set of differential equations that express the time derivatives of the orbital elements by partial derivatives of the disturbing function (e.g., Murray and Dermott 1999).

Using this formalism, one thus arrives at a system of coupled differential equations for the orbital elements. To study the long-term evolution of the orbits, one can ignore all short-period terms of the disturbing function; these will average out to zero over sufficiently long time intervals. Solving the simplified set of differential equations, which contains only the *secular* (i.e., long-period) terms, one typically obtains solutions that are periodic in a and e, and contain linear terms for (orbital precession) and ω (precession of the periastron). Additional complications occur, however, if the e orbital periods of the two planets are nearly commensurate, i.e., if their ratio is close to a ratio of two small integers. The terms in the disturbing function corresponding to such a *mean motion resonance* do not average to zero, because the disturbance occurs always at approximately the same orbital phase. The situation is analogous to a simple harmonic oscillator

$$m\ddot{x} + m\omega_0^2 x = F\cos\omega_d t \tag{10}$$

driven at a forcing frequency ω_d near the natural frequency ω_0. For $\omega_d \neq \omega_0$ the solution to (10) is

$$x = \frac{F}{m(\omega_0^2 - \omega_d^2)}\cos\omega_d t + C_1\cos\omega_0 t + C_2\sin\omega_0 t\,. \tag{11}$$

We see that for $\omega_d \approx \omega_0$ the response can be very large even for a small driving force F. A famous example in the Solar System is the 3:2 mean motion resonance between the orbits of Pluto and Neptune. The angle $\phi \equiv 3\lambda_P - 2\lambda_N - \omega_P$, where λ_P and λ_N are the mean longitudes of Pluto and Neptune, and ω_P Pluto's longitude of perihelion, librates about 180° with a period of 19,670 years (Cohen and Hubbard 1965). This mechanism prevents close encounters of Pluto with Neptune and thus stabilizes the orbit of Pluto. A second type of resonance, called *secular resonance* arises if one of the precession rates ($\dot{\omega}$ or $\dot{}$) equals an eigenfrequency of the system.

A slightly different general formalism for calculations of the tidal, rotational, and dynamical evolution of planetary systems has been developed by Mardling and Lin (2002). It involves calculating the evolution of the orbital angular momentum vector and of the Runge–Lenz vector[15] of the inner orbit. Since these vectors are constant for unperturbed orbits, their components vary slowly compared to the orbital period. The secular evolution of the orbital elements can therefore be obtained by time-averaging the rates of change of

[15] The Runge–Lenz vector points in the direction of periastron and has a magnitude equal to the eccentricity.

the orbital angular momentum and Runge–Lenz vectors over the inner orbit. The resulting equations are quite complicated, but they can be used to implement fairly efficient and flexible computer programs for dynamical simulations of extrasolar planet systems.

Long-Term Stability

An important question about a multiple planetary system is its long-term stability. It is very difficult to prove that a system is stable in the sense that all planets remain bound for all time. One therefore frequently restricts the analysis to the weaker *Hill stability*, which means that the planets cannot undergo close approaches, which otherwise might disrupt the system. It can be shown that two planets in initially circular co-planar orbits cannot enter each other's Hill spheres for definition) if (Gladman 1993)

$$\Delta \equiv (a_2 - a_1)/a_1 > 2\sqrt[6]{3}(\mu_1 + \mu_2)^{1/3} \approx 2.4(\mu_1 + \mu_2)^{1/3} \ , \qquad (12)$$

where μ_1 and μ_2 are the ratios of the planetary masses to the mass of the central star, and a_1 and a_2 the orbital radii. This criterion gives a useful first indication that systems such as HD 168443 and 47 UMa are stable, unless their inclinations are extremely small, which would make μ_1 and μ_2 much larger than their minimum values (Marcy et al. 2001b; Fischer et al. 2002a). One caveat is that for a large *relative* inclination between the two planets a more stringent criterion than (12) would have to be applied (Ida and Makino 1993).

Orbital Resonances

More detailed studies of the stability of the known multiple systems make use of the analytic formalism sketched above, and of numerical integrations of the orbits. Orbital resonances of different types play an important role in at least five systems (υ And, 55 Cnc, GJ 876, HD 12661, and HD 82943), as indicated in Table 5. Orbital dynamics of extrasolar planets has thus become a rich field, in which many new results can be expected as more multiple systems are discovered, and as improved data become available for the systems that are already known.

Soon after the discovery of υ And c and υ And d, the first stability analyses of this system were carried out (Laughlin and Adams 1999; Lissauer and Rivera 2001). More recent studies have made use of updated planetary parameters from continued monitoring observations. They indicate that the outer two planets occupy nearly edge-on orbits with low relative inclination (Lissauer and Rivera 2001; Chiang et al. 2001; Chiang and Murray 2002). The two planets seem to inhabit a secular resonance, in which $\Delta\omega \equiv \omega_D - \omega_C$ librates about 0°. It is worth pointing out that a detailed analysis is needed to find this possible resonance mechanism in υ And.

The 2:1 mean motion resonance in the GJ 876 system is more obvious, as it is directly reflected in the orbital periods of the two planets. The resonance certainly helps to stabilize the system; stable configurations can be found with both high and low values of the relative inclinations of the two orbits (Rivera and Lissauer 2001; Ji et al. 2002). This is quite remarkable in view of the relatively large orbital eccentricities, but the configuration of GJ 876 remains in fact stable even at much larger eccentricities (Lee and Peale 2003). It has already been pointed out that the interaction between the two planets in the GJ 876 system is sufficiently strong to produce a detectable deviation from a model with two Keplerian orbits (see Sect. 4.1). It should thus be possible to determine the true planet masses (without the $\sin i$ factor) from the strength of this interaction, but this procedure does not yet lead to reliable results, even under the assumption of coplanarity (Laughlin and Chambers 2001; Nauenberg 2002a).

In some cases, the current uncertainties of the orbital parameters prevent clear statements on the dynamical state of a given system, because regular and chaotic orbits can be located close to each other in parameter space (e.g. HD 12661, Kiseleva-Eggleton et al. 2002; Lee and Peale 2003; Goździewski and Maciejewski 2003). In the HD 82943 system, which contains two planets in a 2:1 mean motion resonance, the nominal best-fit Keplerian orbits represent an unstable system, which should self-destruct quickly (Goździewski and Maciejewski 2001). An ad-hoc adjustment of the argument of periastron ω of the inner by about 30° leads to a stable system, however, with a model radial-velocity curve that is nearly indistinguishable from the original one. Similarly, the stability of the 47 UMa system depends critically on the eccentricity of the outer planet, which is poorly constrained by the current data, and on the relative inclination of the two planets, which cannot be determined with radial-velocity measurements (Laughlin et al. 2002; Goździewski 2002).

It has been speculated that planets could also be found in a 1:1 resonance, similar to Jupiter's Trojan asteroids (Laughlin and Chambers 2002; Nauenberg 2002b). For an exact 1:1 resonance the radial-velocity signature would be indistinguishable from that of a single planetary companion, but if there are slight deviations from the exact resonance, the pair of planets can execute horseshoe-type orbits around the Lagrangian points. A good example for this type of motion in the Solar System is given by Saturn's moons Janus and Epimetheus (see e.g. Murray and Dermott 1999). In that case, deviations from the single-planet radial-velocity curve can become quite significant over a few orbital periods. If planets in a 1:1 resonance exist, they could therefore be detected in the near future; it is even possible that one or the other of the known "single" planets will over time turn out to be a Trojan pair.

Commensurabilities between orbital periods can be set up during the early evolution of a planetary system, through orbital migration (see Sect. 3.2). If both planets open a gap in the disk, the outer planet migrates more quickly and approaches the inner planet, until it becomes locked in a 2:1 resonance (Snellgrove et al. 2001). This resonance can be maintained through

the subsequent evolution of the system. It is also possible that the planets get locked up in other resonances (e.g., 3:1, 4:1, 5:1, or 5:2), depending on their masses, initial eccentricities, and the time scale for eccentricity damping in the disk (Nelson and Papaloizou 2002). Gathering sufficient statistical information about the incidence of these different resonances might thus provide a useful diagnostic tool of the migration process and of the interactions between the protoplanets and the disk during the formation phase of planetary systems.

4 Radial-Velocity Surveys

The radial-velocity technique has without doubt been the most successful planet detection method so far. In fact, *all* known extrasolar planets around main-sequence stars were discovered in this way[16]. The basis of the radial-velocity method is relatively simple: one obtains a time series of high-resolution spectra of the target star, and searches for periodic variations of the absorption line Doppler shift due to the motion of the star around the center of mass of the star–planet system. It is the exquisite precision of radial-velocity measurements achieved during the past decade that has made the plethora of recent planet detections possible. In this chapter we will discuss the foundations of the method, the design principles of the spectrographs used, and astrophysical limitations of the radial-velocity technique. The properties of the currently known extrasolar planets will be the subject of Sect. 3.

4.1 The Radial-Velocity Technique

Planetary Orbits from Radial Velocities

The orbit of a binary system is defined by seven parameters, the so-called *orbital elements* (see e.g. Batten 1973):

1. P, the orbital period;
2. i, the inclination of the orbital plane with respect to the tangent plane of the sky;
3. Ω, the position angle (measured from North through East) of the line of nodes, which is the intersection of the orbital and tangent planes;
4. ω, the angle between the direction of the ascending node (at which the star crosses the tangent plane while receding from the observer) and the periastron;
5. a, the semi-major axis of the orbit;
6. e, the eccentricity of the orbit;
7. T, the time of passage through periastron.

[16] Editor note added in proof: It is not true anymore, recently, planets have been detected by transit observations (see 6.4)

The radial velocity curve V of the primary star in a spectroscopic binary can be expressed as

$$V = \gamma + K_1 \left[\cos(\nu + \omega) + e \cos \omega\right] \ , \tag{13}$$

where γ is the radial velocity of the center of mass of the system, K_1 the velocity amplitude, and ν the *true anomaly*, i.e., the position angle measured from periastron. The time dependence of $\nu(t)$ is given implicitly by the relations (see e.g. Heintz 1971; Murray and Dermott 1999)

$$\frac{2\pi}{P}\,(t - T) = E - e \sin E \tag{14}$$

and

$$\tan \frac{\nu}{2} = \sqrt{\frac{1+e}{1-e}}\, \tan \frac{E}{2} \ ; \tag{15}$$

the quantity E in these equations is called the *eccentric anomaly*. It is thus clear that the parameters P, T, e, and ω can be determined directly from the shape of the velocity time series. Ω and i, on the other hand, cannot be determined from spectroscopic observations alone. The semi-major axis of the primary around the center of mass is related to K_1 by

$$a_1 \sin i = \frac{P}{2\pi} \sqrt{1 - e^2}\, K_1 \, . \tag{16}$$

According to Kepler's Third Law,

$$a^3 = \left(\frac{P}{2\pi}\right)^2 G\,(m_1 + m_2) \ , \tag{17}$$

where $a \equiv a_1 + a_2$ is the semi-major axis of the relative orbit of the two components. Using $m_1 a_1 = m_2 a_2$ and (16) and (17), we can derive the relation

$$\frac{(m_2 \sin i)^3}{(m_1 + m_2)^2} = \frac{P}{2\pi G}\, K_1^3 (1 - e^2)^{3/2} \, . \tag{18}$$

The left-hand side of this equation is called the *mass function* of the system. If the secondary is a planet, we can use $m_2 \ll m_1$ to simplify (18). This gives

$$m_p \sin i \approx \left(\frac{P}{2\pi G}\right)^{1/3} K_* \, m_*^{2/3} \sqrt{1 - e^2} \ , \tag{19}$$

i.e., we can derive $m_p \sin i$ from the radial-velocity data provided that the mass of the central star m_* is known. In more convenient units one can write

$$m_p \sin i \, [M_{\rm jup}] \approx 3.5 \cdot 10^{-2} K_* [\mathrm{m\,s^{-1}}] \, P^{1/3} [\mathrm{yr}] \ ; \tag{20}$$

this means that Jupiter causes a $12.5\,\mathrm{m\,s^{-1}}$ wobble in the radial velocity of our Sun.

The quantity $m_p \sin i$ is frequently referred to as the "minimum mass" of the planet. In each individual case, the actual mass of the planet may be considerably larger than this lower limit inferred from the radial-velocity technique. In a statistical sense, however, this uncertainty is not as severe as one might think. In a set of randomly oriented orbits $\cos i$ is uniformly distributed between 0 and 1. (It is more likely to observe an object nearly equator-on than nearly pole-on.) This means that in 87% of all cases $\sin i \geq 0.5$, and that in only 0.5% of all cases $\sin i \leq 0.1$. Therefore distributions of $m_p \sin i$ from radial velocity surveys are fairly representative of the true distribution of planetary masses (see also Sect. 3.2).

Multiple Systems

To first approximation, systems with several planets can be represented by a linear superposition of the individual Keplerian orbits. This approximation is a good one if the planetary masses are small, so that their mutual interaction can be neglected. In the case of massive planets, however, this approximation can break down on time scales that are not very long compared to the orbital periods, so that a treatment of the full many-body problem including dynamical resonances is required. In practice, this is best done in several steps (Laughlin and Chambers 2001; Rivera and Lissauer 2001). The starting point is a set of Keplerian fits for each planet; the corresponding orbital elements are called the *osculating elements* at the starting epoch. One can then perform a self-consistent integration of the many-body problem, compute a synthetic radial-velocity curve of the central star from the solution, compare this synthetic curve to the observations, and calculate the corresponding χ^2 value. This procedure can be repeated for different sets of osculating elements; the Levenberg–Marquardt method (e.g. Press et al. 1992) can be used to find the osculating elements that minimize the χ^2.

In this context, it is important to realize that the interaction between the planets depends on their masses and the relative inclination of their orbital planes. If sufficiently precise radial-velocity data are available, it is therefore possible to derive these parameters from dynamical analyses of multiple systems. The uncertainties are fairly large, however, because the parameter space to be searched has many dimensions, especially if the planets are not assumed a priori to be in coplanar orbits. Direct measurements of the relative orientations of the orbits with astrometric methods (see Sect. 9) can provide much better constraints on the dynamical evolution of multiple systems.

4.2 Limitations of the Radial-Velocity Precision

The Principle of Precise Doppler Spectroscopy

To detect the reflex motion of stars orbited by extrasolar planets, it is necessary to determine their radial velocity variations with stunning precision:

a measurement error of $3\,\mathrm{m\,s^{-1}}$ means that the wavelength shift of the stellar absorption spectrum has to be determined to one part in 10^8. The resolving power of modern high-resolution spectrographs is typically of order $R \equiv \lambda/\Delta\lambda \lesssim 100,000$; a precision of 1/1,000 resolution element is therefore required. This is possible only by taking spectra with high signal-to-noise, and averaging over many spectral lines. Several conditions must be met to reach the desired precision of a few $\mathrm{m\,s^{-1}}$:

- The target star must have a sufficient number of absorption lines. This excludes main-sequence stars of spectral earlier than roughly F5 V, which have fewer lines than the cooler stars.
- The stellar absorption lines must be narrow. This again excludes stars with early spectral types and young stars, because they show too much rotational broadening.[17]
- The stellar photosphere must be sufficiently stable. This excludes active (e.g., flaring) stars and pre-main-sequence objects.
- The spectrograph used must be extremely stable, or a suitable calibration technique has to be applied.

For the first three reasons, radial-velocity surveys have concentrated mostly on F, G, and K main-sequence stars. M dwarfs and even brown dwarfs are now also attracting much interest for searches of low-mass planets, because the detection limit for m_p scales with $m_*^{2/3}$ (19). Many K giants are also suitable for precise radial-velocity monitoring, and giant planets have been found orbiting some of them (Frink et al. 2002; Sato et al. 2003).

Photon Noise, the Fundamental Limit

To understand the fundamental limit of the attainable radial-velocity precision, consider first one pixel on the detector of the spectrograph. The intensity change ΔN (measured in detected photo-electrons) in this pixel due to a small variation of the radial velocity ΔV can be written as (Connes 1985; Bouchy et al. 2001)

$$\Delta N \equiv N - N_0 = \frac{\partial N_0}{\partial \lambda}\,\Delta\lambda = \frac{\lambda}{c}\,\frac{\partial N_0}{\partial \lambda}\,\Delta V\,. \tag{21}$$

Solving for ΔV, we obtain

$$\Delta V = \frac{c}{\lambda}\,\frac{N - N_0}{\partial N_0/\partial \lambda}\,. \tag{22}$$

[17] The rotation rate of main-sequence stars is linked to their structure. Stars with $m \lesssim 1.4\,M_\odot$ have outer convection zones; the interplay of convection with rotation leads to differential rotation and drives a dynamo. Magnetic breaking reduces the stellar rotation rate. This leads to a drastic difference in the typical rotation rates between stars earlier and later than F5 V.

In the photon noise-limited case the measurement error on N is proportional to $\sqrt{N}$; the Doppler precision is therefore inversely proportional to $|\partial N_0/\partial\lambda|\,\lambda/\sqrt{N_0}$, which can be taken as a "figure of merit" of the pixel in the stellar spectrum under consideration. When we combine the data from all pixels i in the spectrum, they should get weights $w(i)$ that are proportional to the square of this figure of merit:

$$w(i) \equiv \frac{\lambda^2(i)[\partial N_0(i)/\partial\lambda(i)]^2}{N_0(i)+\sigma_D^2}\ , \tag{23}$$

where we have also included a potential contribution to the noise from the detector σ_D. The radial-velocity change computed from the full spectrum is then

$$\Delta V = \frac{\sum \Delta V(i) w(i)}{\sum w(i)} = c\,\frac{\sum [N(i)-N_0(i)]\sqrt{\frac{w(i)}{N_0(i)+\sigma_D^2}}}{\sum w(i)}\,. \tag{24}$$

One can easily verify that the associated measurement uncertainty $\sigma_{\Delta V}$ can be expressed as

$$\sigma_{\Delta V} = \frac{c}{\sqrt{\sum w(i)}}\,. \tag{25}$$

It is now convenient to introduce a "quality factor" Q defined by

$$Q \equiv \frac{\sqrt{\sum w(i)}}{\sqrt{\sum N_0(i)}} = \frac{\sqrt{\sum w(i)}}{\sqrt{N_{\rm tot}}}\ , \tag{26}$$

where $N_{\rm tot}$ is the total number of detected photons. In the high-flux limit, where detector noise is negligible, Q in independent of the stellar flux; it represents the sharpness and richness in spectral lines of the spectrum. With this definition we can finally write

$$\sigma_{\Delta V} = \frac{c}{Q\sqrt{N_{\rm tot}}}\,. \tag{27}$$

This formulation is well suited for modeling the influence of stellar spectral type, rotational line broadening, and spectrograph resolution on the attainable velocity precision (see Fig. 18). For $v\sin i \lesssim 6\,{\rm km\,s^{-1}}$ the line profiles are broadened by the rotation, which leads to a linear decrease of the average $\partial N_0/\partial\lambda$ and therefore of Q (see (23) and (26)). For larger values of $v\sin i$, neighboring spectral lines start to become blended, which leads to $Q \propto (v\sin i)^{-1}$. At low spectral resolution ($R \lesssim 50,000$) all lines are blended and $Q \propto R$. When the resolution is increased to match the intrinsic (broadened) line width, Q reaches a constant value. Better spectral resolution is therefore beneficial, but only up to $R \approx 100,000$.

Limitations due to Stellar Variability

For radial-velocity measurements with a precision of a small fraction (of order 1/1,000) of the line width, physical processes in the stellar photosphere or

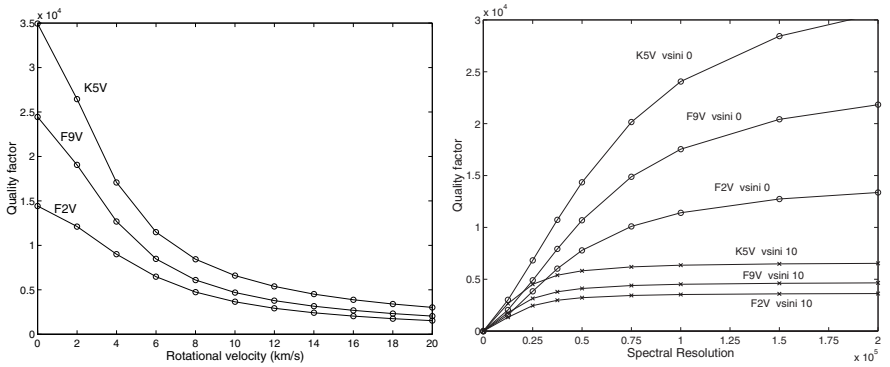

Fig. 18. Quality factor for radial-velocity measurement in the spectral range 3800 Å to 6800 Å. *Left panel*: Dependence on rotational broadening $v \sin i$ for K5 V, F9 V, and F2 V stars, for infinite spectral resolution. *Right panel*: Dependence on spectrograph resolution for K5 V, F9 V, and F2 V stars with $v \sin i = 0$ and $v \sin i = 10\,\mathrm{km\,s^{-1}}$. From Bouchy et al. (2001)

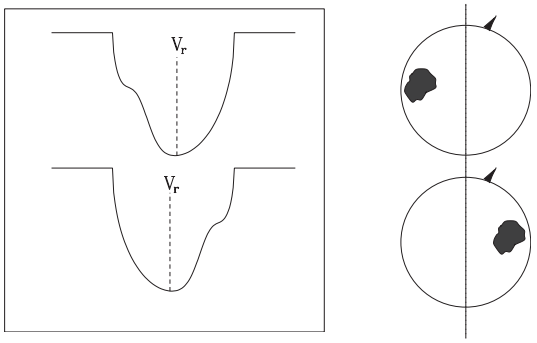

Fig. 19. Illustration of the effect of star spots on line profiles of a rotating star. From Queloz (1999)

chromosphere that affect the line profiles have to be considered carefully. An obvious example are starspots (Fig. 19). When a starspot (or group of spots) rotates into view, it hides part of the approaching side of the star. This causes a bump in the blue wings of absorption lines, which corresponds to a redward shift of the line centroid. When the spot rotates across the meridian, the bump moves from the blue wing to the red wing of the line, now causing a blueshift of the line. If the spot or spot group is long-lived, it will rotate periodically into and out of view, thus mimicking the periodic signal of a planet. To avoid this kind of misinterpretation, one should not rely on the line centroid (or on cross-correlating the observed spectra with a template) alone, but also check for variations of the line depths and shapes. Only if all lines vary synchronously and without changing their profiles is the interpretation of radial-velocity variations as the signature of a planet tenable.

Another indicator for the presence of star spots is of course photometric variability. A case in point is the G0 V star HD 166435 (Queloz et al. 2001). Observations with the ELODIE spectrograph at the Observatoire de Haute Provence revealed low-amplitude radial velocity variations with a period of 3.7987 days, suggestive of a possible planetary companion. Photometric observations uncovered variations with the same period and a one-quarter cycle phase shift, however, as expected for dark photospheric spots. This interpretation is also supported by a detailed analysis of the spectroscopic data, which revealed variations of the line profiles and a loss of coherence of the radial-velocity signal on time scales longer than $\sim$30 days (which gives an indication of the time over which star spots are stable). Photometric variability has also been observed in HD 192263 (Henry et al. 2002), which had been thought to host a $0.75\,M_{\rm jup}$ planet in a 24-day orbit[18]. With a careful analysis it is thus frequently possible to separate true planetary companions from photospheric effects.

There remain difficult cases, however, in which the planetary hypothesis is plausible, but very hard to establish beyond reasonable doubt. A good example for this category is ε Eridani, which shows radial-velocity variations with amplitude $K = 19\,{\rm m s}^{-1}$ and period $P = 6.9\,{\rm yr}$ (Hatzes et al. 2000). These variations can be fit reasonably well with a Keplerian orbit, but the star also displays variations of the Ca II H and K lines indicative of magnetic activity. Further observations will be required to attribute the seven-year variations to either a companion or stellar activity.

In any case, even low-level stellar variability produces background noise, which limits the ultimate precision that can be attained with Doppler observations. The activity of cool stars is directly related to their rotation rate and thus to their age. A high rotation rate usually implies a stronger dynamo and thus stronger magnetic activity (spots, X-ray emission, chromospheric lines, ...). Magnetic breaking reduces the rotation rate and thus the activity. The time scale for this process depends on the mass of the star; low-mass stars (spectral type M) take the most time to slow down and thus show pronounced activity even at fairly old ages. The typical radial-velocity noise due to spots in G dwarfs decreases from $\approx 30 \ldots 50\,{\rm m\,s}^{-1}$ at the age of the Hyades ($\sim 625\,{\rm Myr}$) to $\lesssim 5\,{\rm m\,s}^{-1}$ at the age of the Sun; convective perturbations of the radial velocity can have a similar magnitude (Saar and Donahue 1997).

Good indicators for activity in cool stars are the profiles of the Ca II H and K resonance lines at 3968.5 Å and 3933.7 Å, which consist of a narrow chromospheric emission component (in active stars) superposed on a very broad photospheric absorption line (see Fig. 20, left panels). For a quantitative analysis of these line profiles, one usually uses an "activity index" $R'_{\rm HK}$, which is defined as the ratio of the chromospheric emission in the cores of the Ca II H and K lines to the total bolometric emission of the star (Noyes et al. 1984).

[18] Note of the Editor added in proof: Santos et al. 2003a,b claims that the photometric and line profile are not synchronized and that the planet interpretation still hold.

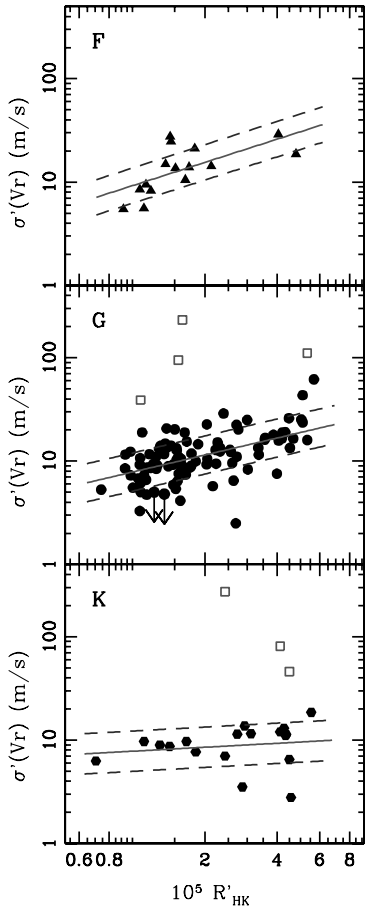

Fig. 20. Plots of the reduced radial-velocity rms, $\sigma'(\mathrm{Vr})$, as a function of the activity index R'_{HK} for F, G, and K dwarfs (for details see text). The solid lines represent the best linear least-squares fit to the data; the rms around the fit is indicated by the dashed lines. Open squares represent stars with planetary systems with $\sigma'(\mathrm{Vr})$ computed without subtracting the orbital solution. From Santos et al. (2000)

Plots of the radial-velocity jitter in F, G, and K dwarfs versus R'_{HK} show that these quantities are indeed correlated (Saar et al. 1998; Santos et al. 2000, see also the right panels in Fig. 20). The same plots also demonstrate that those stars for which planetary companions have been announced clearly stand out from the general distribution; this is another argument that the observed radial-velocity variations cannot be explained by stellar activity.

In addition to this clearly established general correlation of chromospheric activity with radial-velocity variability, the activity indices themselves show temporal variations. Saar and Fischer (2000) find that in 30% of the stars observed in the Lick survey the activity measured in the Ca II λ8662 line

(which is one of the lines of the Ca II infrared triplet) is correlated with simultaneously measured radial velocities, i.e., the level of activity itself produces a shift of the radial velocity. They argue that the main cause of these effects is modification of the mean line bisector shape brought on by long-term, magnetic activity-induced changes in the surface brightness and convective patterns. Taking out this trend before analyzing the radial velocities thus reduces the residual scatter. On the other hand, Paulson et al. (2002) find a similar correlation between $R'_{\rm HK}$ and the radial velocity in only five of 82 stars in the Hyades. It thus remains to be seen to which extent activity indices can be used not only to pre-select intrinsically "quiet" stars, but also to remove activity-induced systematic effects from radial-velocity data.

Stellar pulsations can also produce radial-velocity variations that could be mistaken for the signature of a planet. Again, photometry is a good way of double-checking whether this may be the case. For radial pulsations, there is a straightforward relation between the amplitude ΔR of the radius changes and the radial-velocity curve,

$$\Delta R = \int_0^{T/4} V\,dt\,. \tag{28}$$

Using this relation and

$$\frac{\Delta L}{L} = 2\frac{\Delta R}{R} + 4\frac{\Delta T}{T}\ , \tag{29}$$

which follows directly from the Stefan–Boltzmann Law $L = 4\pi R^2 T_{\rm eff}^4$, it is possible to predict the photometric variability from the observed radial-velocity variations. For example, in the case of 51 Peg, the amplitude of the radial-velocity curve ($59\,{\rm m\,s^{-}}1$, Mayor and Queloz 1995) implies $\Delta R/R = 0.5\%$ and thus $\Delta L/L = 1\%$ (for $\Delta T/T = 0$). The observed photometric stability of better than 0.1% therefore rules out radial pulsations; the assumption that ΔR would be compensated by ΔT such that $\Delta L/L \leq 0.1\%$ is too contrived. No such direct case can be made against non-radial pulsations, which do not necessarily imply detectable photometric variations. However, several strong arguments are also available against this interpretation:

- Only modes with very low amplitudes ($\ll 1\,{\rm m\,s^{-1}}$) and periods $\lesssim 1\,{\rm h}$ are detected in the Sun, and expected for Sun-like stars.
- Mechanisms that could excite a high-order non-radial pulsation mode should also excite many other similar modes. There is no plausible mechanism that could selectively excite one single mode, and thus mimic a planetary signal.
- Detailed studies of stars for which planets have been published have not revealed any changes of the line shapes (e.g. Gray 1998).

Taken together, these arguments rule out pulsations as a plausible explanation for the observed radial-velocity variations.

The situation is somewhat more complicated in the case of giant stars. All evolved stars exhibit intrinsic variability to some degree. A well-known example is the K2 giant α Bootis, which has a complicated variability pattern with an amplitude of $\sim 200\,\mathrm{m\,s^{-1}}$ on time scales down to a few days (Hatzes and Cochran 1994). Giant stars have therefore not been targeted by the planet surveys. A survey of nearby K giants aimed at assessing their suitability as astrometric reference stars has shown, however, that many of these stars have fairly stable radial velocities (Frink et al. 2001). In fact, about 2/3 of the stars observed are drawn from a distribution with a mean radial velocity scatter of $\sim 20\,\mathrm{m\,s^{-1}}$. There is also a correlation with color, which implies that a suitable color selection criterion can reduce the contamination by photospherically unstable targets. The companion of the K2 III giant ι Draconis with a minimum mass of 8.9 AU discovered serendipitously in the survey mentioned demonstrates that detecting planets around giants is indeed possible (Frink et al. 2002).

4.3 Spectrograph Design

Cross-Dispersed Echelle Spectrographs

Precise radial-velocity measurements require a large spectral range covered with high spectral resolution ((27), Fig. 18). These requirements are met best with cross-dispersed echelle spectrographs (Schroeder 2000; Vogt 1987; Baranne 1999). This type of instrument takes its name from the arrangement of two separate dispersing elements. The first is a grating used in high orders, which is responsible for the spectral resolution. To avoid that the overlapping orders of the main grating fall on the same pixel on the detector, a second low-dispersion element (prism, grism or combination of the two) is used in the orthogonal direction. This leads to a format which uses most of the real estate on a CCD chip for a long high-resolution (typically $R \equiv \lambda/\Delta\lambda = 50,000 \ldots 100,000$) spectrum.

Extraordinary measures have to be taken, of course, to obtain a long-term reproducibility of order 1/1,000 pixel for measurements with these instruments. The first requirement is to build the spectrograph as stable as possible. Changes in the spectrograph point spread function (i.e., the observed profile of an infinitely narrow line) due to flexure or thermal expansion can significantly alter the measured position of line centroids, and thus introduce noise in the radial-velocity measurements. In an air-filled spectrograph one also has to take into account changes of the observed "air" wavelength with pressure ($90\,\mathrm{m\,s^{-1}\,mbar^{-1}}$) and temperature ($200 \ldots 300\,\mathrm{m\,s^{-1}\,K^{-1}}$, depending on the observatory elevation). The key to success is referencing all observations to a stable standard, and to eliminate all systematic errors that can enter this process.

When a stellar radial-velocity measurement has been obtained, it must be transformed from the observatory reference system into an inertial reference

frame. Getting a sufficiently precise ephemeris for the motion and rotation of the Earth is no problem, but timing the observation requires some care. As the radial velocity of the Earth can change significantly (up to $2.4\,\mathrm{m\,s^{-1}\,min^{-1}}$) while the shutter is open, we need to know the photon-weighted midpoint of the exposure. Taking this simply as the midpoint between opening and closing the shutter can produce a very significant error if the weather is partly cloudy; it is therefore advisable to monitor the photon flux during the exposure with a separate high-speed photometer.

The Simultaneous Thorium Technique

The classical way of providing good wavelength calibration of astronomical spectra is the simultaneous observation of the star and an emission spectrum from an arc lamp. Spectrographs based on this principle have been used for many years by the Swiss planet search team (ELODIE and CORALIE, Baranne et al. 1996). The new HARPS spectrograph to be installed at ESO's 3.6 m telescope on La Silla has been designed to reach a Doppler precision of $1\,\mathrm{m\,s^{-1}}$ (Pepe et al. 2000)[19]. These instruments use two optical fibers to couple both the star light and the light from the Thorium lamp to the spectrograph input; each "science exposure" thus contains truly simultaneous calibration information. The observable quantities are thus the wavelength differences $\lambda_s(f_1,t_1)-\lambda_T(f_2,t_1)$ between stellar absorption features and Thorium lines. (The argument indicate which fiber was used and the time of the exposure.) In addition, a "calibration exposure" is taken in which Thorium light is coupled into both fibers. The observed Doppler shift between the two Thorium spectra $\lambda_T(f_1,t_2)-\lambda_T(f_2,t_2)$ reflects systematic effects induced by the two different paths through the spectrograph. The double difference $[\lambda_s(f_1,t_1)-\lambda_T(f_2,t_1)]-[\lambda_T(f_1,t_2)-\lambda_T(f_2,t_2)]$ therefore provides a reference of the stellar spectrum to the Thorium lines, which is free of both temporal drifts and systematic differences between the two fibers. Advantages of the Thorium technique are a large usable spectral range and relatively high transmission ($\sim$80% for a well-adjusted fiber).

In addition to providing a convenient way of coupling the telescope to the spectrograph, the optical fibers fulfill the important role of stabilizing the stellar light on the spectrograph slit. In a classical spectrograph, which is attached directly to the telescope, slight displacements of the star with respect to the spectrograph slit can lead to serious shifts of the observed wavelength (see Fig. 21). Keeping the star centered on the slit with the precision required for planet surveys is beyond the capabilities of telescope guide systems. In a fiber-fed spectrograph this problem is reduced substantially by the "scrambling" effect of the fiber: an off-axis illumination of the fiber input still leads to

[19] Editor note added in proof: HARPS is installed and available to the community since October 2003

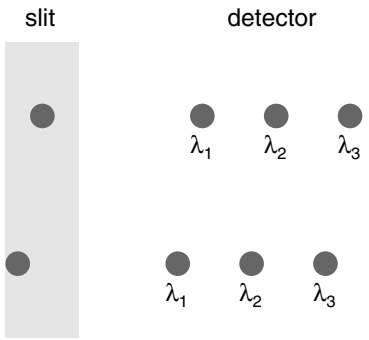

Fig. 21. Stars well-centered (*top*) and poorly centered (*bottom*) on the spectrograph slit. It is apparent that motion of the light centroid perpendicular to the slit (along the dispersion direction) leads to an apparent shift of the observed wavelength

a circularly symmetric output from the fiber (see e.g., Fig. 7 in Queloz 1999). If this azimuthal averaging of the fiber is insufficient, one can further improve the uniformity of the slit illumination by employing a *double scrambler*, which also performs some radial redistribution of the light (Queloz et al. 1999).

The Iodine Absorption Method

An alternative approach to coping with long-term drifts of the spectrograph and unstable illumination of the slit is passing the starlight through an absorbing medium before entry into the spectrograph, thereby superimposing reference absorption lines that experience the same instrumental shifts as the stellar spectrum. It was first suggested to use telluric absorption lines (i.e., absorption lines originating in the Earth's atmosphere) as wavelength references (Griffin and Griffin 1973); modern versions of this technique use an absorption cell in front of the spectrograph slit (Campbell and Walker 1979). A long-term precision of $\sim 15\,\mathrm{m\,s^{-1}}$ has been achieved with an absorption cell filled with hydrogen fluoride gas (Campbell et al. 1988). The main drawbacks of the HF molecule are the fairly small wavelength range covered by the absorption band ($\sim 100\,\text{Å}$), its corrosive nature, and its lethal effect on humans.

After an extensive search for a better absorbing medium, Marcy and Butler (1992) concluded that gaseous iodine is the molecule of choice. It combines the advantages of strong line absorption coefficients (requiring only a short absorption length and low pressure), large wavelength coverage (from 5,000 to 6,300 Å), chemical stability, and low risk to human health. A 10 cm long cell filled with gaseous I_2 at a pressure of 1/100 atm was built for the Hamilton Echelle Spectrograph (Marcy and Butler 1992); similar cells have also been installed at other telescopes. Observations with an iodine cell produce a spectrum in which stellar and iodine lines are heavily blended with each other. The data analysis therefore requires sophisticated modeling of the observed

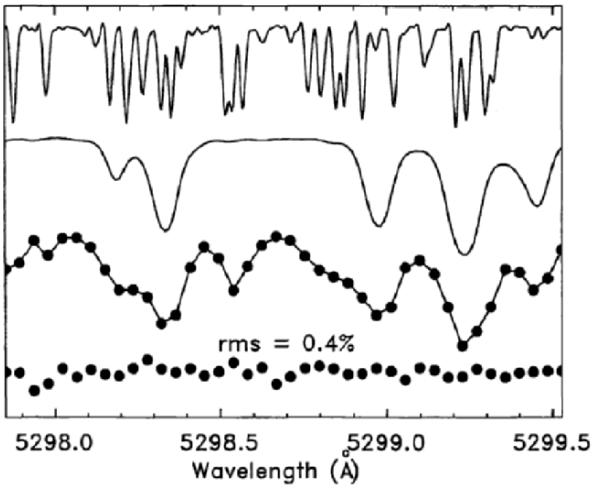

Fig. 22. Modeling of a spectrum observed through an iodine cell. *Top*: The template iodine spectrum. *Second*: The template stellar spectrum (in this case τ Ceti, G8 V). Note how rich this spectrum is; the figure shows only $\sim$1.6 Å of the $\sim$850 Å range used for the Doppler analysis. *Third*: The points are in observation of τ Ceti made through the iodine absorption. The solid line is a model of the observation, composed of the template iodine and stellar spectra. The free parameters consist of the spectrograph point spread function and the Doppler shift of the template star relative to the template iodine. *Bottom*: 10 times the difference between the model and the observation. The model and the observation differ by 0.4% rms. From Butler et al. (1996)

spectrum (see Fig. 22). The data reduction software needs three inputs: the observed spectrum $I_{\rm obs}$, a high-resolution iodine template spectrum $T_{\rm I_2}$, and a high-resolution spectrum $I_{\rm S}$ of the target star (obtained by deconvolving a spectrum with very high signal-to-noise, taken without the iodine cell). The spectra taken through the iodine cell are then modeled as (Butler et al. 1996)

$$I_{\rm obs}(\lambda) = k\,[T_{\rm I_2}(\lambda) I_{\rm S}(\lambda + \Delta\lambda)] * {\rm PSF}\ , \tag{30}$$

where k is a normalization constant and PSF the spectrograph point spread function. The operator $*$ denotes convolution as defined in (238). The Doppler shift $\Delta\lambda$ is obtained by a χ^2 minimization procedure that adjusts this parameter together with twelve others describing the wavelength scale and the shape of the spectrograph point spread function. This technique has consistently produced Doppler measurements with a long-term stability of $3\,{\rm m\,s^{-1}}$ (Butler et al. 1996).

Absolute Astronomical Accelerometry

The simultaneous Thorium and Iodine absorption methods rely on the "random" positions of absorption or emission lines in the spectrum of a molecule or atom. In comparison, it should in principle be advantageous to use the regular transmission comb of an interferometer as a wavelength reference (e.g., Ge et al. 2002). A variation of this concept is the "Absolute Astronomical Accelerometer" proposed by Connes (1985). This instrument is based on two control loops. First, the variable path difference of a tunable Fabry–Perot interferometer is adjusted such that its transmission maxima track the variable Doppler shift of a star. This tracking ability is the main advantage of this concept over the methods described above; it eliminates systematic errors due to the relative shifts of stellar and calibration lines caused by the annual and diurnal variation of the observatory velocity in an inertial reference frame. The second loop involves a tunable laser, which tracks one of the Fabry–Perot transmission peaks. The net result is that the wavelength of the laser line tracks the radial velocity of the star. The beam from the tunable laser is then mixed with a stabilized laser; the change in the beat frequency is the signal that contains the desired information about changes in the stellar Doppler shift. A prototype instrument has been built and tested in the laboratory. It remains to be seen whether wavelength references based on interferometric approaches can reach or surpass the long-term radial-velocity stability that has been demonstrated with gas cells and lamps.

4.4 Radial-Velocity Surveys

The First Planet Detections

In the August 1, 1995 issue of *Icarus* appeared a paper summarizing the results from a 12-year search for Jupiter-mass companions to 21 nearby stars. No planets were found, with limits in the range $m \sin i \leq 1 \ldots 3\, M_{\rm jup}$ for any possible planets with orbital periods up to 15 years (Walker et al. 1995). In retrospect this team was remarkably unlucky; as we now know, their precision and sample size gave them a $> 50\%$ chance to actually discover the first planet around a Solar-type star. But because the sample picked by Walker et al. happened to contain no massive short-period planet, the honor of the first discovery went to Mayor and Queloz (1995), who announced the planet orbiting 51 Peg only three months later. The radial-velocity variations of this star were almost immediately confirmed by Marcy and Butler (1995), who had also started a long-term planet search. This survey soon uncovered two additional planets (Marcy and Butler 1996; Butler and Marcy 1996), providing a first glimpse of the unanticipated diversity of giant planets.

Recent Surveys

Fueled by the unexpected discoveries of giant planets with short orbital periods, the ongoing surveys intensified their efforts, and several new radial-velocity projects were started. More than 2,000 stars are now being monitored with Doppler precisions in the range $3 \ldots 15\,\mathrm{m\,s^{-1}}$. Among the projects that have contributed to the list of known planets are: ELODIE at the Observatoire de Haute Provence (Udry et al. 2001); its improved sister CORALIE at the Swiss Euler Telescope on La Silla (Udry et al. 2000); the Hamilton Echelle Spectrograph at Lick Observatory (Marcy and Butler 1998; Cumming et al. 1999); HIRES at the Keck Observatory (Vogt et al. 2000); the Advanced Fiber-Optic Echelle at Whipple Observatory (Nisenson et al. 1999); the Anglo-Australian Telescope (Butler et al. 2001); the Coudé Echelle Spectrometer at ESO's 3.6 m Telescope on La Silla (Endl et al. 2002); the McDonald Observatory (Cochran et al. 2000); and the Lick bright K giant survey (Frink et al. 2002).

Several attempts have been made to analyze the combined published results from all surveys, in order to obtain statistical information on the distribution of planet masses and periods, and to assess the fraction of stars that have planetary companions (e.g., Nelson and Angel 1998; Tabachnik and Tremaine 2002; Lineweaver and Grether 2002). While such compilations can provide much useful information, of course, one has to keep in mind that it is extremely difficult to estimate the completeness of the underlying data. Many of the long-term surveys have improved their observing techniques over the years, which leads to complicated sensitivity limits as a function of orbital period. Furthermore, the temporal sampling may vary widely from star to star, because "interesting" targets were followed much more frequently than others. A slightly enhanced level of stellar activity may also have an adverse influence both on the detection threshold for planets and on the enthusiasm of the observers to obtain many data points. One finally has to keep in mind that a star without a published planet is not necessarily a star without a detected planet – the observers may just have chosen to wait with the publication until they can get a satisfactory orbital fit. In spite of all these caveats, the amount of information on extrasolar planets that has been gathered with the radial-velocity method is now large enough to enable interesting statistical conclusions. We will come back to this point in the following chapter.

5 Gravitational Microlensing

The detection and monitoring of gravitational microlensing events towards the Galactic bulge and the Large Magellanic Cloud has been used successfully as a tool to study the composition and mass distribution of the Galaxy (Paczyński 1986, 1996; Alcock 2000). The light curves of lensing events involving the linear motion of a point-like lens in front of a point-like source have a

characteristic shape; any deviations from this shape can be used to infer parameters not described by this simple geometry: parallax (affecting the relative path of source and lens), resolution of the stellar disk, or the presence of companions to the source or lens. In the present context, we are mostly interested in the last of these effects: the detection of "binary lenses" with low-mass secondaries.

The monitoring of gravitational microlensing events is arguably the only method that is capable of detecting Earth-like planets from the ground; this is the most important driver behind the further development of this technique. So far, however, no secure planet detection has been made in this way (Sackett 2000). This chapter introduces the theory of gravitational microlensing, first for a single lens, then for the more complicated case of binary lenses. We will then discuss the available results from current microlensing monitoring experiments.

5.1 Theory of Gravitational Microlensing

Gravitational Lensing by a Single Pointlike Lens

According to the General Theory of Relativity, light from a background source passing by a foreground star of mass M with an impact parameter (minimum distance) $r \gg R_S$ is deflected by an angle

$$\alpha = \frac{4GM}{c^2 r} = \frac{2R_S}{r} \ , \tag{31}$$

where we have introduced the *Schwarzschild radius* $R_S \equiv 2GM/c^2$. Derivations of (31) can be found in any textbook on General Relativity.[20] It was realized soon that light bending could lead to multiple images of the same source (Eddington 1920; Chwolson 1924; Einstein 1936); this effect is called "gravitational lensing".

To describe the geometry of a gravitational lens, we use the following notation: θ is the observed position of the source, θ_S the direction to the source in the absence of lensing, D_S and D_L denote the distances of the source and lens from the observer, and $D_{LS} \equiv D_S - D_L$ the distance from the lens to the source. Simple geometry (see Fig. 23) then leads to the relation:

$$\theta_S D_S = r\frac{D_S}{D_L} - D_{LS}\alpha(r) \, . \tag{32}$$

[20] The famous observational verification of (31) for the bending of light in the gravitational field of the Sun during the total eclipse of 1919 (Dyson et al. 1920) played an important role for the popularization and acceptance of General Relativity. Modern precision measurements of light deflection provide tests of extensions and alternatives of General Relativity.

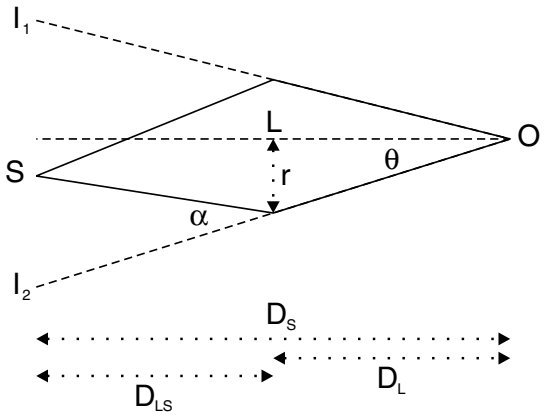

Fig. 23. Geometry of gravitational lensing. Rays from the source S are bent in the lens plane by an angle α, so that the observer O sees two images I_1, I_2

Using $r = D_L\theta$, this can also be written as

$$\theta_S = \theta - \frac{D_{LS}}{D_S}\alpha(r)\,. \tag{33}$$

It is now convenient to introduce the characteristic angle

$$\theta_E \equiv \sqrt{\frac{4GM}{c^2}\frac{D_{LS}}{D_L D_S}}\;; \tag{34}$$

the corresponding characteristic length in the lens plane $r_E \equiv \theta_E \cdot D_L$ is called the *Einstein radius*. Inserting this definition and (31) in (33) gives the *lens equation*

$$\theta^2 - \theta_S\theta - \theta_E^2 = 0\,. \tag{35}$$

This quadratic equation has the two solutions

$$\theta_{1,2} = \frac{1}{2}\left(\theta_S \pm \sqrt{4\theta_E^2 + \theta_S^2}\right)\;, \tag{36}$$

which give the positions of the two images seen by the observer. If the source, lens, and observer lie on a straight line, $\theta_S = 0$, and (36) indicates the presence of two images at positions $\pm\theta_E$. In this case, however, the line containing source, lens, and observer is a symmetry axis. We can use any plane containing this line to draw Fig. 23, and get two images at $\pm\theta_E$ in that plane. The image is thus a circle with radius θ_E around the direction from the observer towards the lens, the so-called "Einstein ring". If $\theta_S \neq 0$, the two solutions of (36) satisfy the inequalities $\theta_1 \geq \theta_E$ and $-\theta_E < \theta_2 < 0$. This means that the two images lie on opposite sides of the observer-lens axis, one of them inside and one of them outside the Einstein ring.

We will see below (40) that if $\theta_S \gg \theta_E$ one of the images is very faint, and the brightness of the other image is hardly affected by the lens. We therefore expect that the lens has noticeable effects only when $\theta_S \lesssim \theta_E$. Then it follows immediately from (36) that the separation of the two images is

$$\theta_1 - \theta_2 = \sqrt{4\theta_E^2 + \theta_S^2} \approx 2\theta_E \,. \tag{37}$$

If we insert typical numbers for observations of stars in the Galactic bulge into (34), we obtain

$$\theta_E = 1.1\,\mathrm{mas} \cdot \left(\frac{M}{M_\odot}\right)^{1/2} \left(\frac{7\,\mathrm{kpc}}{D_S}\right)^{1/2} \left(\frac{D_\mathrm{LS}}{D_L}\right)^{1/2} . \tag{38}$$

The separation of the two images is thus of order a few milliarcseconds. This is usually too small to be resolved, and all we can observe is the combined brightness of the two images (but see Sect. 5.4). It is this situation that is usually called "microlensing".

To compute the observed flux from the images we first note that the surface brightness is not changed by the lensing process.[21] What is affected, though, is the solid angle of the image $\mathbf{\Delta\Omega}$ subtended on the sky. The flux of an infinitesimally small source is simply given by the product of surface brightness and solid angle. The ratio of the observed flux to the flux in the absence of lensing (the "amplification" A, which should more aptly be called "magnification") is thus simply given by $\mathbf{\Delta\Omega}/\mathbf{\Delta\Omega}_S$. Equation (33) defines a mapping from θ_S to θ; the area distortion of this mapping is given by the determinant of the Jacobi matrix J. We therefore get:

$$A_{1,2} = \left(\frac{\mathbf{\Delta\Omega}_S}{\mathbf{\Delta\Omega}_{1,2}}\right)^{-1} = \left.\frac{1}{|\det J|}\right|_{\theta_1,\theta_2} = \left|\frac{\partial\theta_S}{\partial\theta}\right|^{-1}_{\theta_1,\theta_2} . \tag{39}$$

For the calculation of this expression, we introduce the quantity $u \equiv \theta_S/\theta_E$, the source-lens separation in units of the Einstein radius. We have to keep the two-dimensional nature of the lens mapping in mind. Because of the symmetry of the lens, nothing is changed in the direction perpendicular to the plane of

[21] If the curvature radius of space-time is large compared to the wavelength, it can be shown that the photon phase space density along each photon's world line, or equivalently the quantity $I(\nu)/\nu^3$, is conserved. Among the well-known direct consequences are that the bolometric surface brightness of galaxies $I_\mathrm{bol} \propto (1+z)^{-4}$, and that the spectrum emitted by a blackbody (such as the cosmic microwave background) remains a blackbody spectrum with observed temperature $T_\mathrm{obs} = T_\mathrm{em}/(1+z_\mathrm{em})$. For gravitational lensing in our Galaxy we are interested in the special case $z = 0$, i.e., ν is the same for the emitter and observer. For a detailed discussion see Chapter 22 of Misner et al. (1973).

Fig. 23, and the Jacobian can easily be evaluated using polar coordinates; it is given by

$$A_{1,2} = \left| \frac{\theta_{1,2}}{\theta_S} \cdot \frac{d\theta_{1,2}}{d\theta_S} \right| = \frac{u^2+2}{2u\sqrt{u^2+4}} \pm \frac{1}{2} \,. \tag{40}$$

That the first equality is true is also obvious from Fig. 24. The total combined brightness of the two unresolved images is thus

$$A = A_1 + A_2 = \frac{u^2+2}{u\sqrt{u^2+4}} \,. \tag{41}$$

So far we have dealt with a static configuration of a background source being lensed by a foreground object at a normalized projected separation u. Differential Galactic rotation and peculiar motions lead to a relative motion of source and lens, however, with a typical magnitude

$$\dot{\theta} = \frac{v}{D_L} = 12\,\mathrm{mas\,yr^{-1}} \left(\frac{v}{200\,\mathrm{km\,s^{-1}}} \right) \left(\frac{3.5\,\mathrm{kpc}}{D_L} \right) \,, \tag{42}$$

where v is the relative perpendicular velocity of the lens with respect to the source. The typical time scale t_E of a microlensing event is given by the time needed to move by one Einstein radius

$$t_E \equiv \frac{\theta_E}{\dot{\theta}} = 0.13\,\mathrm{yr} \left(\frac{M}{M_\odot} \right)^{1/2} \left(\frac{D_L}{3.5\,\mathrm{kpc}} \right)^{1/2} \left(\frac{D_{LS}}{D_S} \right)^{1/2} \left(\frac{v}{200\,\mathrm{km\,s^{-1}}} \right)^{-1} \,. \tag{43}$$

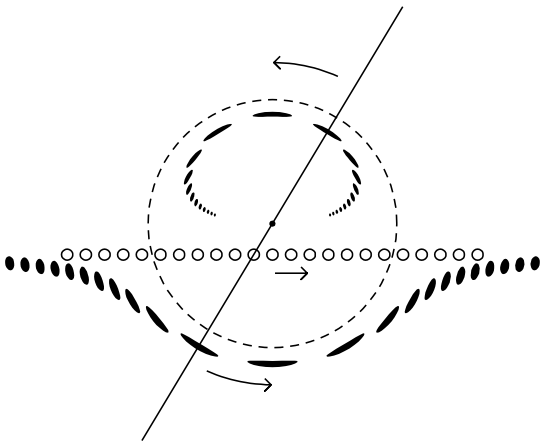

Fig. 24. Location and shape of the two images in a Schwarzschild lens. In this drawing, the lensing mass is indicated with a dot at the center of the Einstein ring, which is marked with a dashed line; the source positions are shown with a series of small open circles; and the locations and the shapes of the two images are shown with a series of dark ellipses. At any instant the two images, the source, and the lens are all on a single line. From Paczyński (1996)

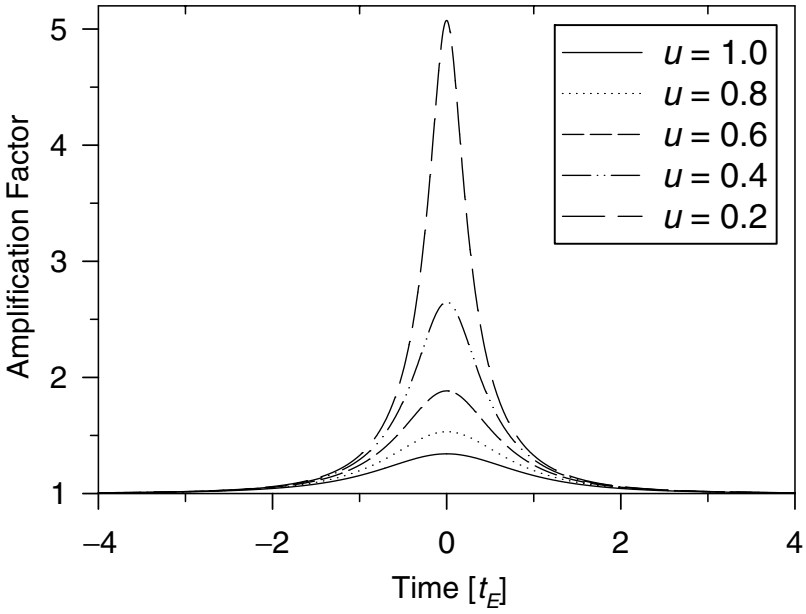

Fig. 25. Microlensing light curves for five different values of the minimum impact parameter. From (41) it follows directly that at the time of crossing the Einstein ring ($u = 1$) the amplification $A = 3/\sqrt{5} = 1.34$, and that the maximum amplification $A(u_{\rm min}) \approx 1/u_{\rm min}$

To first order the motion is linear and can thus be parameterized by:

$$u(t) = \sqrt{(t - t_{\rm min})^2/t_E^2 + u_{\rm min}^2}\,. \tag{44}$$

Here $t_{\rm min}$ and $u_{\rm min}$ are the time of closest approach, and the corresponding impact parameter. Substituting (44) in (41) gives an analytic description of the amplification as a function of time, i.e., of the light curve of the lensing event. Such light curves are shown in Fig. 25 for five different values of $u_{\rm min}$. It is important to note that the light curves of all microlensing events involving a point source and a pointlike lens that moves at a constant rate are completely determined by four parameters: $t_{\rm min}$, t_E, $u_{\rm min}$ and the brightness of the source at $u \to \infty$. Of these parameters, only t_E is related to the properties of the lens. It is therefore only possible to derive a combination of lens mass, distance and transverse velocity in this simple situation.

Lensing Anomalies and Binary Lenses

Real astrophysical sources and lenses are not point sources, of course, and the relative motion is not necessarily rectilinear. We may thus observe lensing "anomalies", i.e., deviations from the simple model described in the previous

section, and attempt to obtain additional information from them. For planet searches, we are mostly interested in binary lenses with very small mass ratio $\mu \equiv m_2/m_1$. The binary lens equation is a straightforward generalization of (33):

$$\vec{\theta}_S = \vec{\theta} - \frac{D_{LS}}{D_S}\left(\vec{\alpha}_1(\vec{r}_1) + \vec{\alpha}_2(\vec{r}_2)\right) , \tag{45}$$

where the two indices 1, 2 refer to the two binary components. We have now written all two-dimensional quantities explicitly as vectors, because there is no plane of symmetry anymore. The analysis of binary lenses is considerably more complicated than that of single lenses. We should expect, however, that source positions for which the determinant of the Jacobi matrix $|\det J| = 0$ have special significance: according to (39) the amplification is infinite for these positions. The locus of the positions for which this condition holds is called a "caustic". The caustic of a point source lens consists only of the point $u = 0$ (see (40)), corresponding to the appearance of an Einstein ring when observer, lens, and source are perfectly aligned. In contrast, the caustics of binary lenses are extended and complicated in shape (Schneider and Weiß 1986; Erdl and Schneider 1993). In the lens plane, the condition $|\det J| = 0$ defines the "critical curves". When the source crosses a caustic at location θ_S, two new highly amplified images appear with positions θ on the corresponding critical curve (or two existing images brighten, merge and disappear if the caustic is crossed in the opposite direction). An example for the case of equal masses of the two binary components is shown in Fig. 26. The left panel shows the structure of the caustic and critical curve, and five possible relative paths of a source with respect to the lens. The source has not been assumed to be pointlike, but rather a uniform disk of diameter $r_s = 0.05\, r_E$. The brightness at any time therefore has to be computed by integrating the amplification as given by (39) over the area subtended by the disk. For each source position along the path, the brightness has been calculated in this way, and plotted versus time in the right panel.

It is apparent from Fig. 26 that a wide variety of light curves are possible for binary lenses (see also Alcock et al. 2000a). The mass ratio μ, the projected binary separation b (in units of r_E), and the angle of the source trajectory with the binary axis are additional free parameters that have to be fit to the observational data. The example of the first binary lens detected, OGLE #7, shows that this can be done fairly well if observations with good sampling and signal-to-noise exist, especially when data points close to caustic crossings are available. OGLE #7 was observed independently by two collaborations (Udalski et al. 1994; Bennett et al. 1995), and the fits to the two disjoint data sets agree well with each other (Alcock et al. 2000a). Additional "anomalous" effects can complicate the interpretation, however. Parallax (due to the annual motion of the Earth around the Sun) leads to a non-linear relative motion of the source and lens, and the orbital motion of the binary can change the caustic structure itself on a time scale comparable to t_E. Confusion, i.e., blending

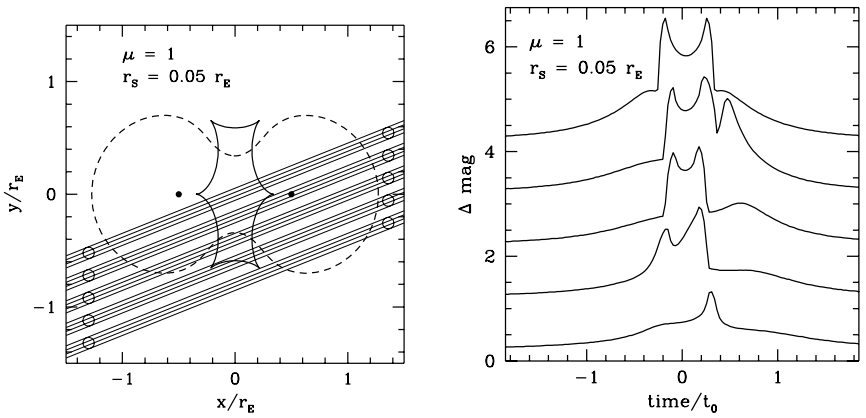

Fig. 26. Microlensing by a binary consisting of two identical point masses, $m_1 = m_2 = 0.5m$, separated by one Einstein ring radius, r_E. The closed figure drawn with a thick solid line in the left panel is the caustic located in the source plane. The closed figure drawn with a thick dashed line is the critical curve. A source placed on a caustic creates an image located on the critical curve. Five identical sources are moving along the straight trajectories, as marked. All sources have radii equal to $r_s = 0.05r_E$, as shown with small open circles. The corresponding light curves are shown in the right panel. The top light curve corresponds to the top trajectory. The sharp spikes are due to caustic crossings by the source. (The light curves are shifted by one magnitude for clarity of the display.) From Paczyński (1996)

with unrelated stars, may adversely affect the photometric measurements, or the source may be a binary star. It is thus necessary to explore the available parameter space fully to avoid misinterpretations of complicated light curves.

5.2 Planetary Systems as Gravitational Lenses

Typical Sizes and Time Scales

A star that has a planetary companion can act as a binary lens with an extreme mass ratio $10^{-6} \lesssim \mu \lesssim 10^{-3}$. Mao and Paczyński (1991) suggested that for $\mu = 10^{-3}$ and a projected separation $\approx r_E$ the detection efficiency should be a few percent, opening the possibility of detecting planets in microlensing surveys. The question about detection thresholds, optimized observing strategies, and the number of expected planet detections has since attracted much interest. Most of the simulations have been done for source stars in the Galactic bulge ($D_S \approx 7 \ldots 8\,\mathrm{kpc}$) lensed either by bulge ($D_L \approx 6\,\mathrm{kpc}$) or disk ($D_L \approx 3 \ldots 4\,\mathrm{kpc}$) stars. It is useful to consider first a few typical numbers for these parameters. The Einstein radius

$$r_E \approx 4\,\mathrm{AU} \left(\frac{M}{M_\odot}\right)^{1/2} \tag{46}$$

is of the order of the orbital radius of Jupiter; this is favorable for the detection of Solar System analogs. The Einstein radius of the planet is related to θ_E by $\theta_p = \sqrt{\mu}\,\theta_E$. One should expect that the influence of the planet is significant over an area with this radius; this is frequently true (Gould and Loeb 1992), but there are important exceptions to this rule (Griest and Safizadeh 1998, see "high-magnification events" below). Assuming for the moment the scaling with $\sqrt{\mu}$, we can derive the planet anomaly duration directly from (43):

$$t_p \approx 2\,\mathrm{days} \left(\frac{m_p}{M_{\mathrm{jup}}}\right)^{1/2} \left(\frac{v}{200\,\mathrm{km\,s^{-1}}}\right)^{-1} . \tag{47}$$

This implies that monitoring with good temporal sampling is required, especially for the detection of Earth-like planets, for which the typical time scale is only a few hours.

A similar scaling argument can be used to estimate the probability that a given planet will actually be detected. At any given time, the probability for amplification by the planet is $\propto (\theta_p/\theta_E)^2 = \mu$, but the total area swept by the planetary Einstein ring while the source sweeps across the Einstein ring of the lensing star is $\propto \theta_p/\theta_E = \sqrt{\mu}$. We thus expect detection probabilities of a few per cent for Jupiter-mass planets ($\sqrt{\mu} \approx 0.03$).

Another important number is the radius of the planetary Einstein ring projected back to the source plane,

$$\tilde{r}_p \equiv \sqrt{\mu}\, r_E \frac{D_S}{D_L} . \tag{48}$$

Numerical values are $\tilde{r}_p \approx 50\,R_\odot$ for $M = M_\odot$ and $m_p = M_{\mathrm{jup}}$, and $\tilde{r}_p \approx 3\,R_\odot$ for $M = M_\odot$ and $m_p = M_\oplus$. These numbers can be compared with the radii of clump giants ($\sim 13\,R_\odot$) and stars near the main-sequence turn-off in the bulge ($\sim 3\,R_\odot$). We see that the effect of the non-zero source size can be safely neglected for Jupiter-like planets, because the star is always much smaller than the planet's Einstein ring radius. For Earth-like planets, however, the radius of the background star is comparable or larger than $\tilde{r}_p$. This means that turn-off stars are much better suited for low-mass planet searches than giants, because in the latter case the planetary anomalies will be strongly smeared out by the large size of the source (see Fig. 30).

Light Curves and Detection Limits

The complicated caustic structure of binary lenses gives rise to a large variety of possible light curves, as discussed above (see Fig. 26); the same is true in the planet case ($\mu \ll 1$). The binary signature is most obvious during caustic crossings, as illustrated in Fig. 27. This figure shows the light curve of a system composed of eight planets with $\mu = 10^{-5}$ located along a straight line, with a source moving with zero impact parameter along this line. Each peak corresponds to the crossing of a planetary caustic; the figure thus demonstrates

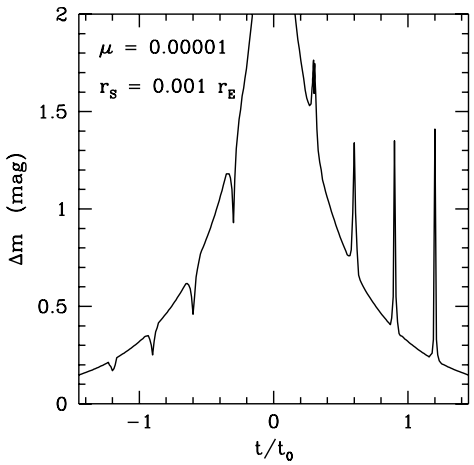

Fig. 27. Simulated light curve of a (very artificial) planetary system, which is made of a star and eight planets, each with mass fraction $\mu = 10^{-5}$, and all located along a straight line. The source with a radius $r_S = 10^{-3} r_E$ is moving along the line defined by the planets, with zero impact parameter. The planets are located at the distances from the star: $b = 0.57,\ 0.65,\ 0.74,\ 0.86,\ 1.16,\ 1.34,\ 1.55,\ 1.76$ in the lens plane, which corresponds to the disturbances in light variations at the times $t/t_0 = -1.2, -0.9, -0.6, -0.3, 0.3, 0.6, 0.9, 1.2$, as shown in the figure. Note that planetary disturbances create local light minima for $b < 1 (t/t_0 < 0)$ and local maxima for $b > 1$ $(t/t_0 > 0)$. From Paczyński (1996)

the effect of Earth-like planets at different separations from the parent star. For more massive planets, significant anomalies can occur even if the source does not cross a caustic (Bolatto and Falco 1994).

To explore the range of possible light curves expected in more realistic cases, one can compute the magnification for every point in the lens plane, and consider "random" paths of the source across this magnification pattern. It is convenient to consider the anomaly $\delta \equiv (A - A_0)/A_0$, where A_0 is the magnification in the absence of planets. This is frequently better than working with A itself, because δ is frequently quite small, especially for small μ. The amplification and corresponding anomaly generated by a Jupiter-like planet are shown in Fig. 28.

The calculation of the magnification pattern and light curves can be repeated for many combinations of the parameters μ (mass ratio), b (projected separation in units of the Einstein radius) and for different source trajectories (see Sackett 1999; Wambsganss 1997). Massive planets are easier to detect, because their anomaly contours cover a larger area on the sky, which makes it more likely that they are intersected by the source trajectory. For a given mass ratio μ, the anomalous regions are largest when $b \approx 1$, i.e., when the star–planet separation is of the order of the Einstein radius (see Fig. 2 in Gaudi and Sackett 2000).

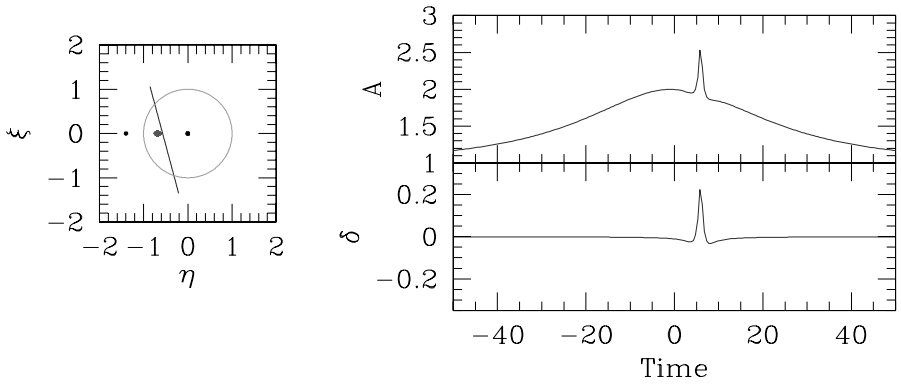

Fig. 28. A background point source travels along a trajectory that just misses the caustic caused by a Jupiter-like planet with mass ratio $\mu = 0.001$ located at $1.3\,r_E$ from its parent star. The three panels show the trajectory and the corresponding amplification A and anomaly δ. The time is given in days. From Sackett (1999)

In the case of planets ($\mu \ll 1$), and if $b \neq 1$, the structure and location of the caustics is quite simple (Griest and Safizadeh 1998). The pointlike single-source caustic becomes a small wedge-shaped caustic still located near the center of the Einstein ring, and one or two new "planetary" caustics appear at locations that depend on the position of the planet. For $b > 1$ there is one planetary caustic located between the lens and the planet, for $b < 1$ there are two planetary caustics on the opposite side of the lens (see Fig. 29).[22] The location x_c of the planetary caustics is approximately given by

$$x_c \approx \left(b^2 - 1\right)/b\,. \tag{49}$$

The caustics are located inside the Einstein ring if $|x| < 1$, i.e., if $1/2\left(\sqrt{5}-1\right) < b < 1/2\left(\sqrt{5}+1\right)$ or $0.618 < b < 1.618$. This region for b is called the "lensing zone"; it has considerable importance for planet searches that are follow-up observations of microlensing surveys. A microlensing event is recognized when the amplification A exceeds a certain value; frequently a source position on the Einstein ring ($u = 1$) corresponding to $A = 1.34$ (41) is used as a detection threshold. If the planetary caustics are located inside the Einstein ring, there is a chance that they will be crossed by the source during the course of the event; if they are located far outside the Einstein ring, however, there will only be a small anomaly during the lensing event. From (46) we therefore

[22] We can now also understand Fig. 27 better. For the planets at $b < 1$, the source passes between the two planetary caustics, located at the opposite side from the star. The amplification in this region is negative, leading to the dips. For the planets at $b > 1$, the source passes through the planetary caustic, which causes the peaks.

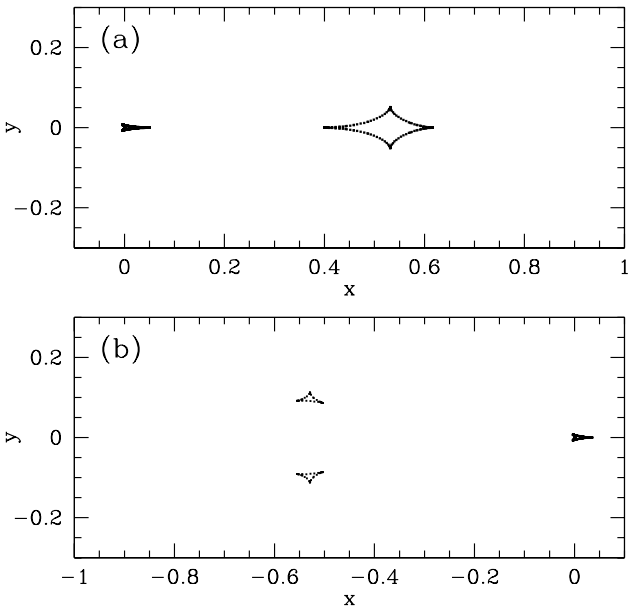

Fig. 29. Caustics for $\mu = 0.003$, showing the central caustic near the origin and the larger planetary caustics. The top panel is for a planet at $b = 1.3$, the bottom panel for $b = 1/1.3 = 0.769$. From Griest and Safizadeh (1998)

see that microlensing is most sensitive to planets with projected separations in the range $2.5 \ldots 6.5\,\mathrm{AU}$ around Solar-type stars, or at $1.4 \ldots 3.5\,\mathrm{AU}$ from $0.3\,M_\odot$ dwarfs. Planets with larger orbital radii may spend some fraction of their orbital time in the lensing zone, depending on the orbital inclination.

The detection of planets in the lensing zone can be hampered by the smearing caused by the non-zero source radius R_S, as mentioned above. If the source covers the entire planetary caustic, $\delta \propto (\tilde{r}_p/R_S)^2$. On the other hand, the typical time scale is longer than given by (47), by a factor $R_S/\tilde{r}_p$. The effect of the finite source size is shown in Fig. 30 for a few typical cases. It is evident from this figure that the ability to detect planets with $\mu \lesssim 10^{-4}$ depends critically on the source radius; the peak deviation from the single-lens light curve is strongly reduced in particular for giant stars.

For each of the light curves in Fig. 30 (and similar light curves for equally probable orientations and impact parameters) we can now ask the question whether the planet would be detected by a photometric monitoring program. The answer will generally be "yes" if $|\delta|$ exceeds a certain threshold (set by the photometric precision) for a minimum time (determined by the time sampling of the photometry). Representative detection probabilities for a model planet system with one planet per factor of 2 in distance from the lens star are listed in Table 8, to illustrate the effect of the source radius. In the real world

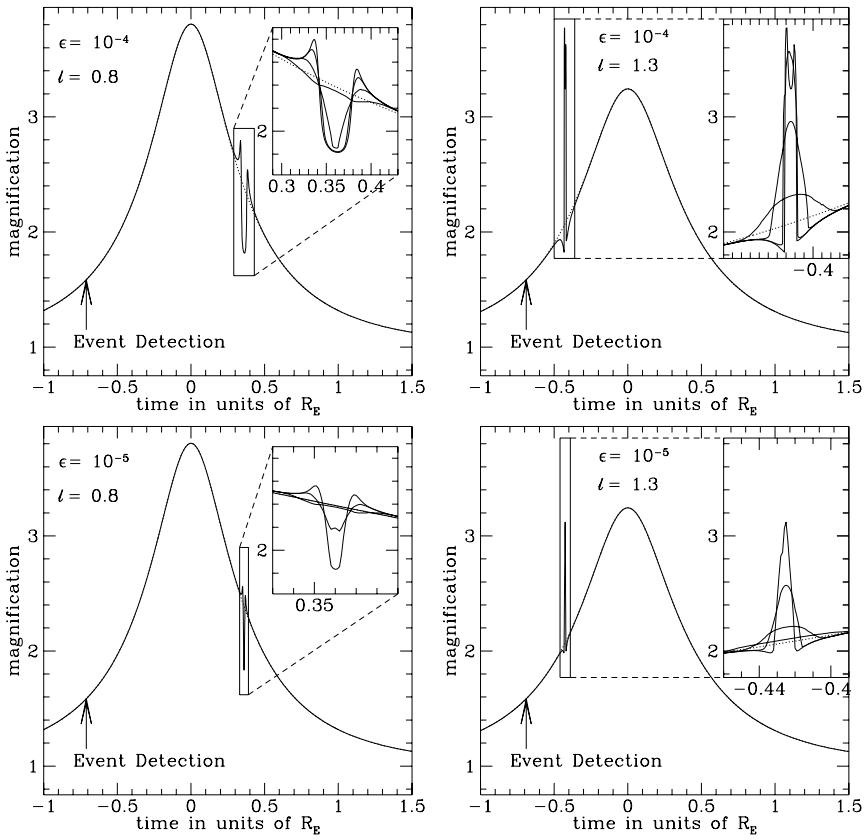

Fig. 30. Theoretical microlensing light curves that show planetary deviations are plotted for mass ratios $\mu = 10^{-4}$ and 10^{-5} and separations of $b = 0.8$ and 1.3. The main plots are for a normalized stellar radius $r_S \equiv (R_S/r_E)\cdot(D_L/D_S) = 0.003$ while the insets show light curves for radii of 0.006, 0013, and 0.03 as well. The amplitude of the maximum deviation from the dotted single-source light curve decreases with increasing r_S. For each of these light curves, the source trajectory is at an angle of $\arcsin 0.6 = 36.^{\circ}9$ with respect to the star–planet axis. The impact parameter $u_{\min} = 0.27$ for the $b = 0.8$ plots and $u_{\min} = 0.32$ for the $b = 1.3$ plots. For these parameters the source trajectory crosses the x-axis near x_c. From Bennett and Rhie (1996)

additional complications are caused by varying seeing conditions, which lead to night-to-night variations of the photometric precision, and by gaps in the data during daytime or due to clouds.

Estimation of Planet Parameters from Microlensing

The next question that needs to be addressed is whether it is possible to use microlensing light curves not only to detect planets, but also to determine

Table 8. Planetary detection probability during microlensing events

r_S	μ	$p(2)\%$	$p(4)\%$	$p(10\%)$	$p(20\%)$
0.003	10^{-4}	0.188	0.144	0.094	0.052
0.006	10^{-4}	0.238	0.159	0.085	0.043
0.013	10^{-4}	0.201	0.118	0.052	0.014
0.03	10^{-4}	0.120	0.035	0.012	0.000
0.003	10^{-5}	0.060	0.034	0.014	0.004
0.006	10^{-5}	0.052	0.026	0.005	0.002
0.013	10^{-5}	0.019	0.008	0.001	0.000
0.03	10^{-5}	0.002	0.000	0.000	0.000

Probabilities p are shown as a function of the threshold for $|\delta|$ and for different values of the normalized source star radius $r_S \equiv (R_S/r_E) \cdot (D_L/D_S)$ and the mass ratio μ. Idealized "factor of 2" planetary systems with one planet per factor of 2 in distance from the lens star are assumed. A planet is considered to be detected if $|\delta|$ is larger than the threshold for a period of time longer than $t_E/400$. The r_S values of 0.003 and 0.006 correspond to a turn-off source star with disk and bulge lenses, respectively, while the r_S values of 0.013 and 0.03 correspond to a giant source with disk and bulge lenses. From Bennett and Rhie (1996)

some of their parameters. First of all, the orbital velocity of planets in the lensing zone is small compared to the typical transverse velocity of the lens with respect to the source ($v \approx 200\,\mathrm{km\,s^{-1}}$). This means that their mass ratio $\mu = (\theta_p/\theta_E)^2$ can be estimated roughly from the duration of the planetary anomaly t_p and the duration of the main event t_E, namely

$$\mu \approx (t_p/t_E)^2\,. \tag{50}$$

The time difference between the peak of the anomaly and the main peak gives an indication of the location of the planetary caustic within the Einstein ring, and thus an estimate of the projected separation b. It is clear from (49), however, that planets with separations b and $1/b$ give rise to caustics at nearly identical positions. The degeneracy between these two cases can be broken by high-quality light curves, because the structure of the magnification pattern near the caustic is different for $b < 1$ and $b > 1$ (compare e.g., the left ($b = 0.8$) and right ($b = 1.3$) panels in Fig. 30). A second degeneracy exists because the source can pass either between the star and the caustic or further from it. This degeneracy is more difficult to break with observational data, but its influence on the estimates for b and μ is relatively small (Gaudi and Gould 1997).

Finite-source effects create additional complications, because they make the duration of the anomaly longer. If we use (50), we will therefore overestimate μ, potentially by a large factor. The differences between the light curves of a large source crossing the caustic of a low-mass planet and of a point source crossing a higher-mass caustic can be very subtle; in some cases photometry with precision much better than 1% (and sufficient time resolution)

would be needed to distinguish between these cases. If this cannot be done, μ may be uncertain by a factor $\sim 1/\delta_{\max}$, or as large as a factor of 20 for anomalies with a maximum deviation $\delta_{\max} = 5\%$ (Gaudi and Gould 1997). If additional information is available, it can be used, however, to alleviate this unsatisfactory situation. For example, the typical time t_S it takes the source to cross a caustic is given by its angular diameter, divided by the relative proper motion of the lens and source. The angular diameter can be estimated from dereddened colors and magnitudes, so it is possible to determine t_S if the relative proper motion can be measured. If $t_S < t_p$, (50) can be used safely to estimate μ, otherwise not. Another possibility is using multi-color (e.g., visible and near-IR) light curves to get a handle on finite-source effects. This idea is based on the fact that stellar limb darkening is generally much stronger at shorter wavelengths. If the source is large compared to the caustic structure, one therefore expects noticeable changes in the color as the center and the limb of the star are amplified by different factors. This color change may be much larger and easier to measure than the details in the shape of the light curve (Gaudi and Gould 1997).

The discussion in the previous paragraphs has tacitly assumed that an observed "blip" in a microlensing light curve is due to a planetary companion of the lens. In practice, it may not at all be easy to establish that this is indeed the case. For example, if the *source* is a binary with a magnitude difference $\Delta V = 5 \ldots 10$ (such as a clump giant primary with a G ... M main-sequence secondary), a single lensing star may produce a light curve mimicking a planetary anomaly. An analysis of the likelihood of such events shows that they may be a significant contaminant in samples of putative planetary lenses, unless precautions are taken to distinguish between the two possibilities; binary sources can be recognized in multi-color light curves, or by spectroscopic follow-up observations (Gaudi 1998).

If all goes well, it is thus possible to identify planetary microlensing anomalies, and to extract μ and b from the observations. What we would really like to know are the mass of the planet m_p and its orbital radius a. With reasonable statistical assumptions about the distribution and velocities of lenses and sources and the measured value of t_E one can estimate the Einstein radius r_E and the lens mass m_L to a factor of ~ 5. It is thus possible to determine $m_p = \mu \cdot m_L$ and a lower limit to $a \geq b \cdot r_E$ with the same uncertainty.

Searching for Planets in High-Magnification Events

We have seen above (Fig. 29) that a binary lens with $\mu \ll 1$ gives rise to a small central caustic as well as one or two larger "planetary" caustics. The size of the central caustic along the x-axis is given to a good approximation by (Griest and Safizadeh 1998)

$$u_c \approx \frac{\mu\, b}{(b-1)^2} \ , \tag{51}$$

provided that b is not close to unity. Since u_c is of order μ (and not $\sqrt{\mu}$ as the size of the planetary caustics) one should think that the central caustic is rather unimportant for planet searches. The arguments and calculations discussed in the previous sections were indeed done for the planetary caustics. Griest and Safizadeh (1998) pointed out, however, that observations that concentrate on high-magnification events offer a good chance to detect planets due to the proximity of the source path to the central caustic. The argument is based on (41): if $A \gg 1$, then $u \approx 1/A$. If there is a shift or distortion of the caustic of size du due to a planet, we will observe an anomaly $\delta \equiv dA/A \approx -Adu$. For planets in the lensing zone $0.618 \leq b \leq 1.618$ we thus expect deviations of order

$$\delta \approx u_c A \approx \mu A \ , \tag{52}$$

where we have somewhat pessimistically set $b/(b-1)^2 \approx 1$. For example, because all values of $u_{\rm min}$ are equally probable, 5% of all microlensing events will have $u_{\rm min} \leq 0.05$ and $A_{\rm max} \geq 20$. A $\mu = 10^{-3}$ planet will thus produce a 2% anomaly near the peak of such events, which should not be too hard to detect. More detailed simulations show indeed that the detection probability for such planets is close to 1 (Griest and Safizadeh 1998).

The very high detection rate for planets during high-magnification events has two important consequences. First, if no anomaly is observed in a well-sampled light curve, the presence of Jupiter-like planets in the lensing zone can be safely excluded. Monitoring of a modest number of such events can thus establish the abundance of such planets. Second, if the lensing star has multiple planets in the lensing zone, each one of them will cause a detectable distortion of the central caustic. The resulting light curve will likely be complicated and difficult to interpret, but there is a good chance that systems with multiple planets can be found in this way (Gaudi et al. 1998). This is not the case in the "traditional" approach, because it would require a fortuitous alignment of two planets for the source to cross the planetary caustics of both of them.

Gravitational Lensing of Planets

We have discussed above that binary sources may be a significant problem for searches for planetary companions to the lens. We can of course turn the argument around and ask whether it is possible to take advantage of the situation if it is not the lens, but rather the source that is accompanied by a planet. In this case the star and the planet are both amplified by the same lens, but if there are any caustic crossings, they will occur at different times for the star and the planet. The peak amplification of the planet is larger because of the smaller radius of the planet. For Jupiter-size planets in close orbits (0.05 AU) with near-unity albedo, the maximum fractional deviation of the light curve above that expected when the source star does not have a planetary companion can get close to 1% (Graff and Gaudi 2000). The

typical time it takes the planet to cross the caustic is $\lesssim 1$ h. It should be possible to search for these "blips" near the caustic crossing time of the parent star, but a large amount of observing time on big telescopes would be needed. Planets with much larger orbital radius a cannot be detected in this way, because the amount of light reflected by the planet scales with a^{-2}. (Note that now the anomaly is caused by the light, not the mass of the planet.)

Observations of lensed planets with future giant telescopes (which are needed to get good SNR and time resolution for fairly faint sources) could reveal a number of interesting effects. The shape of the illuminated fraction of the planet has a strong influence on the light curve; crescents can produce higher peak amplification than half or full disks, partly offsetting the smaller fraction of reflected light (Ashton and Lewis 2001). Reflection by condensed particles in the planetary atmosphere leads to partial polarization of the light from the planet (Seager et al. 2000); the amplification of the planet with respect to the star during a caustic crossing may in favorable cases enhance the total polarization to a detectable level (a fraction of a percent, Lewis and Ibata 2000). Lensing of crescent-shaped planets should also lead to characteristic polarization signatures. Studies of these phenomena could provide information on the particles in the planets' atmospheres and complement observations of "reflected light" from planets in the Solar neighborhood (Sect. 6.6).

5.3 Microlensing Planet Searches

Search Strategies

The probability that a given star in the Galactic bulge is being lensed at a given time is very low (about 10^{-6}). Surveys that monitor a very large number of stars are therefore necessary to detect the occasional brightening of a star due to microlensing. Several such projects were launched in the 1990s, and have now detected more than 1,000 microlensing events: EROS (Aubourg et al. 1993; Derue et al. 2001), OGLE (Udalski et al. 1993, 2000), MACHO (Alcock et al. 1993, 2000b), and MOA (Bond et al. 2001). The temporal sampling – typically one observation per night – of these surveys is inadequate to find planetary anomalies directly. They issue alerts of ongoing events, however, allowing more frequent follow-up observations by teams that have formed specifically for this purpose, for example PLANET (Albrow et al. 1998), GMAN (Alcock et al. 1997), and MPS (Rhie et al. 1999). The PLANET collaboration, for example, uses 1 m class telescopes located in Chile, Tasmania, Australia, and South Africa to achieve round-the-clock coverage of selected ongoing events. It is clear that southern sites are preferred for the monitoring of fields in the Galactic bulge, but because of the large number of small telescopes in the northern hemisphere, and their favorable longitude distribution, searches for Jupiter-like planets from the north are also feasible (Tsapras et al. 2001).

As we have seen in Sect. 5.2, even Earth-mass planets produce large anomalies under favorable conditions (caustic crossings, small background star). Monitoring projects that achieve $\sim 1\%$ photometric precision with $\sim$hourly sampling therefore have sufficient *sensitivity* to detect planets over a large range of masses, but the *efficiency* depends sensitively on the planet mass, and on the detailed photometric performance – most detectable planetary anomalies result from non-caustic crossing events (Gaudi and Sackett 2000). The predicted number of planets that should be detected by monitoring projects depends strongly on the assumptions made, including how planetary masses and separations vary with lens mass (Peale 1997). Since the detection efficiency for any given event depends strongly on $u_{\min}$, it is important to know the distribution of this parameter in the observed sample of microlensing events (Gaudi and Sackett 2000). The chances to find planets with follow-up monitoring can be substantially increased if the original survey produces a large number of high-amplification events (Bond et al. 2002a).

The two-step strategy – search for microlensing using wide-field cameras, and follow-up with targeted observations of "alerted" events – offers currently the best chances to detect planets through microlensing anomalies. This approach has the disadvantage that the high-quality follow-up data are obtained only after the alert; on the rising wing of the event only the monitoring observations are available. The lack of densely sampled data for the first part of the light curve hampers the ability to discriminate between planets and other types of anomalies. In the future it may be possible to use one and the same experiment to detect and monitor microlensing events by conducting frequent observations of a large sample of stars. In such a survey with uniform time coverage it is of course possible to reconstruct the past behavior of any "interesting" star. Next generation of dedicated survey telescopes equipped with wide-field cameras, such as the VLT Survey Telescope, could conduct an efficient survey for low-mass planets, whose anomalies last only a few hours. The largest difficulty of such a project are the gaps in the light curves during daytime and periods of bad weather.

This problem could be overcome by an orbiting telescope, for example the proposed Galactic Exoplanet Survey Telescope (GEST), Bennett and Rhie 2000). A diffraction-limited 1.5 m telescope with a 1° field-of-view and a gigapixel CCD array could monitor $\sim 2 \cdot 10^8$ stars in the Galactic bulge, and observe $\sim 12,000$ microlensing events during a 2.5 yr mission lifetime. A mission like GEST could detect 10 to 20 Earth-mass planets at 1 AU separation if all stars have such companions. The detection efficiency is even better at somewhat larger separations, and thousands of gas-giant planets could be found. A microlensing survey from space would thus be a powerful way to determine the abundance of terrestrial and giant planets in the Galaxy.

Putative Planet Detections

A few claims of microlensing planet detections have appeared in the literature, but none of them has stood up to further scrutiny. It is nevertheless instructive

to take a look at a few examples, because we can see some of the difficulties that anyone attempting to establish the planetary nature of a microlensing anomaly will face.

MACHO 97-BLG-41 was a very unusual event with a complicated light curve (see Fig. 31), which clearly indicates a multiple lens, but cannot be modeled with static binary models. Bennett et al. (1999) interpreted the caustic structure as coming from a triple system consisting of a stellar binary with $\sim 1.8\,\mathrm{AU}$ separation, orbited by a Jovian planet ($m = 3.5 \pm 1.8 M_{\mathrm{jup}}$) at $\sim 7\,\mathrm{AU}$. The PLANET collaboration showed, however, that their own data on this event could be modeled by a normal binary, whose orbital motion changes the orientation and separation of the two stars between the times of the two caustic crossings (Albrow et al. 2000b). Furthermore, this model also provides a stunningly good fit to the MACHO/GMAN data (Fig. 31, right panel), on which Bennett et al. (1999) had based their claim of a planet detection. This alone does not disprove the existence of a Jovian planet in this system, but the PLANET data are inconsistent with the particular model of Bennett et al. (1999), and the existence of a simple, plausible binary model that explains all data on MACHO 97-BLG-41 strongly suggests that this is the correct interpretation of this event.

The microlensing event MACHO 98-BLG-35, which reached a peak amplification factor of almost 80, was monitored intensely by the MOA, MPS, and PLANET teams. Based on the MPS and MOA data, Rhie et al. (2000) reported evidence for a planet with mass fraction $4 \cdot 10^{-5} \leq q \leq 2 \cdot 10^{-4}$.

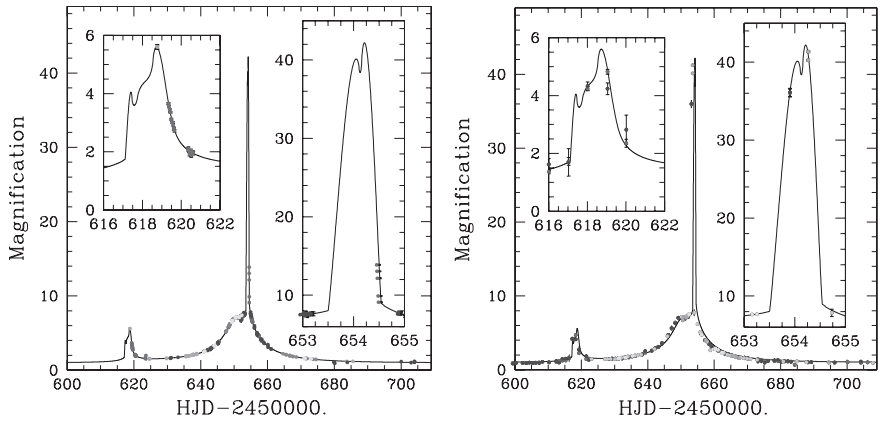

Fig. 31. Rotating binary model for MACHO 97-BLG-41, with data points from PLANET (*left panel*) and from MACHO/GMAN (*right panel*). The fit was obtained using only the PLANET data; the MACHO/GMAN data did not enter the fit and were simply superposed to the model in the right-hand panel. Nevertheless, the model reproduces the MACHO/GMAN data extremely well even in regions where no PLANET data are available. From Albrow et al. (2000b)

A reanalysis of the same observations with an improved photometric algorithm, and inclusion of additional PLANET data, showed, however, that a planet with these parameters can be ruled out (Bond et al. 2002b). This reanalysis revealed apparent lower-level anomalies, which can be fitted by models with one, two or three planets, all with masses $q < 3 \cdot 10^{-5}$. The fact that several different models give fits of similar quality raises the suspicion that they actually fit noise in the data. The problem is that the inclusion of planets lowers the χ^2 by a formally significant amount, but systematic deviations mimicking small planetary anomalies may well be present in the light curves (Gaudi et al. 2002). The evidence for one or more planets associated with MACHO 98-BLG-35 therefore remains tentative at best.

A convincing fast anomaly was observed in MACHO 99-BLG-47 (Albrow et al. 2002). According to (50) the short duration of the anomaly should indicate small secondary mass, but in this case the light curve can be modeled better with a binary star, in which both components have nearly equal masses and either a very small or very large separation (compared to θ_E). Albrow et al. (2002) also show that this interpretation is much more likely than a solution with a planet, which would require rather extreme parameters for the peak amplification, event duration, and blending.

We are thus led to conclude that the ambiguities and degeneracies mentioned in Sect. 5.2 can easily conspire with observational uncertainties to mimic planetary anomalies. Establishing a secure planet detection will require an extremely careful analysis. For Jupiter-mass planets, the main challenge is exhausting the full parameter space in modeling complex light curves (e.g., due to rotating binaries). Terrestrial planets tend to produce anomalies at the threshold of statistical significance, and small random wiggles in the light curves can easily give rise to spurious detections.

Limits on the Abundance of Planets

Whereas microlensing observations have not been successful yet at making a convincing planet detection, they can nevertheless be used to establish useful statistical limits on the abundance of massive planets in the Galaxy. The starting point is the exclusion of planetary anomalies at a certain significance level from a well-sampled microlensing light curve. High-magnification events are well-suited for this purpose: since *every* planet in the lensing zone gives rise to an anomaly close to the peak, the absence of any such anomaly proves the absence of planets in the lensing zone (to a certain mass threshold, which depends on the photometric errors). This argument was used by Rhie et al. (2000) to conclude that there could not be any Jupiter-mass planets in the lensing zone of the event MACHO 98-BLG-35, which has already been discussed above. Similar, somewhat weaker constraints could be placed on the existence of companions in OGLE 1998-BUL-14 (Albrow et al. 2000a).

This argument can of course be extended from individual cases to a combined analysis of a well-understood sample of microlensing events (Gaudi et al. 2002). The first step in this analysis is the selection of a clean sample of events, based on criteria that reject events with sparse light curves, poorly determined parameters, or non-planetary anomalies. For each "good" event, one then searches for deviations of the light curves from the best-fitting point-source/point-lens (PSPL) model. This is done through an exhaustive search of the parameter space of possible binary models and source trajectories, followed by a χ^2 analysis. For each set of parameters $\mathcal{P}$ a synthetic light curve is computed and compared with the data, giving $\chi^2_{\mathcal{P}}$. If $\chi^2_{\mathcal{P}}$ was significantly smaller than $\chi^2_{\rm PSPL}$, i.e., $\chi^2_{\mathcal{P}} - \chi^2_{\rm PSPL} < -\Delta\chi^2_{\rm thresh}$, we would conclude that we have found a planet with parameters $\mathcal{P}$. On the other hand, if $\chi^2_{\mathcal{P}}$ is significantly larger than $\chi^2_{\rm PSPL}$, we can rule out the existence of such a planet. By integrating over the possible source trajectories we can then determine the probability with which we would have detected a planet with given projected separation b and mass ratio μ. After repeating this procedure for each event in the sample, we can determine the maximum fraction of stars f that can have planets with parameters b and μ, which is still consistent with the non-detections at a certain confidence level (see Fig. 32). The reliability of the result of this procedure clearly depends on the correct modeling of subtleties like finite-source effects, and on the adoption of a realistic threshold $\Delta\chi^2_{\rm thresh}$ at which differences in χ^2 are regarded "significant".

Five years of photometric data collected by the PLANET collaboration have been analyzed in this way (Albrow et al. 2001; Gaudi et al. 2002). Of all observed events, 43 fulfill the selection criteria used by the authors and form the basis of the statistical arguments. At 95% confidence, less than 25% of the lenses have companions with mass ratio $\mu = 10^{-2}$ in the lensing zone (see Fig. 32). With the help of a model for the mass, velocity and space distribution of bulge lenses, this result can be converted to a statement about Jupiter-mass companions of M dwarfs in the Galactic bulge: less than 33% of the $\sim 0.3\,M_\odot$ stars have companions with $m_p \geq M_{\rm jup}$ and $1.5\,{\rm AU} < a < 4\,{\rm AU}$.

5.4 Astrometric and Interferometric Observations of Microlensing Events

The Photocenter of a Single Lens

Our discussion of gravitational microlensing has so far focused on the observable change in the combined brightness from all images. A look at Fig. 24, however, suggests that the change of position with time may also be detectable. If the resolution is insufficient to separate the two images, an astrometric observation will measure the position of the "center of light". To compute the deviation $\Delta\theta$ from a straight-line motion, we add the positions of the two images, weighted by their respective brightness, and subtract the position of the source in the absence of lensing, $\theta_E u$. From (36) and (40) we thus get

$$\Delta\theta = \frac{1}{A}\left(\theta_1 A_1 + \theta_2 A_2\right) - \theta_E u = \frac{u}{u^2+2}\theta_E\,. \tag{53}$$

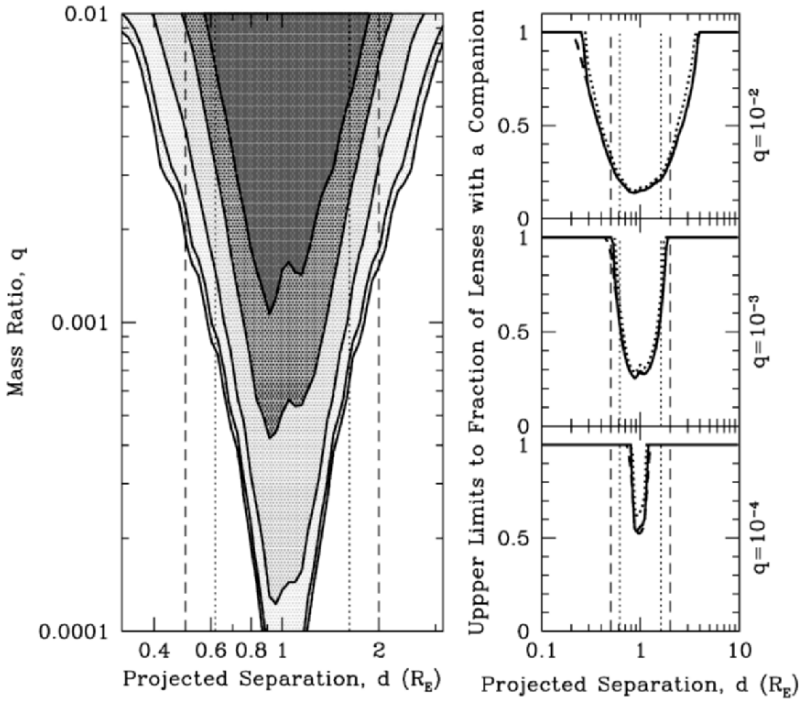

Fig. 32. *Left panel*: Exclusion contours (95% confidence level) for the fractions of primary lenses with a companion derived from the PLANET sample of 43 events, as a function of the mass ratio and projected separation of the companion. Solid black lines show exclusion contours for $f = 75\%$, 66%, 50%, 33%, and 25% (outer to inner). The dotted and dashed vertical lines indicate the boundaries of the lensing zone and extended lensing zone, respectively. *Right panel*: Cross sections through the left panel, showing for three different mass ratios the upper limit to the fraction of lenses with a companion as a function of projected separation. The solid line is derived from the point-source efficiencies with a threshold of $\Delta\chi^2_{\rm thresh} = 60$. The dotted line is derived from the point-source efficiencies with a threshold of $\Delta\chi^2_{\rm thresh} = 100$. The dashed line is finite-source efficiencies with a threshold of $\Delta\chi^2_{\rm thresh} = 60$. The dotted vertical lines indicate the boundaries of the lensing zone $0.6 \leq d \leq 1.6$. The dashed vertical lines indicate the extended lensing zone, $0.5 \leq d \leq 2$. From Gaudi et al. (2002)

The function $u/(u^2 + 2)$ has a maximum for $u = \sqrt{2}$; the corresponding astrometric deviation is

$$\Delta\theta_{\rm max} = \frac{1}{2\sqrt{2}}\,\theta_E \approx 0.4\,{\rm mas}\,, \tag{54}$$

where we have used the numerical estimate from (38). All microlensing events with $u_{\rm min} \leq \sqrt{2}$ therefore produce astrometric signatures with a peak amplitude that depends only on θ_E and has a value ($\sim$0.4 mas) that is well within

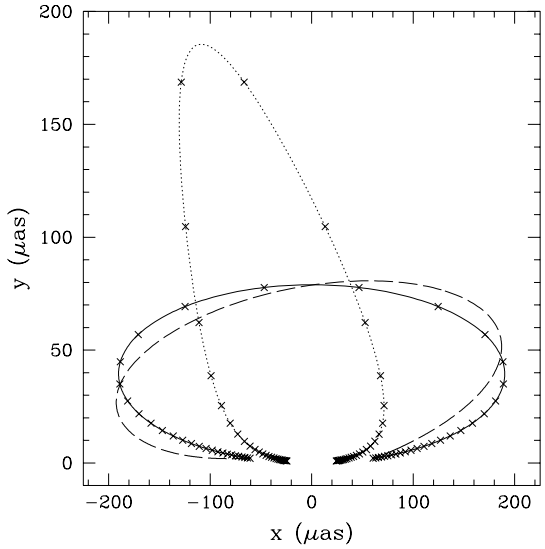

Fig. 33. Astrometric microlensing of a single star. The solid line shows a simple single-lens curve with $u_{\rm min} = 0.3$, and $t_E = 40\,$days. The curve is plotted over one year, with x's marking each week, so only the 5 or so weeks at largest y have magnification greater than 1.34. The dashed line shows the same with an example parallax effect included, while the dotted line shows the effect of blending (blend fraction $f_b = 60\%$). From Safizadeh et al. (1999)

reach of precise astrometric methods (see Sect. 9). It is not difficult to show (e.g. Boden et al. 1998) that the two-dimensional motion $\Delta\vec{\theta}$ is an ellipse with eccentricity

$$e = \sqrt{\frac{2}{u_{\rm min}^2 + 2}}\ ; \tag{55}$$

for very small $u_{\rm min}$ the motion becomes nearly one-dimensional ($e \to 1$). The solid line in Fig. 33 shows this ellipse for the case $u_{\rm min} = 0.3$; note that the two axes in this figure have different scales, and that the motion of the photocenter is fastest at the time of closest approach. Parallax and blending (i.e., contributions from the lens or from unrelated nearby stars to the total light) lead to distortions of this simple shape, as illustrated in Fig. 33. The combined analysis of photometric an astrometric information can help to resolve some of the degeneracies between the possible source, lens, and planet parameters pointed out in Sect. 5.1 and 5.2 (Han 2002).

Planetary Signatures

It is to be expected, of course, that planetary companions of the lens lead to modifications of the astrometric signature, similar to those of the light curve.

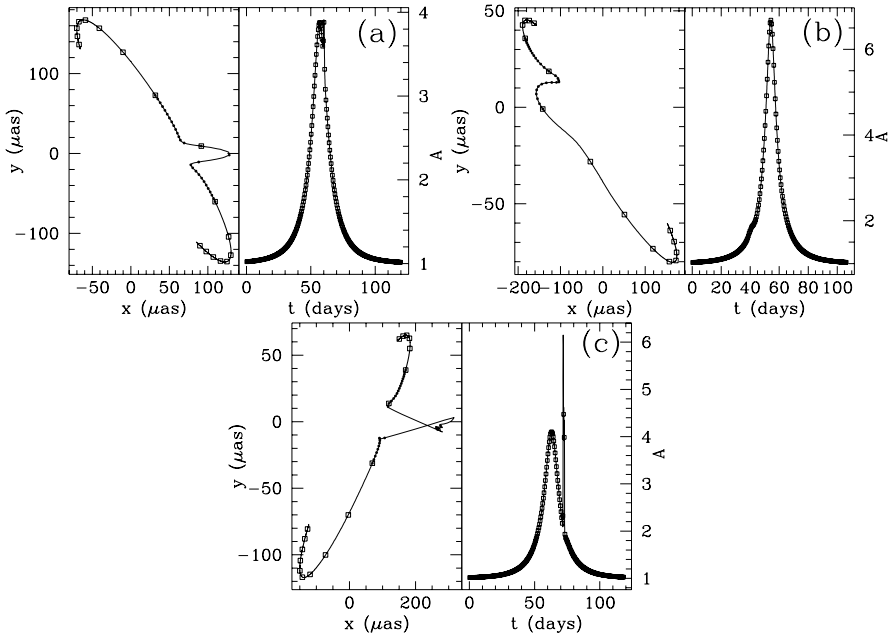

Fig. 34. Some examples of planetary astrometric and photometric curves. All examples assume $\mu = 10^{-3}$, with a primary lens angular Einstein radius of 550 μas, which corresponds to a Saturn-mass planet orbiting a $0.3\,M_\odot$ star. Panel (**a**) has $b = 1.3$, panel (**b**) has $b = 0.7$, and panel (**c**) shows a caustic-crossing event with $b = 1.3$. In the astrometric panels one square is plotted per week, so the durations of the deviations are of order a few days. Dots are plotted every 12 h during the deviation. From Safizadeh et al. (1999)

This is indeed the case, as can be seen in Fig. 34, which displays astrometric and photometric curves for a Saturn-mass planet. The excursions due to the planets are much shorter in duration than the overall microlensing events; this should simplify distinguishing them from the global distortions of the astrometric motion due to parallax and blending effects. The signature of the planet is particularly strong if a caustic-crossing occurs.

As for microlensing light curves, finite-source effects tend to smear out the planetary signal; this is important especially for low-mass planets (Safizadeh et al. 1999, see Fig. 35). If the source star is not too large, caustic-crossing events reach peak deviations of a few hundred μas, but only for a very short time. The peak amplitude of events for which no caustic crossings occur is much smaller. Still, the detection probability for Saturn-mass planets (for which finite-source effects are not important) in the lensing zone is quite high, provided that an astrometric accuracy of a few μas can be achieved. It should thus be possible in principle to search for planetary events with the Space Interferometry Mission (Sect. 9.6). The best observing strategy will probably

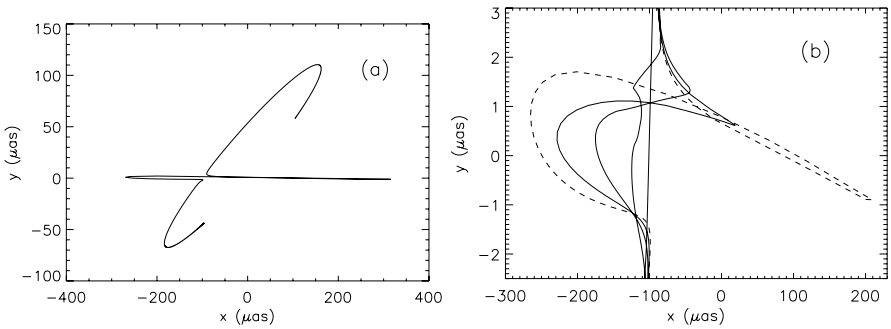

Fig. 35. Astrometric motion for Earth-mass caustic crossing. Panel (**a**) shows the center-of-light motion for a point source, crossing a caustic associated with an Earth-mass planet at $b = 0.825$. The primary lens is $0.3\,M_\odot$ at $D_L = 4\,\mathrm{kpc}$, and source at $D_S = 8\,\mathrm{kpc}$. Panel (**b**) shows a close-up view of the planetary deviation, with finite-size source. The dotted line plots the center-of-light motion for a $1\,R_\odot$ source. The solid lines depict the center-of-light motion for more realistic sizes typical of Galactic bulge stars, namely 3, 5, 9, and $30\,\mathrm{R}_\odot$. Note the extreme anisotropy of the axes on the graph. For $t_E = 40\,\mathrm{days}$ the duration of the deviation is about 20 h, with the center of the source spending roughly 90 min inside the caustic. From Safizadeh et al. (1999)

be obtaining dense temporal sampling close to the peak of high-magnification events, because this gives a relatively high probability of caustic crossings at a time that can be predicted several days in advance.

Resolution of the Individual Images

The separation of the two individual images in a microlensing event (37) and (38) is comparable to the resolution achievable with a long-baseline interferometer. For example, an interferometer with a baseline length of 200 m operating in the H band (1.6 μm) has a fringe spacing $\lambda/B = 1.6\,\mathrm{mas}$. It thus seems possible to fit "binary" models to interferometric data and to determine the separation and flux ratio of the two images (Delplancke et al. 2001). The principal challenge of such observations is the relative faintness of the lensed stars (even while they are being magnified), which necessitates using a brighter star within the isoplanatic radius to co-phase the interferometer (see Sect. 9.5). For small impact parameter u, the images become noticeably distorted (see Fig. 24), approaching the Einstein ring for $u \to 0$. The VLT interferometer could provide sufficient sensitivity and uv plane coverage to produce true images showing these effects in favorable cases.

Interferometric imaging could also reveal the appearance and disappearance of image pairs during the crossings of planetary caustics. Measuring the individual motions and fluxes of the images in planetary microlensing events could certainly remove most of the ambiguities and uncertainties that arise in

the interpretation of light curves. Detailed simulations will be needed to determine the resolution, response time, imaging speed, and sensitivity required for such observations. The capabilities of the present interferometer arrays are likely not sufficient, but obtaining "movies" of planetary lensing events could be an interesting addition to the science case for a future large interferometric facility (e.g., Ridgway and Roddier 2000).

6 Planetary Transits and Searches for Light Reflected by Planets

If a planetary system happens to be oriented in space such that the orbital plane is close to the line-of-sight to the observer, the planets will periodically transit in front of the stellar disk. Photometric or spectroscopic observations of these eclipses can be used to infer orbital and physical parameters of the planets. The first part of this chapter deals with the basic parameters of transits that can be derived from simple geometric considerations. Summaries of ongoing observing programs, and of photometric space missions that are currently under development, follow in the next sections. The last part of the chapter discusses the prospects of detecting the light reflected by extrasolar planets without spatially resolving them from their parent stars.

6.1 The Geometry of Transits

The Probability of Transits

The first question that we would like to answer is about the probability that eclipses occur in a set of planetary systems with randomly oriented equatorial planes. For simplicity, the following discussion is restricted to circular orbits, although the case of strongly eccentric orbits would certainly be relevant, too (see Sect. 3.3). It is obvious that eclipses occur if and only if

$$a \cos i \leq R_* + R_p \ , \tag{56}$$

where a and i are the orbital radius and inclination, and R_* and R_p the radii of the star and planet. In a set of randomly oriented orbits $\cos i$ is distributed between 0 and 1, as already mentioned in Sect. 4.1. The probability $p_{\rm trans}$ that transits occur therefore follows directly from (56),

$$p_{\rm trans} = \frac{R_* + R_p}{a} \approx \frac{R_*}{a} \, . \tag{57}$$

Numerical values of the transit probability for the planets in the Solar System are listed in Table 9. Typical values range from $\approx 5 \cdot 10^{-3}$ for the terrestrial planets to a few times 10^{-4} for the gas giants. Together with the time between

Table 9. Transit probabilities, maximum durations, and depths for the planets in the Solar System as seen by a distant observer

planet	probability	duration [h]	depth
Mercury	$1.2 \cdot 10^{-2}$	8	$1.2 \cdot 10^{-5}$
Venus	$6.4 \cdot 10^{-3}$	11	$7.6 \cdot 10^{-5}$
Earth	$4.7 \cdot 10^{-3}$	13	$8.4 \cdot 10^{-5}$
Mars	$3.1 \cdot 10^{-3}$	16	$2.4 \cdot 10^{-5}$
Jupiter	$8.9 \cdot 10^{-4}$	30	$1.1 \cdot 10^{-2}$
Saturn	$4.9 \cdot 10^{-4}$	40	$7.5 \cdot 10^{-3}$
Uranus	$2.4 \cdot 10^{-4}$	57	$1.3 \cdot 10^{-3}$
Neptune	$1.5 \cdot 10^{-4}$	71	$1.3 \cdot 10^{-3}$
Pluto	$1.2 \cdot 10^{-4}$	82	$2.7 \cdot 10^{-6}$

successive transits of each planet – which is obviously equal to the orbital period P – these numbers elucidate the main difficulty of searches for planetary occultations: thousands of stars have to be monitored for many years in order to find a few eclipses, if no prior knowledge about the orbital inclination for individual systems is available. The two main exceptions to this rule will be discussed in Sect. 6.4. These are searches for "hot Jupiters", which have small a and P and thus high transit probability and short intervals between occultations, and searches in binary systems with known inclination of the binary orbit, which assume that potential planets may likely be coplanar with the stellar pair.

Transit Duration

The next question to be addressed is how long transit events for a given planet last. The transit duration $t_{\rm trans}$ is given by the expression

$$t_{\rm trans} = \frac{P}{\pi} \arcsin \left(\frac{\sqrt{(R_* + R_p)^2 - a^2 \cos^2 i}}{a} \right) , \tag{58}$$

where P is the orbital period of the planet. (The expression within the square root follows from Pythagoras' Theorem, the projection of the relevant segment of the orbit beeing approximately straight.) For $a \gg R_* \gg R_p$ (58) can be simplified to

$$t_{\rm trans} = \frac{P}{\pi} \sqrt{\left(\frac{R_*}{a} \right)^2 - \cos^2 i} \, . \tag{59}$$

The maximum values for the Solar System (corresponding to $i = 90°$) are again tabulated in Table 9; they range from a few hours to a few days, which is quite favorable for monitoring campaigns.

Transit Depth and Shape of the Light Curve

The variation of the stellar brightness during the eclipse is clearly very important because it sets the photometric precision that must be achieved to detect the transit. To first approximation, in which we can treat the star as a disk of uniform brightness, the relative change of the observed flux $\Delta\mathcal{F}/\mathcal{F}$ is given by

$$\frac{\Delta\mathcal{F}}{\mathcal{F}} = \frac{\pi R_p^2 \mathcal{B}_*}{\pi R_*^2 \mathcal{B}_* + \pi R_p^2 \mathcal{B}_p} \approx \left(\frac{R_p}{R_*}\right)^2 , \tag{60}$$

where $\mathcal{B}_*$ and $\mathcal{B}_p$ are the surface brightness of the star and planet, respectively; in almost all cases of interest $\mathcal{B}_p \ll \mathcal{B}_*$. For the secondary eclipse (when the planet is behind the star), the numerator of (60) has to be replaced with $\pi R_p^2 \mathcal{B}_p$. The secondary eclipse is therefore a factor $\mathcal{B}_p/\mathcal{B}_*$ shallower than the primary eclipse, which means that it is normally much more difficult to observe.

To compute the shape of the light curve during the ingress and egress of the eclipse, we first define $x \equiv d - R_*$, where d is the projected separation of the planet from the star. The time dependence of d is

$$d(t) = a\sqrt{\sin^2 \omega t + \cos^2 i \cos^2 \omega t} , \tag{61}$$

with $\omega \equiv 2\pi/P$. In the uniform disk approximation, the change of the brightness is proportional to the fraction of the disk covered by the planet. If we make the additional assumption that $R_p \ll R_*$, the segment of the limb of the star across the planet can be regarded as a straight line. Half of the occulted area is then given by the "pie slice" with angle α, minus the (approximate) triangle with sides x and R_p shown in Fig. 36. We therefore get

$$\begin{aligned} A_{\rm cov} &\approx 2 \cdot \left(\tfrac{1}{2}\alpha R_p^2 - \tfrac{1}{2} x \sqrt{R_p^2 - x^2}\right) \\ &= R_p^2 \arccos\left(\frac{x}{R_p}\right) - x\sqrt{R_p^2 - x^2} . \end{aligned} \tag{62}$$

(Note that during ingress and egress $-R_p \le x \le R_p$.) Inserting $x(t) = d(t) - R_*$ from (61) then gives the desired shape of the light curve.

If the planet cannot be regarded as small compared to the star, or if the effects of limb darkening are taken into account, one has to perform an integration over the occulted part of the stellar disk, which is best done in polar coordinates. So our goal is to integrate twice the length of the dotted arc in Fig. 36 from $d - R_p$ to R_*. The length of this arc is $r_*\beta$. Application of the Law of Cosines to the triangle formed by R_p, d, and r_* gives

$$R_p^2 = d(t)^2 + r_*^2 - 2 r_* d(t) \cos\beta . \tag{63}$$

We can therefore formally write

$$A_{\rm cov} = 2 \int_{\max(0, d(t) - R_p)}^{\min(R_*, d(t) + R_p)} r_* dr_* \arccos\big[\Theta(t)\big] , \tag{64}$$

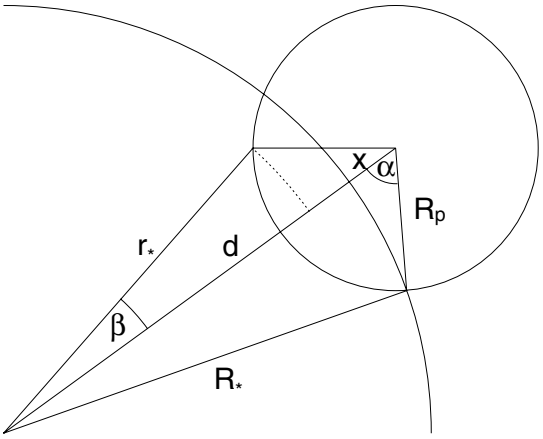

Fig. 36. The area occulted by a planet and definition of the geometric quantities used in the computation of transit light curves

where

$$\Theta(t) \equiv \begin{cases} \frac{d^2(t)+r_*^2-R_p^2}{2r_* d(t)} & \text{for } r_* > R_p - d(t) \\ -1 & \text{otherwise}\,. \end{cases} \tag{65}$$

The shape of the "dip" in the light curve can then be computed numerically from

$$\frac{\Delta\mathcal{F}(t)}{\mathcal{F}_0} = -\frac{A_{\rm cov}(t)}{\pi R_*^2}\ , \tag{66}$$

where we have still neglected stellar limb darkening. After the initial steep decline during ingress the brightness remains constant while the whole disk of the planet transits in front of the star (provided that $a\cos i \leq R_* - R_p$); Equation (60) applies to this phase. Its duration $t_{\rm flat}$ is given by

$$t_{\rm flat} = \frac{P}{\pi}\arcsin\left(\frac{\sqrt{(R_*-R_p)^2 - a^2\cos^2 i}}{a}\right)\ . \tag{67}$$

This expression can be derived in the same way as (58); we just have to move the planet's disk such that it touches the stellar disk from the inside. The brightness increase during egress is symmetric to the drop at the beginning of the transit.

Stellar Limb Darkening

For a more detailed quantitative analysis of planetary transit light curves it is important to consider the effects of stellar limb darkening, i.e., the variation of the brightness from the center of the disk to the edge. Limb darkening is due to the fact that the light we receive comes from optical depths $\tau \lesssim 1$

in the stellar photosphere. At the center of the disk, the line of sight penetrates vertically into the atmosphere; close to the limb it enters at an oblique angle and therefore reaches a given value of τ at a larger height. The light from the disk center therefore comes from deeper and – because of the temperature gradient in the photosphere – hotter layers. We thus expect that the brightness decreases from the center to the limb, and that this decrease depends on the atmospheric structure and observing wavelength, with a tendency to be stronger at shorter wavelengths. A common parameterization of limb darkening is

$$\mathcal{B}_\lambda(\mu) = \mathcal{B}_\lambda(0) \cdot \left[1 - c_1(\lambda)(1-\mu) - c_2(\lambda)(1-\mu)^2\right] \ , \tag{68}$$

where $\mathcal{B}$ is the surface brightness, and μ the cosine of the angle of incidence of the line of sight on the local stellar surface. μ is related to the separation r_* from the center of the disk by

$$\mu \equiv \sqrt{1 - \left(\frac{r_*}{R_*}\right)^2} \ . \tag{69}$$

Depending on the accuracy needed for a particular application, (68) is frequently used without the quadratic term, or generalized to higher polynomial orders; other functional forms have also been suggested for the description of limb darkening (Hestroffer 1997). The numerical coefficients in the limb darkening laws ($c_{1,2}(\lambda)$ or equivalent) can be computed from stellar model atmospheres (e.g., Claret et al. 1995; Díaz-Cordovés et al. 1995; Van Hamme 1993) or measured with stellar long-baseline interferometry (Quirrenbach et al. 1996). For observations of planet transits, the type of the parent star is usually fairly well known, so that tabulated limb darkening coefficients can be used to predict $\mathcal{B}_\lambda(\mu)$ quite accurately. Equation (64) can then be modified to read

$$\Delta\mathcal{F}_\lambda = 2 \int_{\max(0,d(t)-R_p)}^{\min(R_*,d(t)+R_p)} r_* dr_* \mathcal{B}_\lambda(r_*) \arccos\left[\Theta(t)\right] . \tag{70}$$

For quantitative predictions of transit light curves, this integral has to be calculated numerically. We can, however, immediately draw a few important qualitative conclusions. First, the central depth of the light curve depends on $\cos i$, even if the planetary disk fully eclipses the star. If $\cos i = 0$, the planet blocks the bright central part of the star in mid-eclipse, which means that the transit is deeper than expected from a uniform disk model. Conversely, if the planet transits at a high stellar latitude, the transit is shallower. Second, the light curves do not have a flat bottom but look more rounded than for a uniform stellar disk. Third, because of the wavelength-dependence of limb darkening, the light curves are not achromatic, but show distinct color variations. It should be emphasized that these effects are by no means negligible; they can readily be observed with high-precision photometry in transits of giant planets. An illustrative example is shown in Fig. 37.

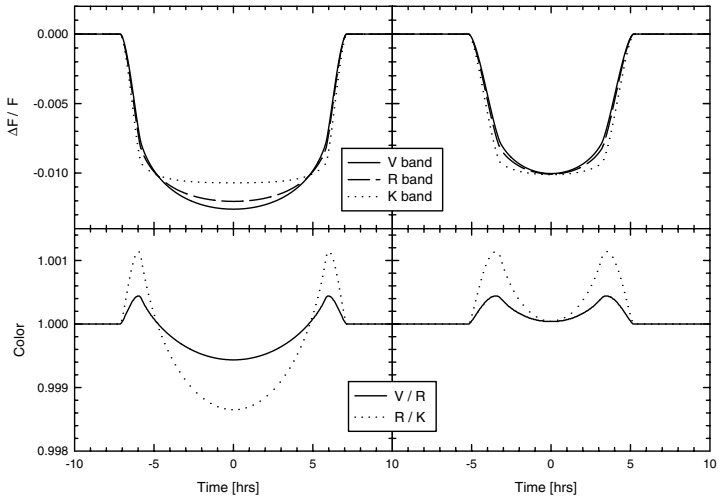

Fig. 37. Light curves (*top*) and color variation (*bottom*) for the transit of a Jupiter-like planet in a 1 AU orbit around a Solar-type star. The effect of limb darkening is more pronounced at shorter wavelengths. The transit is longer and deeper for $i = 90^\circ$ (*left*) than for $i = 89.8^\circ$ (*right*). The color curves are defined as $(\mathcal{F}_V/\mathcal{F}_R)/(\mathcal{F}_V/\mathcal{F}_R)_0$ and $(\mathcal{F}_R/\mathcal{F}_K)/(\mathcal{F}_R/\mathcal{F}_K)_0$, respectively; the subscript 0 denotes the color outside the transit. For $i = 90^\circ$ the color changes from blue at the beginning of the eclipse to red at mid-eclipse, and back to blue. For $i = 89.8^\circ$ the color is always blue. Note that the color variations have only about 10% of the amplitude of the photometric variations. The light curves were computed using the limb darkening coefficients for $T_{\rm eff} = 5750\,{\rm K}$, $\log g = 4.5$ from Claret et al. (1995) and Díaz-Cordovés et al. (1995)

Parameter Determination from Transits

Aside from the period, which can easily be measured from observations of multiple transits, the radius R_p is the most accessible parameter from transit data. If the stellar radius is known, R_p follows immediately from (60). There is an important pitfall, however: eclipsing binary stars can mimic transits by planets in two ways. First, grazing eclipses, in which the projected distance d always remains substantially larger than $R_1 - R_2$, can be very shallow because only a small fraction of the secondary ever covers the primary. This source of contamination of planet searches can normally be eliminated by photometric observations with high signal-to-noise, because the transit duration and shape do not match those expected for planetary transits. The second problem are eclipsing binaries in triple systems, or blended with unrelated stars in dense fields. The dilution of the transit depth by the additional light will lead to an underestimate of R_p, and in extreme cases to mistaking a stellar secondary for a planet. This problem is very hard to diagnose with photometry alone; spectroscopic follow-ups of candidates from photometric searches are therefore required to establish that they are indeed of planetary nature.

A mass measurement with the radial-velocity technique is extremely desirable in any case, because the size does not change very much over the mass range $1\,M_{\rm jup} \lesssim m \lesssim 100\,M_{\rm jup}$.

It has been pointed out that Kepler's Law and (58), (60), and (67) are four equations that relate the four observables P, $t_{\rm trans}$, $t_{\rm flat}$, and $\Delta\mathcal{F}$ to the four quantities R_p/R_*, a/R_*, $a\cos i/R_*$, and the density of the star ρ_*, which can therefore be determined directly from high-SNR photometric data (Seager and Mallén-Ornelas 2002). If in addition a stellar mass-radius relation is assumed, one can solve for M_*, R_*, a, i, and R_p. This analysis neglects limb darkening and thus does not provide the best possible estimates of these parameters, but it can be useful to pre-select candidate planetary transits from photometric monitoring campaigns for telescope time-consuming radial-velocity follow-up.

Planetary Radii and Transmission Spectra

Planetary transits offer a unique opportunity to obtain information on the planet's atmosphere through transmission spectroscopy, i.e., by measuring the radius of the planet as a function of wavelength. The observable quantity is $\mathcal{R}(\lambda)$, the ratio of the flux during the transit to that outside of transits:

$$\mathcal{R}(\lambda) \equiv \frac{\mathcal{F}_{\rm trans}(\lambda)}{\mathcal{F}_0(\lambda)}\,. \tag{71}$$

The integrated light of a star–planet system consists of three separate contributions: (1) the light from the star that reaches the observer directly, (2) starlight that is reflected by the illuminated part of the planetary disk, and (3) thermal emission from the planet. Separating these three components, (71) can be written as

$$\mathcal{R} = \frac{\mathcal{F}_0 + \delta\mathcal{F}}{\mathcal{F}_0} = 1 + \frac{\delta\mathcal{F}_{\rm direct} + \mathcal{F}_{\rm therm} + \mathcal{F}_{\rm refl}}{\mathcal{F}_0}\,, \tag{72}$$

where it is implicitly understood that all quantities depend on λ. Quantitative estimates of the relative importance of these three contributions show that for observations of transits of "hot Jupiters" in the visible wavelength range the second and third term can be neglected; thermal emission from the planet has to be taken into account only at $\lambda \gtrsim 2.5\,\mu$m, and the variation of the reflected light is small because only a small illuminated crescent is visible around the time of transit (Brown 2001). Several separate effects contribute to $\delta\mathcal{F}_{\rm direct}$, however, which have to be treated correctly. First of all, the shape of stellar lines changes during the transit, as the planet blocks light from different parts of the stellar disk. This effect will be discussed below (Sect. 6.3). Intrinsic variations of the flux on the time scale of a few hours are only of order 10^{-5} for the Sun (see Sect. 6.2), but larger effects occur in younger and magnetically more active stars. It is possible to discriminate against them, however, because they don't repeat consistently from one transit to the next, and because they show a wavelength dependence only in the vicinity of strong stellar lines.

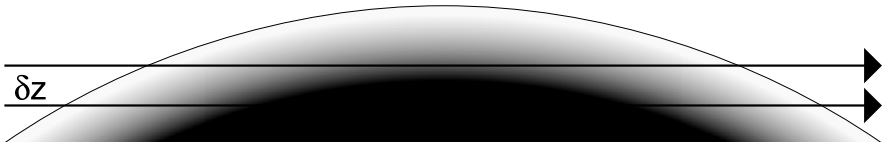

Fig. 38. Two rays separated by δz passing tangentially through a planetary atmosphere with scale height H. The opacity along the higher ray is approximately $\exp(-\delta z/H)$ smaller than along the lower ray

In regions of the spectrum away from prominent stellar lines, $\mathcal{R}(\lambda)$ is therefore affected mostly by $(\delta\mathcal{F}/\mathcal{F})_{\rm atmos}$, the part of the obscuration due to rays that pass through the atmosphere of the planet. The characteristic fractional coverage of the atmosphere projected against the stellar disk is given by the area of an annulus one atmospheric scale height H thick around the planet, divided by the area of the stellar disk. We therefore have

$$\left(\frac{\delta A}{A}\right)_{\rm atmos} = \frac{2\pi R_p H}{\pi R_*^2} = \frac{2R_p(kT/g\mu)}{R_*^2} \ , \tag{73}$$

where T and g are the temperature and surface gravity of the planet, and μ the mean molecular weight of the atmospheric constituents. For an atmosphere of H_2 with $g = 10^3\,{\rm cm\,s^{-2}}$, $T = 1,400\,{\rm K}$, $R_p = 1.4\,R_{\rm jup}$, and $R_* = R_\odot$ the numerical value is $\delta A/A = 2.4 \cdot 10^{-4}$.

If we assume that the most important opacity sources are well-mixed in the atmosphere, we can now estimate the variation of $\delta\mathcal{F}/\mathcal{F}$ with wavelength. If σ_1 and σ_2 are the opacities per gram of material at λ_1 and λ_2, the optical depth at λ_1 along a ray 1 will be approximately equal to the optical depth at λ_2 along a ray 2 if these rays are separated by $\delta z = H\ln(\sigma_1/\sigma_2)$ (see Fig. 38). The difference between the occulted flux at the two wavelengths can therefore be written as (Brown 2001)

$$\left(\frac{\delta\mathcal{F}}{\mathcal{F}}\right)_1 - \left(\frac{\delta\mathcal{F}}{\mathcal{F}}\right)_2 \approx \ln\left(\frac{\sigma_1}{\sigma_2}\right) \times \left(\frac{\delta A}{A}\right)_{\rm atmos} . \tag{74}$$

The opacity in strong molecular or atomic absorption lines may be $\sim 10^4$ times more than in the nearby continuum, which means that the two rays in Fig. 38 have to be separated by almost ten scale heights to make the line optical depth along the upper ray equal to the continuum optical depth along the lower ray. According to (73) and (74) this results in an observed line depth of $\approx 2 \cdot 10^{-3}$ with respect to the stellar flux.

Oblateness, Rings, Moons, and Starspots

With very precise photometry, it should be possible to search for deviations from the expected shape of the transit light curve given by (70). The giant

planets in the Solar System are significantly non-spherical because of their fast rotation rates, and their rotation axes are strongly inclined with respect to the normal of their orbital planes. The ingress and egress in light curves of transits by such a planet are asymmetric, unless $i = 90°$ exactly (Hui and Seager 2002; Seager and Hui 2002). The expected deviations from the light curve of an eclipse by a spherical planet is a few times 10^{-5}, which may be detectable with photometric space missions (see Sect. 6.5).

The transits of giant extrasolar planets also offer a chance to look for rings and moons around them. An opaque ring would potentially have a large cross-section, but its projected area is strongly reduced if it lies in the orbital plane of the planet. Giant moons would also produce characteristic dips or discontinuities in transit light curves, depending on their orbital parameters (Sartoretti and Schneider 1999). The space missions designed to detect transits by Earth-like planets will by definition also be sensitive to moons with radius $R_s \approx 1R_\oplus$. An alternative way of looking for moons of transiting planets is timing of the eclipses. A satellite of mass m_s in an orbit with radius a_s around a planet with mass m_p and orbital radius and period a_p and T_p will give rise to shifts in the occultation times of order

$$\tau \approx \frac{a_s m_s}{m_p} \frac{T_p}{2\pi a_p} \,. \tag{75}$$

With sub-second timing of the transit times of a Jupiter-like planet, the COROT satellite should be capable of detecting satellites similar to the Galilean moons (Sartoretti and Schneider 1999).

An interesting question thus concerns the dynamical stability of moons around close-in planets. It is well-known from the Earth–Moon system that tidal interaction causes a satellite to move inward or outward, depending on whether its orbital period is shorter or longer than the rotation period of the planet. For hot Jupiters, the planetary rotation rate is regulated by tidal interaction with the star (not with the moon), which keeps the torque on the moon strong. This leads to a loss of moons with masses $\gtrsim 10^{-6}\, M_\oplus$ around planets with orbital radii $a \leq 0.1\,\mathrm{AU}$ (Barnes and O'Brien 2002). These considerations therefore predict that moons should not be found around those planets that are most easily detected in transit surveys.

Finally, the transit light curves might show bumps if the planet happens to move across a star spot. This effect will be difficult to detect, however, because the amplitude cannot exceed the fractional area of the stellar disk covered by the spot, and because the variations are erratic and do not repeat from one transit to the next.

6.2 Photometric Error Sources

The precision of photometric observations is ultimately limited by photon noise and (for fainter stars) by the sky background. These errors can be reduced by increasing the exposure time T, because they scale with $T^{-1/2}$.

There are a number of systematic effects, however, which frequently prevent one from reaching the theoretical limit. The most important of these error sources will be discussed in the following sections.

Stellar Noise

Looking at our own Sun, we can identify many distinct mechanisms that cause variations of the emitted flux: oscillations, sunspots, flares, prominences, and variability of the granulation. Stellar activity varies strongly with spectral type and age; for example, many M dwarfs display flares with amplitudes up to one magnitude or even more (Allard et al. 1997 and references therein). Many types of binaries also show periodic brightness variations, either due to eclipses or to distortions of the stellar shape. Transit searches that monitor many thousands of stars are therefore very good at detecting variable stars (e.g., Street et al. 2002), but one has to be able to distinguish these from the planetary transits one is looking for. Fortunately, transits have a short duration and a very characteristic shape, so that it is possible to search for these events in light curves that are otherwise flat within the noise. The Sun shows variations up to $\sim$0.15% on time scales close to the rotation period due to spots, but there is very little power on time scales shorter than a day, which are typical for planetary transits (Borucki and Summers 1984 and references therein). Therefore, at least around Solar-type stars, it should be possible to clearly discriminate transits events from stellar variability, even for planets as small as $\sim 1\,R_\oplus$.

Atmospheric Noise

The Earth's atmosphere limits the precision that can be reached in ground-based photometric observations. The most important effects to consider are scintillation, changes of the extinction with time and zenith angle, and seeing variations. Scintillation is strongest for small apertures and short integration times (see Sect. 7.8). Quantitative estimates (e.g., Dravins et al. 1998) scaled to the parameters relevant for planet transit searches indicate that scintillation noise is normally not a limiting factor. Variations of the extinction due to changes in the air mass during the observations, or to non-photometric conditions, are more difficult to deal with. CCD photometry generally offers substantial advantages over classical one-channel or two-channel photometers. If the CCD field is sufficiently large, many stars can be measured simultaneously. It is then possible to perform differential photometry by dividing the observed brightness of each star by that of a set of calibration stars or by the median of all stars in the field. This eliminates extinction fluctuations and variations due to changes in the air mass to first order. If there are enough stars in the field, it it possible to perform an even better photometric calibration by taking color terms into account. Because of the limited dynamic range of the CCD, the useful range between the brightest and faintest stars

spans not more than ~ 4 magnitudes. Somewhat ironically, CCD photometry is therefore more difficult for very bright stars, since for these the contrast to the potential comparison stars within the field-of-view is usually much too large.

Changes in the width and shape of the point spread function due to short-term or night-to-night variations of the seeing (see Sect. 7.4) can be a serious error source, in particular in crowded regions of the sky. Images of stars than can be cleanly separated in a good night may be blended together when the seeing is bad, which can introduce severe photometric errors. Three main photometric methods are currently in use: aperture photometry, which measures the flux within a circle of specified radius at the location of each star; psf-fitting photometry, which fits a model of the point spread function with variable intensity at the position of each star (e.g. DAOPHOT, Stetson 1987); and image subtraction techniques, in which a reference image of the field is subtracted from each frame before the photometry is carried out (Alard and Lupton 1998). In each technique it is possible to take psf variations into account. Image subtraction algorithms generally appear to give better results than psf-fitting (Mochejska et al. 2002). If the stars are well separated from each other, aperture photometry gives excellent results; a precision of ~ 0.2 mmag for light curves of bright stars binned and averaged over 4.5 h has been demonstrated with this technique (Everett and Howell 2001).

It was suggested early on that it might be easier to look for the characteristic color changes (see Fig. 37) than for the absolute brightness changes; it was assumed that the most important errors would cancel in a differential measurement between two filters (Rosenblatt 1971). This is not true, however, because scintillation and variable extinction are intrinsically chromatic, which makes it quite difficult to detect the color changes, which are approximately ten times smaller than the photometric variations.

Instrumental Noise

While modern CCD cameras do not contribute significantly to the noise of ground-based photometric measurements, instrumental effects have to be taken into account at the $\sim 10^{-5}$ precision level that can be achieved in space. The most troublesome difficulty is the effect of intra-pixel variations of the quantum efficiency; if a point source of constant intensity is scanned across a pixel, variations in the electron count by several percent may be observed. The best strategy to mitigate this problem is avoiding steep gradients in the focal plane illumination by defocusing the telescope, so that the light from each star is sampled by $\sim 5 \times 5$ pixels. Laboratory tests have shown that a photometric precision of 10^{-5} can indeed be achieved with a careful calibration of systematic effects (Robinson et al. 1995). A difficulty of the defocusing approach is the increased sensitivity to contamination by faint background stars. There is a requirement on the stability of the telescope pointing to avoid background stars moving onto and off the pixels used for each target star. In addition,

faint eclipsing binaries near the target star may mimic planetary transits; these false detections have to be eliminated by follow-up observations (see Sect. 6.4).

6.3 HD 209458

Light Curve and System Parameters

Photometric follow-up observations of stars with known radial-velocity variations lead to the discovery of the first transiting planet, HD 209458 B (Henry et al. 2000; Charbonneau et al. 2000). This detection of dips in the light curve with the "right" shape and at the "right" times provides an important confirmation of the existence of extrasolar planets, dispelling the last doubts about the interpretation of the radial-velocity variations. In addition, the transit light curve provides for the first time access to physical parameters of a planet other than $m \sin i$, as explained in Sect. 6.1. Observations of HD 209458 with the STIS spectrograph on the Hubble Space Telescope have produced an exquisite transit light curve, with typical photometric errors of $\sim 1.1 \cdot 10^{-4}$ for each 60 s data point (Brown et al. 2001, see Fig. 39). This corresponds to a signal-to-noise of ∼150 for the transit depth of 1.64%.

The mass of HD 209458 has been estimated to be $m_* = 1.1 \pm 0.1\, M_\odot$ from its metallicity and location in the HR diagram, using stellar evolutionary models (Mazeh et al. 2000). With this mass as input value, the transit light curve can be used to determine the radii R_* and R_p, and the orbital inclination

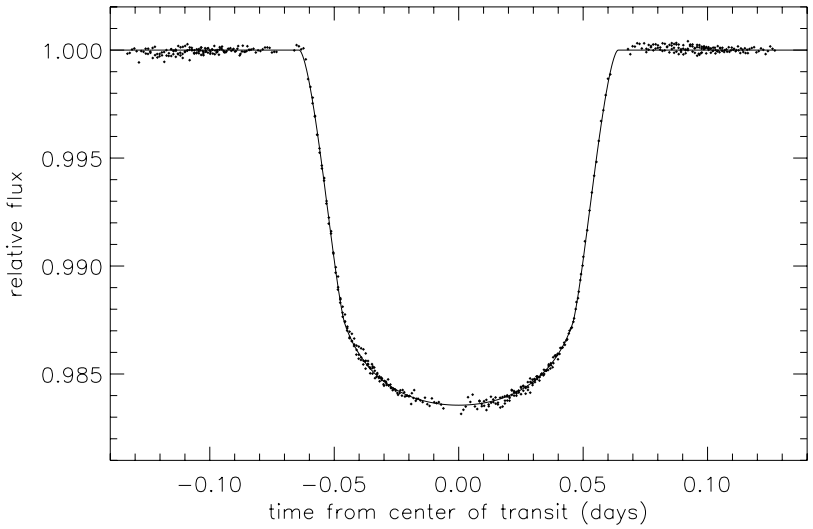

Fig. 39. Phased light curve of the transits of HD 209458 B from observations with the STIS spectrograph on the Hubble Space Telescope. From Brown et al. (2001)

Table 10. Parameters of HD 209458 and its planet, compiled from Henry et al. (2000); Mazeh et al. (2000); Queloz et al. (2000a); Robichon and Arenou (2000) and Brown et al. (2001)

parameter	value
m_*	$1.1 \pm 0.1\, M_\odot$
R_*	$1.146 \pm 0.050\, R_\odot$
$v \sin i_*$	3.7 ± 0.5
P	3.524739 ± 0.000014
a	0.046 ± 0.001 AU
$i_{\rm orb}$	$86\overset{\circ}{.}68 \pm 0\overset{\circ}{.}14$
m_p	$0.69 \pm 0.05\, M_{\rm jup}$
R_p	$1.347 \pm 0.060\, R_{\rm jup}$
ψ	$0 \pm 30^\circ$

The table lists mass, radius, and projected rotation velocity of the star, period, radius, and inclination of the orbit, mass and radius of the planet, and the angle between the equatorial plane of the star and the planetary orbit

$i_{\rm orb}$. The orbital period P, orbital radius a, and planetary mass m_p are known from the radial-velocity observations. Values for these parameters are listed in Table 10, together with the stellar rotation velocity $v \sin i_*$, and the angle ψ between the equatorial plane of the star and the planet's orbit (from an analysis of spectroscopic data during transit, see below).

The Radius of HD 209458 b

Knowledge of both the mass and radius of HD 209458 B is a key for comparisons with physical models of planets. A first important conclusion is that it is made predominantly of hydrogen; a rocky or icy planet with a mass of $m_p = 0.69\, M_{\rm jup}$ would have a radius smaller than HD 209458 B by a factor of 3 to 4 (Burrows et al. 2000). A more detailed analysis shows that the radius of HD 209458 B is also larger than that expected for an isolated $0.69\, M_{\rm jup}$ planet; this is a consequence of the retardation of contraction by the stellar irradiation. An interesting conclusion is that HD 209458 B must have migrated to its present position early on (or even been born there); if it had dwelled more than $\sim 10^7$ years at a distance $\gtrsim 0.5$ AU, it would already have contracted during that time to a radius smaller than the presently observed value (Burrows et al. 2000). The exact radius–age relation for a given mass and external irradiation depends sensitively on the Bond albedo (for definition see Sect. 6.6);

uncertainties in this quantity therefore limit our current ability to compare the observational data to detailed model calculations. The most recent models that take into account the irradiation of HD 209458 B by its parent star predict a radius ∼20% lower than the observed value; an additional source of energy might therefore be required (Baraffe et al. 2003).

The effective planetary radius is a function of wavelength due to variations of the atmospheric opacity (see Sect. 6.1). For the parameters of HD 209458 B, dramatic variations in the occulted area are expected due to alkali metal lines in the visible (Seager and Sasselov 2000), and to H_2O absorption in the near-IR (Hubbard et al. 2001). The HST STIS data mentioned above show indeed that the transit of HD 209458 is deeper at the wavelength of the 589 nm Na resonance doublet than in the adjacent continuum; the difference is $(2.32 \pm 0.57) \cdot 10^{-4}$ with respect to the stellar flux (Charbonneau et al. 2002). These measurements constitute the first detection of an extrasolar planet atmosphere, and confirm the important role of alkali metal absorption for the spectra of "hot Jupiters". A detection of the first overtone band of CO near 2.3 μm may also be possible with the same technique (Brown et al. 2002). A quantitative interpretation of the observed wavelength dependence of the planetary diameter is only possible with the inclusion of non-LTE effects in the modeling of the upper atmosphere (Barman et al. 2002).

It has been suggested that even planets that do not show photometric eclipses could exhibit "transits" of absorption features. Because of the strong stellar irradiation and interaction with the stellar wind, the mass loss from planets like 51 Peg b may be appreciable and lead to and "exosphere" similar to the tails of comets. Because of its larger size, this exosphere would periodically eclipse the star for a larger range of inclinations than the planet itself. A search toward 51 Peg with the Short Wavelength Spectrometer of the ISO satellite did not result in the detection of any absorption lines from atoms, molecules or their ionization or dissociation products from a transiting cloud (Rauer et al. 2000). Similar observations toward 51 Peg and HD 209458 in the visible have not lead to any detections, either (Bundy and Marcy 2000; Moutou et al. 2001). The existence of extended exospheres around strongly irradiated planets thus remains speculative.

Rossiter–McLaughlin Effect

Planetary eclipses affect not only the observed brightness of the star, but also the shape of spectral lines. A prograde planet orbiting in the equatorial plane transits in front of the approaching side of the star during the first half of the transit, and in front of the receding side during the second half. This causes successive dips in the blue-shifted and red-shifted wings of the spectral lines analogous to those due to star spots. If the apparent "radial velocity" of the star is determined from the centroids of the lines, the selective occultation causes an anomaly, whose shape and amplitude depend on the stellar rotation velocity and the geometry of the transit. This effect has been observed in

HD 209458, and used to place an upper limit of 30° on the angle between the orbital plane and the stellar equatorial plane (Queloz et al. 2000a). In addition to providing information on the transit geometry, observations of the expected spectroscopic anomalies could also help to confirm future detections of transiting planets from photometric surveys.

6.4 Photometric Planet Searches

Requirements of Wide-Field Searches

Whereas the detection of transits in the HD 209458 system resulted from follow-up observations of a known planet discovered by radial-velocity surveys, a number of projects are also underway to conduct searches for transiting planets through photometric monitoring of large numbers of stars. The immediate goal of these surveys is the detection of transiting "hot Jupiters". The probability that a planet in a 0.05 AU orbit will show transits is ∼10% (57); if a few percent of all Sun-like stars have such planets, one should therefore expect a few detections per 1,000 stars monitored. The expected transit depths (∼1%, (60)), durations (∼3 h, (59)), and repeat rates (once every ∼4 days) are all favorable for ground-based surveys.

A wide-field search for transiting planets has to accomplish two separate tasks (1) identification of transit candidates, i.e., stars with regularly spaced brightness dips that are consistent with planetary transits and (2) confirmation that the features in the light curve are indeed due to a transiting planet. To identify transit candidates, it is first necessary to obtain a large number of exposures of the target field, and to perform a careful photometric analysis as described in Sect. 6.2. Then one has to search for small dips in the many resulting light curves. The most straightforward approach is performing an automated search for individual dips that have a depth and duration consistent with the expectations for a planetary transit. If there are least three dips in the light curve of one star, a candidate has been found, provided also that the differences between the times at which the dips occur are small multiples of one common interval (the orbital period). For light curves with only two dips, one can predict the epochs at which further transits should occur, and thus establish more candidates with follow-up observations at these times.

If the signal-to-noise ratio in the light curves is too small for a straightforward detection of individual transits, it is still possible to search for the periodic signal of multiple transits. One possibility is folding each light curve with a set of trial periods, and cross-correlating the folded time series with a template transit light curve. If the cross-correlation coefficient for a star–period pair exceeds a certain pre-defined threshold, one has found a transit candidate. This method has the disadvantage that the sensitivity of the search is reduced for transits with a duration that is significantly different from the one assumed in the template. This difficulty can be avoided by using a "box-fitting" algorithm (Kovács et al. 2002). This algorithm searches for signals

that alternate periodically between two discrete levels, with much less time spent at the lower level. This is a fair description of planetary transits for the purposes of a search algorithm.

Once a sample of transit candidates has been established, careful follow-up observations have to be performed to discriminate true planets from false alarms. The most important source of contamination are stellar eclipsing binaries, which can mimic planetary transits in a number of ways:

- Eclipses at grazing incidence. If the impact parameter $d_{\min}$ is in the range $R_1 - R_2 < d_{\min} < R_1 + R_2$, the secondary (radius R_2) will never be fully in front of the primary (radius R_1). This produces a shallow eclipse, which could be mistaken for the transit of a smaller object.
- Eclipses of an evolved primary by a main-sequence secondary. The diameter ratio between an evolved star and a low-mass main-sequence star can be similar to that between a main-sequence star and a giant planet; this leads to eclipses of identical depth and similar shape for these two types of systems.
- Eclipses in triple systems. The light curve of a hierarchical triple system, in which the primary is a Solar-type star, and the secondary an eclipsing pair of late-type dwarfs can be very similar to that of a Solar-type star orbited by a giant planet.

In principle, these can all be recognized as false alarms from the light curves alone, since the details of the transit (total duration, duration of ingress and egress, limb darkening profile and color variation) differ from those expected for planets. While this may be a possibility for very accurate photometry from space missions, observations from the ground will normally not reach the required precision. It is therefore necessary to follow-up the candidates with other techniques, which in many cases turns out to be much more difficult than finding the candidates in the first place. Medium-resolution spectroscopy can be used to determine the spectral type and luminosity class of the primary for each candidate event, and thus to weed out eclipses of evolved stars by main-sequence secondaries. Radial-velocity monitoring with relatively low precision (a few times $100\,\mathrm{m\,s^{-1}}$) can rule out eclipses by stellar secondaries at grazing incidence; a null result at this level means that there cannot be a stellar-mass secondary in an edge-on orbit of a few AU or less. For a clear confirmation of the planetary nature of the secondary, a positive detection of the radial velocity variation of the primary is needed. However, this normally requires radial velocity measurements with a precision of a few times $10\,\mathrm{m\,s^{-1}}$. This is quite difficult in the case of faint primaries, but necessary to reject cases like eclipses in triple systems.

Planet candidates from photometric transit data without mass determinations from radial-velocity (or astrometric) measurements are also of very limited use because the radius carries very little information on the physical nature of the transiting object (see Fig. 40). For example, a radius of $0.15\,R_\odot$ can correspond to a low-mass star of $\sim 0.1\,M_\odot$, to a brown dwarf, or to a

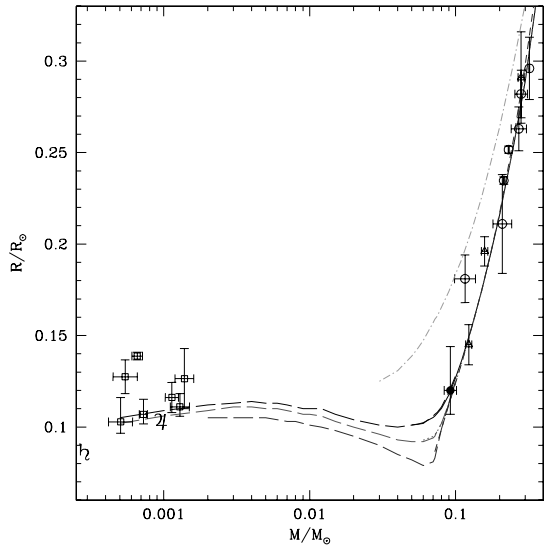

Fig. 40. Mass vs. radius for observed low-mass stars and giant planets and theoretical isochrones. Eclipsing binaries are shown as *circles* (OGLE-TR-122b in black), interferometric data as *open triangles*. **?** isochrones for masses from 0.06 to 1.4 $M_{\odot}$ are plotted for 5 Gyr (*solid*) and 0.1 Gyr (*dash-dotted*). *Dashed lines* represent the Baraffe et al. (2003) CON models for masses for 0.5, 1 and 5 Gyr, from top to bottom (Figure added during proof by Editor)

planet with a mass as low as $2\,M_{\rm jup}$, depending on age. If one assumes that the companion is several billion years old, the range of possible masses for a given radius is substantially smaller. Because of the shallowness of the mass–radius relation, however, observational errors and uncertainties in the radius of the primary still lead to a very large allowed range for the companion mass. Follow-up observations are therefore required for a proper interpretation of planet candidates discovered by photometric surveys.

Searches in the Field and Toward the Galactic Bulge

The best method to look for transiting planets around nearby stars is monitoring with a small wide-angle telescope, which can produce simultaneous light curves of thousands of stars. The Vulcan project at the Lick Observatory on Mt. Hamilton uses a lens with 12 cm aperture to observe 6,000 stars brighter than 13th magnitude in a $7^{\circ} \times 7^{\circ}$ field with a 4096×4096 pixel CCD (Borucki et al. 2001). A similar setup is used by the STARE group (Brown and Charbonneau 2000). The detection efficiency of these surveys is somewhat limited because of the unavoidable gaps in the light curves during daytime and periods of bad weather. It would therefore be highly desirable to establish a network of small telescopes with good longitude distribution, which could

perform continuous monitoring of many square degrees over several months. The planets detected by this network would be relatively nearby ($\lesssim 300\,\mathrm{pc}$) and could therefore be studied in detail with follow-up spectroscopy.

An alternative approach to simultaneous monitoring of many stars is using wide-field cameras on large telescopes (e.g., Mallen-Ornélas et al. 2003). Modern all-purpose cameras with CCD mosaics typically cover of order $30' \times 30'$, and exposure times of a few minutes at 4 m class telescopes are sufficient for $\lesssim 1\%$ photometry on stars between 16th and 18th magnitude. In the Galactic plane, several 10,000 stars can be monitored in this way. The Galactic bulge possesses an even higher stellar density, but it is necessary to look at even fainter stars ($19 \leq m_I \leq 21$), because the brighter bulge stars have already evolved off the main sequence and therefore have diameters that are too large for planetary transit detection. This means that a $5' \times 5'$ field-of-view is sufficient, but a $\sim$10 m aperture is required for an efficient search for planets transiting bulge stars (Gaudi 2000). However, the microlensing experiments surveying large fields toward the bulge also observe many thousands of disk main sequence stars located in the same sky area. Transits of disk stars with depths of a few percent, corresponding to estimated companion radii down to $\sim 1.5\,R_{\mathrm{jup}}$ have indeed been identified in the OGLE microlensing data set (Udalski et al. 2002).

A substantial problem of deep surveys is the difficulty of confirming planet candidates and measuring their masses, due to the faintness of the primaries and the potentially very high false alarm rate due to blending and grazing eclipses of stellar binaries. Spectroscopic observations of a number of the OGLE candidates from Udalski et al. (2002) have been used to *reject* most of them as low-mass stellar companions (Dreizler et al. 2002), but two candidates still remained in this study. It has recently been claimed that one of them (OGLE-TR-3) and another OGLE candidate (OGLE-TR-56) are indeed transiting extrasolar planets (Dreizler et al. 2003; Konacki et al. 2003), but these "planet detections" appear to be highly dubious[23]. In neither case are the radial-velocity data good enough for a secure detection of the claimed variations; this leaves many alternative interpretations of the eclipses open. More importantly, both papers claim a detection in a part of parameter space where planets are known to be exceedingly rare: the OGLE candidates have periods of $P = 1.1899\,\mathrm{d}$ and $P = 1.2119\,\mathrm{d}$, respectively, whereas the shortest-period planet known from radial-velocity surveys has $P = 2.5486\,\mathrm{d}$ (Udry et al. 2003a). Since the radial-velocity technique is very sensitive to short-period planets, the complete absence of planet-mass objects with periods close to those of OGLE-TR-3 and OGLE-TR-56 from the radial-velocity samples means that they must be very rare. The probability that the first *two* objects

[23] Editor note added in proof: Several planets on short orbits have been since detected in OGLE fields after submission of the manuscript. OGLE-TR-3 has been shown to be a grazing binary system

found with transit searches should reside in a sparsely populated region of parameter space is exceedingly small.

Searches in Open Clusters

Old open clusters are interesting targets for transit searches, since they contain numerous main-sequence stars with uniform distance, age and metallicity (Janes 1996; Quirrenbach et al. 2000). The clusters NGC 2682, NGC 6819, and NGC 7789 were selected for a pilot study with the 1 m Nickel Telescope at Lick observatory and the 1 m Jacobus Kapteyn Telescope at La Palma. The relative small fields of these telescopes (about $6' \times 6'$) allowed monitoring of only a few hundred stars, but the possibility of performing relative photometry at the $\lesssim 0.3\%$ level in the relatively crowded cluster fields was successfully demonstrated (Quirrenbach et al. 2000). This project has therefore been continued at the 2.4 m Isaac Newton Telescope, also at La Palma (Street et al. 2000). This survey, and a similar search in NGC 6791, have detected large numbers of variable stars in the cluster fields (Street et al. 2002; Mochejska et al. 2002). The data reduction has been refined to the point where an accuracy at the theoretical limit is reached for the vast majority of all stars. For more than 10,000 stars in the field of NGC 6819, the precision is sufficient to detect transits down to $1\,R_{\rm jup}$.

Under the assumption that $\sim$1% of all stars have "hot Jupiter" companions, and that $\sim$10% of these produce transits, several transits should have been detected by the observations of NGC 6819. One apparent transiting object with a radius similar to HD 209458 B has indeed been identified (Street et al. 2003). Improvements of the transit detection algorithm will be needed to identify more candidates, to measure the frequency of hot Jupiters in the cluster, and compare it to that in the Solar neighborhood. Measurements in a larger sample of open clusters could help to explore the relation of planet formation to metallicity and other environmental factors.

The Globular Cluster 47 Tucanae

The WFPC2 instrument on the Hubble Space Telescope was used in July 1999 to monitor a field in the globular cluster 47 Tuc continuously for 8.3 days (Gilliland et al. 2000). The core of the cluster was placed on the PC chip of WFPC2; a total of $\sim 34,000$ main-sequence stars with a typical projected distance from the core of $1'$ could thus be monitored. The noise in the light curves of stars in the magnitude range $17.1 \leq m_V \leq 21.5$ ranged from $\sim$0.3% to $\sim$ 3%. Assuming a frequency of "hot Jupiters" identical to that in the Solar neighborhood, some 15 . . . 20 transits should have been detected, but none were actually found. It has thus been established with extremely high significance that hot Jupiters are much less prevalent in 47 Tuc than in the Solar neighborhood.

Two possible reasons for this difference come to mind immediately. The lack of massive close-in planets in 47 Tuc could either be a metallicity effect, or related to the high density of stars in the cluster. 47 Tuc is a massive cluster with a stellar density of $n \sim 10^5\,\mathrm{pc}^{-3}$ in the core, and $n \sim 10^4\,\mathrm{pc}^{-3}$ (corresponding to $\sim 10^3\,M_\odot\,\mathrm{pc}^{-3}$) at $1'$ from the core. The metallicity of 47 Tuc is $[\mathrm{Fe/H}] = -0.7\,\mathrm{dex}$, but the abundance of α elements with respect to Fe is $[\alpha/\mathrm{Fe}] = 0.4\,\mathrm{dex}$ (Salaris and Weiss 1998). Extrapolating the metallicity dependence of the incidence of planets (Sect. 3.4) to even lower values, it certainly appears plausible that "hot Jupiters" are exceedingly rare around stars as metal-poor as those in 47 Tuc. On the other hand, disruption of planet formation by close stellar encounters should also play an important role. It is likely that close-in planets with $a \lesssim 0.3\,\mathrm{AU}$ survive for a Hubble time at a density $n \sim 10^4\,\mathrm{pc}^{-3}$, which corresponds to the "typical" environment of the stars monitored in the WFPC2 observations (Davies and Sigurdsson 2001). One has to consider the history, of the cluster and the planetary system, however. A cluster that starts its life with a certain density will expand when gas that was not used up by star formation is expelled by the winds of OB stars or supernova explosions (Goodwin 1997; Kroupa et al. 2001). This leads to a decrease of the cluster density by a factor of ∼10 or more. It is thus quite possible that all protoplanetary disks or young planetary systems were destroyed in an early high-density phase of 47 Tuc (Bonnell et al. 2001). It is also possible that the ionizing flux of young massive stars destroys all circumstellar disks in the cluster before they can form planets (Armitage 2000); this process seems indeed to be at work near θ^1 Ori C in the Orion Trapezium (Johnstone et al. 1998). More observations in environments that cover different combinations of density and metallicity are needed to settle the question which of these factors is responsible for the lack of close-in planets in 47 Tuc.

Eclipsing Binaries

An interesting way to increase the odds of finding transiting planets are observations of eclipsing binary systems (Schneider and Chevreton 1990). This idea is based on the two assumptions that (a) planets form around close binaries with similar properties to those around single stars and (b) the orbital plane of the planet(s) will be roughly coplanar with the binary orbit. Our current understanding of binary star formation is certainly not good enough to provide compelling arguments either in favor of or against these hypotheses, and observational data do not exist.[24] These uncertainties notwithstanding,

[24] Note that planets around members of wide multiple systems have been found by the radial-velocity method, but close binaries are not suited as targets for this technique because of the large radial-velocity amplitude of the binary orbit.

an extensive observing campaign has been carried out to search for planetary transits in the CM Draconis system (Deeg et al. 1998; Doyle et al. 2000). This is arguably the best-suited binary for such a project, because both components are M4.5 dwarfs, which gives them a combined disk area of $\sim 12\%$ of the Solar disk (see Lacy 1977 for details). This means that a transit of a planet with $3.2\,R_\oplus$ would cause a 1% photometric dip. Furthermore, the inclination is nearly edge-on, $i = 89.82°$, which implies that planets in coplanar orbits with radii up to 0.35 AU would actually cause eclipses (see (56)).

Several 1 m class telescopes have been used to obtain a long time series of photometric measurements of CM Dra, which comprises more than 25,000 data points covering over 1,000 h with an rms precision of 0.2% to 0.7% (Deeg et al. 1998; Doyle et al. 2000). A single transit from a planet significantly larger than $3\,R_\oplus$ would produce a signal that could easily be detected above the noise. Considering the time coverage of the light curve, the detection probability for such planets with orbital periods between 7 and 60 days is $> 90\%$, but no such events were found. Smaller planets could still be detected if they transit several times. The search for multiple transits from the same planet is much more complicated in this case than for a single star, because the orbital motion of the binary causes distortions of the individual dips, and a non-periodic signal from the repeated transits. Instead of doing a periodogram analysis, one therefore has to generate simulated light curves and cross-correlate them with the data, which is time-consuming and CPU intensive. Doyle et al. (2000) conclude that the detection probability for $2.5\,R_\oplus$ planets with periods up to 10 days in the CM Dra light curve from 1994 to 1998 was 80%. They found a few planet "candidates", i.e., assumed sets of planetary radius, orbital period and epoch that would produce several transits matching observed dips near the noise level. Follow-up observations around the predicted transit times of seven of the best candidates in 1999 showed that six of them were not real, however, leaving only one good candidate. Further unpublished observations of this candidate gave somewhat contradictory results: two rather convincing transit dips at the right time followed by three non-transit events that should have been there (L. Doyle, priv. comm.). It is presently unclear whether this is an inconsistency due to instrumental problems, or a detection of real transits of a non-coplanar object whose orbital nodes subsequently precessed from the line-of-sight. One can thus summarize that a fair part of the parameter space for planets $\gtrsim 2.5\,R_\oplus$ with periods between 7 and 60 days in orbits coplanar with the binary has been searched in CM Dra with negative results; no strong statement is currently possible for similar planets in non-coplanar orbits.

A second way to look for the signature of planets in the CM Dra light curve is by timing the eclipse minima (Deeg et al. 2000). This analysis is complementary to the search for transits; it is less sensitive to small planets, but it can detect planets with longer periods and in non-coplanar orbits. The

amplitude of the timing residuals τ for a planet with mass m_p orbiting a binary with component masses $m_{1,2}$ at an orbital radius a is given by[25]

$$\tau = \frac{m_p a \sin i}{(m_1 + m_2)c} \,. \tag{76}$$

The low mass of the components of CM Dra ($m_1 + m_2 = 0.44\,M_\odot$) is favorable, and the light curve obtained for the transit search covers more than 80 primary and secondary eclipses of the binary pair. A periodogram analysis of the 41 eclipses with the best-determined minimum times yields no peak above $\sim$3 s, from which the presence of large planets (e.g., $m_p \sin i \geq 1\,M_{\rm jup}$ for $a = 2\,$AU) can be ruled out according to (76). There is weak evidence for excess power at periods around 1,000 days, which would correspond to a circumbinary planet with $m_p \sin i \approx 1.5 \ldots 3\,M_{\rm jup}$ at $a \approx 1.1 \ldots 1.45\,$AU (Deeg et al. 2000). The significance of this feature in the power spectrum is at present uncertain, but it demonstrates the possibility of using transit timing in eclipsing binaries for searches of Jupiter-like planets.

6.5 Photometric Space Missions

As we have seen in Sect. 6.2, the Earth's atmosphere is the most severe source of photometric noise. Observations from orbiting observatories can thus reach much better precision, as already demonstrated by the HST light curve of HD 209458 (Fig. 39), obtained with an instrument that was not specifically built as a precise photometer. A number of photometric space missions will be launched in the near future; they will provide new opportunities for observations of extrasolar planets.

COROT, MONS, and MOST

Three small photometric satellites are expected to be launched over the next few years: the Canadian MOST spacecraft (Matthews et al. 2000), the Danish MONS (Christensen-Dalsgaard 2002), and the French/ESA project COROT (Baglin et al. 2002). These missions were initially conceived with the primary objective of studying the internal structure of stars through asteroseismology; they have therefore been designed to perform exquisite photometry of very bright stars (better than 10^{-6} for periodic signals like stellar oscillation modes). With the discovery of hot Jupiters it has become obvious that these satellites will also be able to perform precise measurements of the light curves of planetary transits, and to detect their reflected light. Observations of extrasolar planets have therefore been added to their scientific goals. Two different

[25] Note the difference between (75) and (76). In the first case, the invisible body orbits one component of the eclipsing system, and the second factor in the timing equation is the inverse of the orbital velocity of the eclipsing body. In the latter case, the invisible body orbits the eclipsing system at a large distance, and the velocity in the denominator is the speed of light.

modes are possible: a survey mode, which uses a wide-field camera to search for new transiting planets, and a targeted mode to get detailed light curves for stars that are already known to host planets. COROT, for example, will monitor about 6,000 to 12,000 stars in the range $m_V = 11 \ldots 16.5$. It should be able to detect planets with radii as small as $R_p = 1.6\,R_\oplus$, provided that their orbital periods are not longer than 50 days (Rouan et al. 2000). During the first two years of its lifetime, the planet program of MOST will consist of pointed observations of ε Eri, τ Boo, 51 Peg, HD 38529, and HD 209458. More targets may be added later.

Kepler and Eddington

ESA's Eddington (Favata 2002) and NASA's Kepler (Borucki et al. 1997; Koch et al. 1996, 1998) missions are more ambitious than the relatively small satellites described in the last section: they will search for transit of Earth-like planets and thus measure the frequency of other potentially habitable worlds. This requires both a large field-of-view and a sizeable aperture. Eddington will cover a field of at least 6 square degrees with a 1.2 m telescope. Kepler will even observe a 12° field with a 0.95 m aperture; this requires a focal plane covered with 42 CCD detectors.

If most stars have Earth-like planets, both Kepler and Eddington are expected to produce many reliable detections. In addition to detecting up to 50 "Earth twins", each of these missions would be able to measure the reflected light from hundreds of hot Jupiters, and observe transits for ∼100 of them. The success of such an ambitious program clearly depends on a reliable data processing pipeline, which has to find the transiting objects and discriminate them from low-amplitude variables. This is particularly difficult for small planets in short-period orbits; for these objects many transits will be observed, but each one will have a very low signal-to-noise. Understanding the properties of the light curves (after phasing with trial periods and co-adding of the individual transits), and the statistical significance of low-level transit detections, is therefore an important task (Deeg et al. 2000; Jenkins et al. 2002). If the Kepler and Eddington missions can be carried out successfully, they will likely be first to give us an answer to the question whether Earth analogs are common or rare around stars that are similar to our Sun.

6.6 Searches for the Light Reflected by the Planet

Definition of Albedo

For a discussion of the properties of starlight reflected by a planet, we first have to clarify the notion of albedo. The best-known concept is that of the *Bond albedo* A defined by

$$A \equiv \frac{P_{\rm refl}}{P_{\rm incid}} \ , \tag{77}$$

where $P_{\rm incid}$ is the total power of light incident on a surface, and $P_{\rm refl}$ that reflected by it. The Bond albedo governs the equilibrium temperature $T_{\rm eff}$ of a planet heated by its parent star. Balancing radiation losses with internal energy production and insolation gives

$$4\pi R_p^2 \sigma T_{\rm eff}^4 = P_{\rm int} + \pi R_p^2 (1-A) L_* / (4\pi a^2) \ , \tag{78}$$

where R_p is the radius of the planet, a its orbital radius, $P_{\rm int}$ the internal energy production per time interval, and L_* the stellar luminosity. If the internally generated heat is negligible compared to the insolation, the equilibrium temperature is given by

$$T_{\rm eff} = \left[\frac{(1-A)L_*}{16\pi\sigma a^2} \right]^{1/4} . \tag{79}$$

While the Bond albedo is needed to estimate the temperature of a planet, the quantity that is most useful to describe the reflected light is the *geometric albedo* p_λ. It is defined as the reflectivity of a planet at wavelength λ, measured at full phase, i.e., $\alpha = 0$, where α is the angle star–planet–observer. (We neglect the possibility of eclipses in this section.) The observed intensity of the reflected light $\mathcal{F}_\lambda(\alpha)$ can then be written as

$$\mathcal{F}_\lambda(\alpha) = p_\lambda \mathcal{F}_{\rm incid} \left(\frac{R_p}{d} \right)^2 \phi_\lambda(\alpha) \ , \tag{80}$$

where $\mathcal{F}_{\rm incid}$ is the stellar flux incident on the planet, and $\phi_\lambda(\alpha)$ the *phase function*, which describes the phase dependence of the scattering and is normalized such that $\phi_\lambda(0) = 1$. The flux from a star at distance d observed on the Earth is related to the flux incident on a planet at orbital distance a by the inverse square law

$$\mathcal{F}_* = \mathcal{F}_{\rm incid} \left(\frac{a}{d} \right)^2 . \tag{81}$$

Dividing (80) by (81), we thus obtain an expression for the intensity ratio $\epsilon_\lambda(\alpha)$ between the planetary and stellar spectra:

$$\epsilon_\lambda(\alpha) \equiv \frac{\mathcal{F}_\lambda(\alpha)}{\mathcal{F}_*} = p_\lambda \left(\frac{R_p}{a} \right)^2 \phi_\lambda(\alpha) \, . \tag{82}$$

For circular orbits, the phase angle α can be computed from

$$\cos\alpha = -\sin i \sin 2\pi\Phi \ , \tag{83}$$

where $\Phi \in [0,1)$ is the orbital phase, measured from the time of maximum recessional velocity of the star.[26]

[26] Note that different definitions of the zero point of the orbital phase are used in the literature. The one adopted here is customarily used for spectroscopic binaries, whereas in eclipsing binaries the time of the primary eclipse is normally chosen as the zero point of the orbital phase. In analogy to the latter definition, some authors count the phase from the time of inferior conjunction. For circular orbits, the difference between the two definitions is 0.25.

The functions p_λ and $\phi_\lambda(\alpha)$ depend on the properties of the planetary atmosphere. For reference purposes, it is useful to consider the case of a *Lambert sphere*, which scatters all incoming photons isotropically. It can be shown that for a Lambert sphere

$$p_\lambda = 2/3 \ , \tag{84}$$

and

$$\phi_\lambda(\alpha) = \frac{\sin\alpha + (\pi - \alpha)\cos\alpha}{\pi} \, . \tag{85}$$

Inserting these two relations in (82), we get

$$\epsilon_\lambda(\alpha) = \frac{2}{3}\left(\frac{R_p}{a}\right)^2 \left(\frac{\sin\alpha + (\pi - \alpha)\cos\alpha}{\pi}\right) . \tag{86}$$

This equation can be used together with (83) to calculate simple models of the phase variations. Inserting typical numbers of "hot Jupiters" in (86), we expect variations of order

$$\Delta\epsilon = 6 \cdot 10^{-5} \left(\frac{R_p}{R_{\rm jup}}\right)^2 \left(\frac{a}{0.05\,{\rm AU}}\right)^{-2} \tag{87}$$

for planets with edge-on orbits. (For $i < 90°$ the variations are smaller, because we cannot probe the full range of phase angles.) With the precision of the upcoming photometric space missions (Sect. 6.5), it should thus be possible to detect the starlight reflected by hot Jupiters through their phase variations.

For more realistic predictions of the brightness of the reflected light, one has to compute detailed models of the structure of the planetary atmosphere (see below). The presence of strong atomic and molecular absorption features also opens the possibility to obtain spectroscopic information on the planetary atmospheres. It should finally be pointed out that the atmospheric light scattering processes lead to a high degree of polarization of the reflected light at phase angles near or slightly below 90°. The expected signature of a planet in the stellar polarization fraction is only a few times 10^{-6}, however, which is below the current detection limit (Seager et al. 2000).

The Signature of Spectral Features in Reflected Light

It is obvious from (80) that the light reflected by a planet carries spectral imprints both from the stellar photosphere (wavelength dependence of $\mathcal{F}_{\rm incid}$) and from the planet's atmosphere (parameterized by p_λ and $\phi_\lambda(\alpha)$). For observations with low spectral resolution, the description in (80) is sufficient, but at a resolving power $R \gtrsim 1{,}000$ we have to take a more careful look at

the implications of the Doppler effect. First of all, the radial velocity v_p of the planet is given by

$$v_p(\Phi) = K_p \cos 2\pi\Phi = -K_* \frac{m_*}{m_p} \cos 2\pi\Phi \ , \tag{88}$$

where K_p and K_* are the radial-velocity amplitudes of the planet and the star, respectively. If the geometric albedo of the planet can be regarded as constant over a small spectral range in the vicinity of a stellar absorption line, a "ghost image" of this line appears in the reflected spectrum with (time-dependent) amplitude and Doppler velocity given by (82) and (88). The velocity amplitude K_p is typically many $\mathrm{km\,s^{-1}}$ ($K_p = 30\,\mathrm{km\,s^{-1}}$ for $m_* = 1\,M_\odot$ and $a = 1\,\mathrm{AU}$), which is much larger than the line width of old G dwarfs, and much more than the resolution achievable with an Echelle spectrograph. It thus appears promising to search for the reflected "ghost spectrum" in high-signal-to-noise spectra of stars with known planets. The difference of two such spectra taken at different orbital phases should reveal two ghost spectra (one of them positive, the other negative) at the velocities predicted by (88). One important difficulty of such searches is the unknown orbital inclination of the planet. Combining (19) with (88), we obtain

$$v_p(\Phi) = \left(\frac{2\pi G}{P}\right)^{1/3} m_*^{1/3} \sin i \, \cos 2\pi\Phi. \tag{89}$$

This means that only an upper limit to K_p is known a priori. The need to perform a search for the weak reflected line over the velocity range up to this limit increases the probability of a false detection; therefore a higher signal-to-noise ratio is necessary than for the detection of an equally faint line at a known position.

A second consideration concerns the width of the reflected lines. For a planet orbiting in the equatorial plane, the observed Doppler width v_{refl} of the reflected lines is given by

$$v_{\mathrm{refl}} = 2\pi R_* \sin i \left| \frac{1}{P_{\mathrm{rot}}} - \frac{1}{P_{\mathrm{orb}}} \right| = v_* \sin i \left| 1 - \frac{P_{\mathrm{rot}}}{P_{\mathrm{orb}}} \right| \ , \tag{90}$$

where P_{rot} is the rotational period of the star, P_{orb} the orbital period of the planet, and $v_* \sin i$ the Doppler width of the directly observed stellar lines. In the case of "hot Jupiters", P_{rot} and P_{orb} may be similar to each other; it has also been argued that a convective envelope with mass $m \approx 0.01\,M_\odot$ could become tidally locked in less than the age of the system (Marcy et al. 1997). In this case, there is no relative motion between any point on the stellar surface and the planet. The reflected spectrum therefore does not show any broadening due to stellar rotation, in agreement with (90). The broadening caused by the rotation of the planet (which is certainly tidally locked) is small, because of the small planetary radius. The width of the reflected lines is therefore dominated by convective motions in the stellar photosphere, and may thus be substantially smaller than $v_* \sin i$.

The most obvious observing strategy consists in acquiring spectra with good coverage near the phases $\Phi = 0$ and $\Phi = 0.5$, when the separation in velocity space between the "ghosts" and the directly observed lines is largest. One can then subtract a mean spectrum obtained by averaging over all phases from each individual spectrum, and search in these difference spectra for a ghost feature whose location varies sinusoidally with phase according to (88). Alternatively, one can observe around $\Phi = 0.75$, when the velocity separation between the direct and reflected light are smaller than the line width, and search for subtle distortions of the line shape due to the time-varying contribution from the planet (Charbonneau et al. 1998). In either approach, the spectral regions around many stellar lines have to be analyzed together to achieve a signal-to-noise ratio that comes close to the requirement for a planet detection.

Classification of Extrasolar Giant Planets

The geometric albedo of a planet and the wavelength dependence of this albedo are determined by the relative strengths of a few scattering and absorption processes; the most important of these are Rayleigh scattering, molecular absorption, and scattering by atmospheric condensates (e.g. Marley et al. 1999). If the scattering cross section is much larger than the absorption cross section, there is a good chance that an incoming photon will be reflected back (the geometric albedo of an infinitely deep purely Rayleigh-scattering atmosphere is 0.75); conversely a large absorption cross section leads to a low albedo. Rayleigh scattering, with a λ^{-4} wavelength dependence, predominates in the blue, while molecular absorption bands tend to remove red and infrared photons. Cloud-free atmospheres are therefore quite dark at wavelengths $\gtrsim 0.6\,\mu\mathrm{m}$, but water clouds an other condensates can form bright reflecting layers. In addition, Raman scattering and photochemical hazes can reduce the geometric albedo in the ultraviolet and blue.

The atmospheric chemistry and stratification has a dramatic influence on the wavelength-dependent albedo of giant planets, and therefore on their detectability in reflected light. Sudarsky et al. (2000) define five distinct classes, which form a temperature sequence; the boundaries between these classes are given by the temperatures at which various types of condensates can form.

Class I: "Jovian" planets.

At $T_{\mathrm{eff}} \lesssim 150\,\mathrm{K}$, the albedo spectrum is determined mainly by reflection from condensed NH_3, and absorption from molecular CH_4. Ammonia clouds keep the albedo high at $\lambda \lesssim 1.5\,\mu\mathrm{m}$, except in methane absorption bands. At longer wavelengths, the molecular absorption cross sections tend to become larger, which leads to an increased probability of absorption above the cloud deck, and therefore a lower albedo.

Class II: "Water cloud" planets.

At somewhat higher temperatures, $T_{\rm eff} \approx 250\,{\rm K}$, very strongly reflective H_2O clouds develop in the upper atmosphere. Because these clouds form higher in the atmosphere than the NH_3 clouds of class I objects, the visible-wavelength albedo of class II objects is even higher than that of class I planets.

Class III: "Clear" planets.

The atmospheres of planets in the range $350\,{\rm K} \lesssim T_{\rm eff} \lesssim 900\,{\rm K}$ are essentially free of condensates, so that their albedos are determined predominantly by atomic and molecular absorption and Rayleigh scattering. The photons can penetrate to depths where sodium and potassium absorption as well as H_2–H_2 collision-induced absorption play an important role. These processes, as well as absorption by CH_4 and H_2O, make the albedo low through most of the visible region, and almost negligible in the near-IR.

Class IV: "Roasters".

For temperatures $900\,{\rm K} \lesssim T_{\rm eff} \lesssim 1,500\,{\rm K}$, which are expected for the planets with small orbital radii, the equilibrium abundance of alkali metal atoms is even higher than in class III objects. A silicate cloud deck exists at moderate pressures, but it is so deep in the atmosphere that it has no significant effect on the albedo. Absorption above the silicate clouds by sodium and potassium atoms as well as ro-vibrational molecular bands renders class IV objects very dark in the visible and near-IR spectral range.

Class V: "Hot roasters".

If the temperature is very high, $T_{\rm eff} \gtrsim 1,500\,{\rm K}$, the silicate clouds are located much higher in the atmosphere. This increases the albedo dramatically, and makes class V objects much brighter than class IV planets. The transition temperature between these two classes depends on the gravity and is reduced for less massive planets.

In summary, the albedo of extrasolar giant planets depends primarily on their effective temperature; the coolest and hottest objects have much higher albedo than those at intermediate temperatures. The class IV models by Sudarsky et al. (2000) have very low bond albedo ($< 1\%$ for irradiation by a cool star); this has to be taken into account for the computation of $T_{\rm eff}$, of course (see (79)).

For detailed comparisons between observations of reflection spectra and models, the temperature–pressure profile in the atmosphere has to be modeled more carefully (Seager and Sasselov 1998). The best approach is a self-consistent computation of the atmospheric structure that takes the external irradiation into account in the solution of the radiative transfer equation (Barman et al. 2001). The predictive power of these models is limited by the

unknown "weather patterns" in the planetary atmosphere. It is unlikely that the clouds form a homogeneous layer; one should thus expect variations of the albedo across the planetary surface. Large differences in the atmospheric structure may exist between the day- and nightside, especially for close-in planets whose rotation is tidally locked to the orbital period. These differences depend on the efficiency with which the radiation received on the dayside is redistributed across the whole planetary surface, which may be fairly low (see Guillot in this volume). As these considerations show, detailed observations of the reflected spectrum as a function of phase angle have the potential to provide a wealth of information on the atmospheric structure of extrasolar giant planets.

Searches for Reflected Light from "Hot Jupiters"

Several attempts have been made to detect the reflected light from the planet orbiting τ Boo, which has one of the smallest orbital radii of all known planets. Observations with the HIRES spectrograph at the Keck I Telescope gave an upper limit of $\epsilon \lesssim 5 \cdot 10^{-5}$ at $\lambda \approx 480\,\mathrm{nm}$; together with the assumption that $R_p = 1.2\,R_{\mathrm{jup}}$, this limits the geometric albedo near 480 nm to $p \lesssim 0.3$ (Charbonneau et al. 1999). Based on initial observations with the William Herschel Telescope (WHT) at La Palma in a similar wavelength range, a tentative detection of τ Boo B was reported by Collier Cameron et al. (1999). This claim was retracted, however, when further observations by the same group did not confirm the initial result. The latter data imply a geometric albedo $p < 0.22$, under the simplifying assumptions of a wavelength-independent albedo in the range $387.4\,\mathrm{nm} \leq \lambda \leq 586.3\,\mathrm{nm}$, a Lambert-sphere phase-function (85), and a radius $R_p = 1.2\,R_{\mathrm{jup}}$. This appears to exclude the existence of a high cloud deck as expected for class V "hot roasters", but is compatible with the predicted class IV "roaster" spectra.

A complementary set of observations in the infrared at wavenumbers near $3{,}044\,\mathrm{cm}^{-1}$ has been performed with the Infrared Telescope Facility (IRTF) on Mauna Kea (Wiedemann et al. 2001). The goal of this experiment was not the detection of the reflected ghost of the stellar absorption spectrum, but rather the detection of methane absorption features in the planetary atmosphere, which should also vary with time according to (88). The analysis of this data set yielded a weak signal at the $\epsilon \approx 2 \cdot 10^{-4}$ level, but the attribution of this signal to the planet remains very doubtful.

The innermost planet of υ And may offer even better prospects of detection than τ Boo B, because it has a lower mass and may thus be a class V rather than a class IV object. A search for reflected light from υ And B with the WHT did not result in a clear detection, however (Collier Cameron et al. 2002). The observations give an upper limit of $R_p \leq 1.51\,R_{\mathrm{jup}}$ if υ And B indeed has a class V spectrum, but the constraint is weaker, $R_p \leq 2.23\,R_{\mathrm{jup}}$ if the spectrum is assumed to be of class IV. In summary therefore, the searches for reflected light from hot Jupiters conducted so far have given only upper limits on the

combination of albedo and planetary diameter that are consistent with the expectations for these objects.

7 The Effects of Atmospheric Turbulence on Astronomical Observations

Turbulence in the Earth's atmosphere is a major obstacle to the detection of planets with coronographic and interferometric methods from the ground. It limits the contrast achievable with high-resolution imaging and the precision of astrometric measurements. Atmospheric turbulence also determines many of the key design parameters of adaptive optics systems and interferometers: site selection, operating wavelength, aperture size, temporal bandwidth of the servo loops, and integration times. It is therefore important to understand how turbulence is generated in the atmosphere, and how its effects on the propagation of light can be quantified. This chapter gives a brief outline of atmospheric turbulence in the "standard" Kolmogorov model; more detailed treatments of this topic have been given by Roddier (1981, 1989), Fried (1994), and Hardy (1998).

7.1 The Kolmogorov Turbulence Model

Eddies in the Turbulent Atmosphere

The properties of fluid flows are determined primarily by the well-known Reynolds number $\mathcal{R} = VL/\nu$, where V is the fluid velocity, L a characteristic length scale, and ν the kinematic viscosity of the fluid. For air, $\nu \approx 1.5 \cdot 10^{-5}\,\mathrm{m^2\,s^{-1}}$, so that atmospheric flows with wind speeds of a few $\mathrm{m\,s^{-1}}$ and length scales of several meters to kilometers have $\mathcal{R} \gtrsim 10^6$ and are therefore almost always turbulent. The turbulent energy is generated by eddies on a large scale L_0; these large eddies spawn a hierarchy of smaller eddies. Dissipation is not important for the large eddies, but the kinetic energy of the turbulent motion is dissipated in small eddies with a typical size l_0. The characteristic size scales L_0 and l_0 are called the *outer scale* and the *inner scale* of the turbulence. There is considerable debate over typical values of L_0; it is probably a few tens to hundreds of meters in most cases (Buscher et al. 1995; Davis et al. 1995; Conan et al. 2000; Linfield et al. 2001; Quirrenbach 2002b). l_0 is of order a few millimeters.

In the so-called *inertial range* between l_0 and L_0, there is a universal description for the turbulence spectrum, i.e., of the turbulence strength as a function of the eddy size, or of the spatial frequency κ. This somewhat surprising result is the underlying reason for the importance of this simple turbulence model, which was developed by Kolmogorov, and is therefore generally known as *Kolmogorov turbulence*.

The spatial structure of a random process can be described by *structure functions*. The structure function $D_x(R_1, R_2)$ of a random variable x measured at positions R_1, R_2 is defined by

$$D_x(R_1, R_2) \equiv \left\langle \left| x(R_1) - x(R_2) \right|^2 \right\rangle \tag{91}$$

(see also (246)). In words: the structure function measures the expectation value of the difference of the values of x measured at two positions R_1 and R_2. For example, the temperature structure function $D_T(R_1, R_2)$ is the expectation value of the difference in the readings of two temperature probes located at R_1 and R_2. In the following paragraph, a simple argument based on dimensional analysis will be used to derive structure functions for the Kolmogorov model.

The Structure Function for Kolmogorov Turbulence

The only two relevant parameters (in addition to l_0 and L_0) that determine the strength and spectrum of Kolmogorov turbulence are the rate of energy generation per unit mass ε, and the kinematic viscosity ν. The units of ε are $\mathrm{J\,s^{-1}\,kg^{-1}} = \mathrm{m^2\,s^{-3}}$, and those of ν are $\mathrm{m^2\,s^{-1}}$. Under the assumption that the turbulence is homogeneous and isotropic, the structure function of the turbulent velocity field, $D_v(R_1, R_2)$, can only depend only on $|R_1 - R_2|$, and can therefore be written as:

$$\begin{aligned} D_v(R_1, R_2) &\equiv \left\langle \left| v(R_1) - v(R_2) \right|^2 \right\rangle \\ &= \alpha \cdot f\big(|R_1 - R_2| \,/\, \beta\big) \ , \end{aligned} \tag{92}$$

where f is some as yet unspecified dimensionless function of a dimensionless argument. It is immediately clear that the dimensions of α must be velocity squared, and those of β length. Since α and β depend only on ε and ν, it follows from dimensional analysis that

$$\alpha = \nu^{1/2}\varepsilon^{1/2} \quad \text{and} \quad \beta = \nu^{3/4}\varepsilon^{-1/4} \, . \tag{93}$$

In addition, the structure function must be independent of ν in the inertial range, because dissipation does not play a role here. This is possible only if f has the functional form

$$f = k \cdot \big(|R_1 - R_2| \,/\, \beta\big)^{2/3} \tag{94}$$

with a dimensionless numerical constant k, because only in this case the dependence on ν drops out in the expression of the structure function:

$$D_v(R_1, R_2) = \alpha \cdot k \cdot \big(|R_1 - R_2| \,/\, \beta\big)^{2/3} = C_v^2 \cdot |R_1 - R_2|^{2/3} \, , \tag{95}$$

where $C_v^2 \equiv \alpha \cdot k/\beta^{2/3} = k \cdot \varepsilon^{2/3}$. We have thus derived the important result mentioned above, namely a universal description of the turbulence spectrum. It has only one parameter C_v^2, which describes the turbulence strength.

Structure Function and Power Spectral Density of the Refractive Index

The turbulence, with a velocity field characterized by (95), mixes different layers of air, and therefore carries around "parcels" of air with different temperature. Since these "parcels" are in pressure equilibrium, they must have different densities ρ, and therefore different indices of refraction n. The "parcels" are carried along by the velocity field of the turbulence. The temperature fluctuations therefore also follow Kolmogorov's Law with a new parameter C_T^2:

$$D_T(R_1, R_2) = C_T^2 \cdot |R_1 - R_2|^{2/3} \ ; \tag{96}$$

note that this is completely analogous to (95). From the Ideal Gas Law, and $N \equiv (n-1) \propto \rho$, it follows that the structure function of the refractive index is

$$D_n(R_1, R_2) = D_N(R_1, R_2) = C_N^2 \cdot |R_1 - R_2|^{2/3} \ , \tag{97}$$

with C_N given by

$$C_N = \left(7.8 \cdot 10^{-5} P[\mathrm{mbar}]/T^2[\mathrm{K}]\right) \cdot C_T \, . \tag{98}$$

It should be noted that (97) contains a complete description of the statistical properties of the refractive index fluctuations, on length scales between l_0 and L_0. It is possible to calculate related quantities such as the power spectral density Φ from the structure function D. Now we write $R \equiv R_1 - R_2$, and use the relation between the structure function and the covariance (247), and the Wiener-Khinchin Theorem (245). In this way we obtain from (97):

$$C_N^2 \cdot R^{2/3} = D_N(R) = 2 \int_{-\infty}^{\infty} d\kappa \big[1 - \exp(2\pi i \kappa R)\big] \Phi(\kappa) \, . \tag{99}$$

Calculating $\Phi(\kappa)$ from this relation is a slightly non-trivial task[27]; the result is:

$$\Phi(\kappa) = \frac{\Gamma(\frac{5}{3}) \sin \frac{\pi}{3}}{(2\pi)^{5/3}} \, C_N^2 \kappa^{-5/3} = 0.0365 \, C_N^2 \kappa^{-5/3} \, . \tag{100}$$

We have thus obtained the important result that the power spectrum of Kolmogorov turbulence follows a $\kappa^{-5/3}$ law in the inertial range.[28]

---

[27] See Tatarski (1961). Note that his definition of the power spectral density has an additional factor $\frac{1}{2\pi}$, and that his ω corresponds to $2\pi\kappa$.

[28] Note: We have defined $R = |R_1 - R_2|$ and κ as one-dimensional variables, and consequently used a one-dimensional Fourier transform in (99). Sometimes three-dimensional quantities $\vec{R}$ and $\vec{\kappa}$ are used instead. Then a three-dimensional Fourier transform with volume element $4\pi |\vec{\kappa}|^2 \, d|\vec{\kappa}|$ has to be used in (99), and the result is a power spectrum $\Phi(|\vec{\kappa}|) \propto |\vec{\kappa}|^{-11/3}$.

7.2 Wave Propagation Through Turbulence

The Effects of Turbulent Layers

We now look at the propagation of an initially flat wavefront through a turbulent layer of thickness δh at height h. The phase shift produced by refractive index fluctuations is

$$\phi(x) = k \int_h^{h+\delta h} dz\, n(x,z)\ , \tag{101}$$

where $k \equiv 2\pi/\lambda$ is the wavenumber corresponding to the observing wavelength. For layers that are much thicker than the individual turbulence cells, many independent variables contribute to the phase shift. Therefore the Central Limit Theorem implies that ϕ has Gaussian statistics.

We will now use the statistical properties of the refractive index fluctuations, which were calculated in Sect. 7.1, to derive the statistical behavior of the wavefront $\psi(x) = \exp i\phi(x)$. We first express the coherence function $B_h(r)$ of the wavefront after passing through the layer at height h in terms of the phase structure function (see Sect. 11 for definitions):

$$\begin{aligned} B_h(r) &\equiv \left\langle \psi(x)\psi^*(x+r) \right\rangle \\ &= \left\langle \exp i\big[\phi(x) - \phi(x+r)\big] \right\rangle \\ &= \exp\left(-\tfrac{1}{2} \left\langle \big|\phi(x) - \phi(x+r)\big|^2 \right\rangle \right) \\ &= \exp\left(-\tfrac{1}{2} D_\phi(r) \right)\ . \end{aligned} \tag{102}$$

Here we have used the fact that $[\phi(x) - \phi(x+r)]$ has Gaussian statistics with zero mean, and applied the relation

$$\left\langle \exp(\alpha\chi) \right\rangle = \exp\left(\tfrac{1}{2}\alpha^2 \left\langle \chi^2 \right\rangle \right) \tag{103}$$

for Gaussian variables χ with zero mean, which can easily be verified by carrying out the integral over the distribution function.

Calculation of the Phase Structure Function

The next step is the computation of $D_\phi(r)$. We start with the covariance $B_\phi(r)$, which is by definition (240):

$$\begin{aligned} B_\phi(r) &\equiv \left\langle \phi(x)\phi(x+r) \right\rangle \\ &= k^2 \int_h^{h+\delta h} \int_h^{h+\delta h} dz'\, dz''\, \left\langle n(x,z') n(x+r,z'') \right\rangle \\ &= k^2 \int_h^{h+\delta h} dz' \int_{h-z'}^{h+\delta h-z'} dz\, B_N(r,z)\,. \end{aligned} \tag{104}$$

Here we have introduced the new variable $z \equiv z'' - z'$, and the covariance $B_N(r,z)$ of the refractive index variations. For δh much larger than the correlation scale of the fluctuations, the integration can be extended from $-\infty$ to ∞, and we obtain:

$$B_\phi(r) = k^2 \delta h \int_{-\infty}^{\infty} dz\, B_N(r,z)\,. \tag{105}$$

Now we can use (247) again, first for $D_\phi(r)$, then for $D_N(r,z)$ and $D_N(0,z)$, and get:

$$\begin{aligned} D_\phi(r) &= 2\big[B_\phi(0) - B_\phi(r)\big] \\ &= 2k^2\delta h \int_{-\infty}^{\infty} dz\, \big[B_N(0,z) - B_N(r,z)\big] \\ &= 2k^2\delta h \int_{-\infty}^{\infty} dz\, \Big[\big(B_N(0,0) - B_N(r,z)\big) - \big(B_N(0,0) - B_N(0,z)\big)\Big] \\ &= k^2\delta h \int_{-\infty}^{\infty} dz\, \big[D_N(r,z) - D_N(0,z)\big]\,. \end{aligned} \tag{106}$$

Inserting from (97) gives:

$$\begin{aligned} D_\phi(r) &= k^2\delta h\, C_N^2 \int_{-\infty}^{\infty} dz\, \left[\left(r^2+z^2\right)^{1/3} - |z|^{2/3}\right] \\ &= \frac{2\Gamma(\frac{1}{2})\Gamma(\frac{1}{6})}{5\Gamma(\frac{2}{3})}\, k^2\delta h\, C_N^2\, r^{5/3} \\ &= 2.914\, k^2\delta h\, C_N^2\, r^{5/3}\,. \end{aligned} \tag{107}$$

This is the desired expression for the structure function of phase fluctuations due to Kolmogorov turbulence in a layer of thickness δh.

Wavefront Coherence Function and Fried Parameter

We are now in a position to put the results of the previous sections together. Inserting (107) into (102), we get:

$$B_h(r) = \exp\left[-\tfrac{1}{2}\left(2.914\, k^2 C_N^2\, \delta h\, r^{5/3}\right)\right]\,. \tag{108}$$

This expression can now be integrated over the whole atmosphere. In the process, we also take into account that we are not necessarily looking in the vertical direction. Introducing the zenith angle z, this leads to:

$$B(r) = \exp\left[-\tfrac{1}{2}\left(2.914\, k^2 (\sec z) r^{5/3} \int dh\, C_N^2(h)\right)\right]\,. \tag{109}$$

To simplify the notation, it is now convenient to define the *Fried parameter* r_0 by

$$r_0 \equiv \left[0.423\, k^2 (\sec z) \int dh\, C_N^2(h)\right]^{-3/5} , \tag{110}$$

and we can write

$$B(r) = \exp\left[-3.44 \left(\frac{r}{r_0}\right)^{5/3}\right] \quad , \quad D_\phi(r) = 6.88 \left(\frac{r}{r_0}\right)^{5/3} . \tag{111}$$

We have thus derived fairly simple expressions for the wavefront coherence function and the phase structure function. They depend only on the Fried parameter r_0, which in turn is a function of turbulence strength, zenith angle, and wavelength. The significance of the Fried parameter will be discussed further in Sect. 7.4.

7.3 The Effect of Turbulence on Astronomical Images

Optical Image Formation

The complex amplitude A of a wave ψ diffracted at an aperture P with area $\mathit{\Pi}$ is given by Huygens' principle, which states that each point in the aperture can be considered as the center of an emerging spherical wave. In the far field (i.e., in the case of Fraunhofer diffraction), the spherical waves are equivalent to plane waves, and we can write down the expression for the amplitude as a function of position α in the focal plane:

$$A(\alpha) = \frac{1}{\sqrt{\mathit{\Pi}}} \int dx\, \psi(x) P(x) \exp(-2\pi i \alpha x/\lambda) . \tag{112}$$

Here we describe the aperture P by a complex function $P(x)$. In the simple case of a fully transmissive and aberration-free aperture $P(x) \equiv 1$ inside the aperture, and $P(x) \equiv 0$ outside. Introducing the new variable $u \equiv x/\lambda$ we can write this as a Fourier relation:

$$A(\alpha) = \frac{1}{\sqrt{\mathit{\Pi}}}\, FT\big[\psi(u) P(u)\big] . \tag{113}$$

The normalization in (112) and (113) has been chosen such that the illumination S in the focal plane is given by the square of the wave amplitude:

$$S(\alpha) = \big|A(\alpha)\big|^2 = \frac{1}{\mathit{\Pi}} \Big|FT\big[\psi(u) P(u)\big]\Big|^2 . \tag{114}$$

Applying the Wiener-Khinchin Theorem (245) to this equation we get

$$S(f) = \frac{1}{\mathit{\Pi}} \int du\, \psi(u) \psi^*(u+f) P(u) P^*(u+f) . \tag{115}$$

This equation describes the spatial frequency content $S(f)$ of images taken through the turbulent atmosphere, if ψ is identified with the wavefront after passing through the turbulence. Taking long exposures (in practice this means exposures of at least a few seconds) means averaging over many different realizations of the state of the atmosphere:

$$\begin{aligned}\left\langle S(f)\right\rangle &= \frac{1}{\Pi}\int du\left\langle\psi(u)\psi^*(u+f)\right\rangle P(u)P^*(u+f)\\ &= B(f)\cdot T(f)\,.\end{aligned} \tag{116}$$

Here we have introduced the *telescope transfer function*

$$T(f) = \frac{1}{\Pi}\int du\,P(u)P^*(u+f)\,. \tag{117}$$

Equation (116) contains the important result that for long exposures the optical transfer function is the product of the telescope transfer function and the atmospheric transfer function, which is equal to the wavefront coherence function $B(f)$.

Diffraction-Limited Images and Seeing-Limited Images

The resolving power $\mathcal{R}$ of an optical system can very generally be defined by the integral over the optical transfer function. For the atmosphere–telescope system this means:

$$\mathcal{R} \equiv \int df\,S(f) = \int df\,B(f)T(f)\,. \tag{118}$$

In the absence of turbulence, $B(f)\equiv 1$, and we obtain the *diffraction-limited* resolving power of a telescope with diameter D:

$$\begin{aligned}\mathcal{R}_{\rm tel} &= \int df\,T(f) = \frac{1}{\Pi}\int\int dudf\,P(u)P^*(u+f)\\ &= \frac{1}{\Pi}\left|\int du\,P(u)\right|^2 = \frac{\pi}{4}\left(\frac{D}{\lambda}\right)^2\,.\end{aligned} \tag{119}$$

The last equality assumes a circular aperture and shows the relation of $\mathcal{R}$ to the more familiar Rayleigh criterion $1.22\cdot\lambda/D$. Working with $\mathcal{R}$ instead of using the Rayleigh criterion has the advantage that $\mathcal{R}$ is a well-defined quantity for arbitrary aperture shapes and in the presence of aberrations.

For strong turbulence and rather large telescope diameters, $T = 1$ in the region where B is significantly different from zero, and we get the *seeing-limited* resolving power:

$$\begin{aligned}\mathcal{R}_{\rm atm} &= \int df\,B(f) = \int df\exp\left[-3.44\left(\frac{\lambda f}{r_0}\right)^{5/3}\right]\\ &= \frac{6\pi}{5}\Gamma(\tfrac{6}{5})\left[3.44\left(\frac{\lambda}{r_0}\right)^{5/3}\right]^{-6/5} = \frac{\pi}{4}\left(\frac{r_0}{\lambda}\right)^2\,.\end{aligned} \tag{120}$$

Here we have used (111) with $r = \lambda f$ for the wavefront coherence function $B(f)$.

7.4 Fried Parameter and Strehl Ratio

The Significance of the Fried Parameter r_0

A comparison of (119) and (120) elucidates the significance of the Fried parameter for image formation, and reveals the reason for the peculiar choice of the numerical constant 0.423 in (110): *The resolution of seeing-limited images obtained through an atmosphere with turbulence characterized by a Fried parameter r_0 is the same as the resolution of diffraction-limited images taken with a telescope of diameter r_0.* Observations with telescopes much larger than r_0 are seeing-limited, whereas observations with telescopes smaller than r_0 are essentially diffraction-limited. It can also be shown that the mean-square phase variation over an aperture of diameter r_0 is about $1\,\mathrm{rad}^2$ (more precisely, $\sigma_\phi^2 = 1.03\,\mathrm{rad}^2$). These results can be captured in an extremely simplified picture that describes the atmospheric turbulence by r_0-sized "patches" of constant phase, and random phases between the individual patches. While this picture can be useful for some rough estimates, one should keep in mind that Kolmogorov turbulence has a continuous spectrum ranging from l_0 to L_0, as described by (100).

The scaling of r_0 with wavelength and zenith angle implied by (110) has far-reaching practical consequences. Since

$$r_0 \propto \lambda^{6/5} \ , \tag{121}$$

it is much easier to achieve diffraction-limited performance at longer wavelengths. For example, the number of degrees of freedom (the number of actuators on the deformable mirror and the number of subapertures in the wavefront sensor) in an adaptive optics system must be of order $(D/r_0)^2 \propto \lambda^{-12/5}$. An interferometer works well only if the wavefronts from the individual telescopes are coherent (i.e., have phase variances not larger than about $1\,\mathrm{rad}^2$); therefore the maximum useful aperture area of an interferometer is $\propto \lambda^{12/5}$ (unless the wavefronts are corrected with adaptive optics). Equation (121) implies that the width of seeing-limited images, $\theta \approx 1.2 \cdot \lambda/r_0 \propto \lambda^{-1/5}$, varies only slowly with λ; it is somewhat better at longer wavelengths. In addition, we see from (110) that $r_0 \propto (\sec z)^{-3/5}$; the seeing gets worse with increasing zenith angle.

From this discussion it should be clear that the value of r_0 – given by the integral over C_N^2 – is a crucial parameter for high-resolution observations. At good sites, such as Mauna Kea or Cerro Paranal, r_0 is typically of order 20 cm at 500 nm, which corresponds to an image FWHM of $0\farcs6$. The scaling of r_0 with λ (121) implies that in the mid-infrared ($\lambda \gtrsim 10\,\mu\mathrm{m}$) even the 10 m Keck Telescopes are nearly diffraction-limited, whereas a 1.8 m telescope has $D/r_0 \sim 2$ at $\lambda = 2\,\mu\mathrm{m}$ and $D/r_0 \sim 5$ at $\lambda = 800\,\mathrm{nm}$. It should be noted that at

any given site r_0 varies dramatically from night to night; at any given time it may be a factor of 2 better than the median or a factor of 5 worse. In addition, the seeing fluctuates on all time scales down to minutes and seconds; this has to be taken into account in calibration procedures and in the design of servo loops for adaptive optics systems and of fringe trackers for interferometers.

Strehl Ratio

The quality of an aberrated imaging system, or of the wavefront after propagation through turbulence, is often measured by the *Strehl ratio* S. This quantity is defined as the on-axis intensity in the image of a point source divided by the peak intensity in a hypothetical diffraction-limited image taken through the same aperture. For a circular aperture with an aberration function $\psi(\rho, \theta)$, which describes the wavefront distortion (in units of μm or nm) as a function of the spherical coordinates (ρ, θ), the Strehl ratio is given by:

$$S = \frac{1}{\pi^2} \left| \int_0^1 \int_0^{2\pi} \rho \, d\rho \, d\theta \, e^{ik\psi(\rho,\theta)} \right|^2 . \tag{122}$$

From this equation it is immediately clear that $0 \leq S \leq 1$, that $S = 1$ for $\psi = \text{const.}$, that $S \ll 1$ for strongly varying ψ, and that for any given (varying) ψ the Strehl ratio tends to be larger for longer wavelengths (smaller k). In the case of atmospheric turbulence, only the statistical properties of ψ are known. If the r.m.s. phase error $\sigma_\phi \equiv k \, \sigma_\psi$ is smaller than about 2 rad, S can be approximated by the so-called *extended Marechal approximation*:

$$S = e^{-\sigma_\phi^2} . \tag{123}$$

We have seen above ((111) and discussion of the significance of r_0) that

$$\sigma_\phi^2 = 1.03 \left(\frac{D}{r_0} \right)^{5/3} . \tag{124}$$

Equations (123) and (124) show that the Strehl ratio for images obtained with a telescope of diameter $D = r_0$ is $S = 0.36$; for $D \gtrsim r_0$ the Strehl ratio decreases precipitously with telescope diameter. Equivalently, S decreases sharply with decreasing wavelength, since $r_0 \propto \lambda^{6/5}$.

If $S \gtrsim 0.1$ in an imaging application, deconvolution algorithms can usually be applied to obtain diffraction-limited images, but the dynamic range and signal-to-noise ratio are worse than for $S \sim 1$. For example, because of spherical aberration, the Hubble Space Telescope has $S \approx 0.1$ without corrective optics. Before the installation of COSTAR and WFPC2 in the first servicing mission, the imaging performance of HST was severely affected by the flawed optics, although diffraction-limited images could be obtained with image restoration software. In an interferometer, the maximum fringe contrast is roughly proportional to the Strehl ratio if no corrective measures (adaptive

optics or mode filtering with pinholes or single-mode fibers) are taken. Planet detection with imaging requires an extremely high dynamic range, which usually means that a Strehl ratio close to 1 is desired.

7.5 Temporal Evolution of Atmospheric Turbulence

Taylor Hypothesis and τ_0

So far we have discussed the spatial structure of atmospheric turbulence and its effects on image formation. Now we turn to the question of temporal changes of the turbulence pattern. A convenient approximation assumes that the time scale for these changes is much longer than the time it takes the wind to blow the turbulence past the telescope aperture. According to this *Taylor hypothesis of frozen turbulence*, the variations of the turbulence caused by a single layer can therefore be modeled by a "frozen" pattern that is transported across the aperture by the wind in that layer. If multiple layers contribute to the total turbulence, the time evolution is more complicated, but the temporal behavior of the turbulence can still be characterized by a time constant

$$\tau_0 \equiv r_0/v \ , \tag{125}$$

where v is the wind speed in the dominant layer. With typical wind speeds of order $20\,\mathrm{ms}^{-1}$, $\tau_0 \approx 10\,\mathrm{ms}$ for $r_0 = 20\,\mathrm{cm}$. The wavelength scaling of τ_0 is obviously the same as that of r_0, i.e., $\tau_0 \propto \lambda^{6/5}$.

7.6 Temporal Structure Function and Power Spectra

It is sometimes necessary to quantify the temporal behavior of phase fluctuations at a given point in space. If Taylor's hypothesis is valid, we can of course convert the spatial structure function (111) into a temporal structure function:

$$D_\phi(t) = 6.88 \left(\frac{t}{\tau_0}\right)^{5/3} . \tag{126}$$

A calculation similar to the one leading to (100) can be carried out to compute the temporal phase power spectrum

$$\Phi_\phi(f) = 0.077\,\tau_0^{-5/3} f^{-8/3} . \tag{127}$$

This equation tells us which residual phase errors we have to expect if we try to correct atmospheric turbulence with a servo loop of a given bandwidth (e.g., in an adaptive optics system or an interferometric fringe tracker). For example, if we could correct the turbulence perfectly up to a limiting frequency f_0, and not at all at higher frequencies, we would obtain a phase variance that can be computed by integrating (127) from f_0 to ∞. For a more realistic calculation, we have to multiply the phase power spectrum with the response function of the servo loop.

The Long-Exposure and Short-Exposure Limits

Observations with exposure time $t \gg \tau_0$ average over the atmospheric random process; these are the *long exposures* for which (116) and (120) are applicable. In contrast, *short exposures* with $t \ll \tau_0$ produce images through a single instantaneous realization of the atmosphere; these *speckle images* contain information at high spatial frequencies up to the diffraction limit, which can be extracted from series of such images with computer processing (e.g., bispectrum analysis). The parameter τ_0 is also of great importance for the design of adaptive optics systems and interferometers. All control loops that have to reject atmospheric fluctuations – AO control loops, angle trackers, fringe trackers – must have bandwidths larger than $1/\tau_0$. Together r_0 and τ_0 set fundamental limits to the sensitivity of these wavefront control loops: a certain number of photons must arrive per r_0-sized patch during the time τ_0 for the wavefront sensor (or fringe sensor) to work. This implies that the sensitivity scales with $r_0^2 \cdot \tau_0 \propto \lambda^{18/5}$ (for equal photon flux per bandpass).

7.7 Angular Anisoplanatism

The light from two stars separated by an angle θ on the sky passes through different patches of the atmosphere and therefore experiences different phase variations. This *angular anisoplanatism* limits the field corrected by adaptive optics systems and causes phase decorrelation for off-axis objects in interferometers. To calculate the effect of anisoplanatism, we trace back the rays to two stars separated by an angle θ from the telescope pupil. They coincide at the pupil, and their separation $r(d)$ at a distance d is $\theta \cdot d$. At zenith angle z, the distance is related to the height h in the atmosphere by $d = h \sec z$. To calculate the phase variance between the two rays, we insert this relation in

$$\left\langle |\phi(0) - \phi(r)|^2 \right\rangle = D_\phi(r) = 2.914\, k^2 \sec z\, \delta h\, C_N^2\, r^{5/3} \tag{128}$$

(see (phasestruct)), integrate over the height h, and obtain:

$$\begin{aligned} \left\langle \sigma_\phi^2 \right\rangle &= 2.914\, k^2 (\sec z) \int dh\, C_N^2(h)\, (\theta h \sec z)^{5/3} \\ &= 2.914\, k^2 (\sec z)^{8/3} \theta^{5/3} \int dh\, C_N^2(h) h^{5/3} \\ &= \left(\frac{\theta}{\theta_0}\right)^{5/3} , \end{aligned} \tag{129}$$

where we have introduced the *isoplanatic angle* θ_0, for which the variance of the relative phase is $1\,\mathrm{rad}^2$:

$$\theta_0 \equiv \left[2.914\, k^2 (\sec z)^{8/3} \int dh\, C_N^2(h) h^{5/3} \right]^{-3/5} . \tag{130}$$

By comparing the definitions for the Fried parameter r_0 and for θ_0, (110) and (130), we see that

$$\theta_0 = 0.314\,(\cos z)\,\frac{r_0}{H}\ , \tag{131}$$

where

$$H \equiv \left(\frac{\int dh\, C_N^2(h) h^{5/3}}{\int dh\, C_N^2(h)}\right)^{3/5} \tag{132}$$

is the *mean effective turbulence height*. Equations (isodef) and (131) show that the isoplanatic angle is affected mostly by high-altitude turbulence; the anisoplanatism associated with ground layers and dome seeing is very weak. Moreover, we see that θ_0 scales with $\lambda^{6/5}$, but it depends more strongly on zenith angle than r_0. For $r_0 = 20\,\mathrm{cm}$ and an effective turbulence height of 7 km, (131) gives $\theta_0 = 1.8\,\mathrm{arcsec}$. For two stars separated by more than θ_0 the short-exposure point spread functions (or point spread functions generated by adaptive optics) are different. In contrast the long-exposure point spread functions, which represent averages over many realizations of the atmospheric turbulence, are nearly identical even over angles much larger than θ_0.

It should be pointed out that these calculations of anisoplanatism give results that are somewhat too pessimistic. The reason is that a large fraction of the phase variance between the two rays considered is a piston term (i.e., a difference in phase that is constant across the aperture), which doesn't lead to image motion or blurring.[29] Moreover, anisoplanatism is less severe for low spatial frequencies, which most adaptive optics systems correct much better than high spatial frequencies. The degradation of the Strehl ratio with off-axis angle is therefore not quite as bad as suggested by inserting (129) in (123).

7.8 Scintillation

The geometric optics approximation of light propagation that was used in Sect. 7.2 is only valid for propagation pathlengths shorter than the *Fresnel propagation length* $d_F \equiv r_0^2/\lambda$. In other words, the *Fresnel scale*

$$r_F \equiv \sqrt{\lambda L} = \sqrt{\lambda h \sec z}\ , \tag{133}$$

where L is the distance to the dominant layer of turbulence, must be smaller than the Fried scale r_0. For $r_0 = 20\,\mathrm{cm}$ and $\lambda = 500\,\mathrm{nm}$, $d_F = 80\,\mathrm{km}$. This is significantly larger than the height of the layers contributing much to the C_N^2 integrals, and the geometric approximation is a good first-order approach at good sites for visible and infrared wavelengths, as long as the zenith angle is

[29] Note, however, that piston terms have to be taken into account in interferometry, where they are responsible for fluctuations in the relative delay between the two stars.

not too large. ($d_F \propto \lambda^{7/5}$ for Kolmogorov turbulence; therefore the geometric approximation is even better at longer wavelengths.) However, if the propagation length is comparable to d_F or longer, the rays diffracted at the turbulence cells interfere with each other, which causes intensity fluctuations in addition to the phase variations. This phenomenon is called *scintillation*; it is an important error source in high-precision photometry unless the exposure times are sufficiently long to average over the fluctuations. Since scintillation is an interference phenomenon, it is highly chromatic. This effect can be easily observed with the naked eye: bright stars close to the horizon twinkle strongly and change color on time scales of seconds.

Although scintillation is weak for most applications of adaptive optics and interferometry, it has to be taken into account when high Strehl ratios are desired. High-performance adaptive optics systems designed for the direct detection of extrasolar planets have to correct the wavefront errors so well that intensity fluctuations become important. In interferometers that use fringe detection schemes based on temporal pathlength modulation and synchronous photon detection, scintillation noise has to be considered when very small fringe amplitudes are to be measured.

The effects of scintillation can be quantified by determining the relative intensity fluctuations $\delta I/I$; for small amplitudes $\delta I/I = \delta \ln I$. A calculation similar to the one in Sect. 7.2 gives the variance of the log intensity fluctuations:

$$\sigma^2_{\ln I} = 2.24\, k^{7/6} (\sec z)^{11/6} \int dh\, C_N^2(h) h^{5/6} \,. \tag{134}$$

This expression is valid only for small apertures with diameter $D \ll r_F$. For larger apertures, scintillation is reduced by averaging over multiple independent subapertures. This changes not only the amplitude of the intensity fluctuations, but also the functional dependence on zenith angle, wavelength and turbulence height. The expression

$$\sigma^2_{\ln I} \propto D^{-7/3} (\sec z)^3 \int dh\, C_N^2(h) h^2 \,, \tag{135}$$

which is valid for $D \gg r_F$ and $z \lesssim 60°$, shows the expected strong decrease of the scintillation amplitude with aperture size; note that it is independent of the observing wavelength. For larger zenith angles the assumption $\delta \ln I \ll 1$ is no longer valid, the fluctuations increase less strongly with $\sec z$ than predicted by (135), and eventually saturate.

7.9 Turbulence and Wind Profiles

We have seen in the preceding sections that the most important statistical properties of seeing can be characterized by a few numbers: the Fried parameter r_0, the coherence time τ_0, the isoplanatic angle θ_0, and the scintillation index $\sigma_{\ln I}$. For the design and performance evaluation of high-angular-resolution instruments it is of great importance to have reliable statistical

information on these parameters. Therefore extensive seeing monitoring campaigns are normally conducted before decisions are made about the site selection for large telescopes and interferometers, or about the construction of expensive adaptive optics systems. Having access to the output of a continuously running seeing monitor which gives the instantaneous value of r_0 (and ideally also of the other seeing parameters) is also very convenient for debugging and for optimizing the performance of high-resolution instruments.

From equations (110), (125), (129), and (134) it is obvious that all seeing parameters can easily be calculated from moments

$$\mu_m \equiv \int dh\, C_N^2(h) h^m \tag{136}$$

of the turbulence profile $C_N^2(h)$, and (in the case of τ_0) from moments

$$v_m \equiv \int dh\, C_N^2(h) v^m(h) \tag{137}$$

of the wind profile $v(h)$. More complicated analyses such as performance estimates of adaptive optics systems with laser guide stars and of multi-conjugate AO systems also rely on knowledge of $C_N^2(h)$ and $v(h)$. In-situ measurements of these profiles with balloon flights and remote measurements with SCIDAR[30] or related methods are therefore needed to fully characterize the atmospheric turbulence. Figure 41 shows profiles measured on Cerro Paranal, the site of the European Southern Observatory's Very Large Telescope observatory. The decrease of C_N^2 with height is typical for most sites; frequently wind shear at altitudes near 10 km creates additional layers of enhanced turbulence. The highest wind speeds normally occur at heights between 9 and 12 km. Extensive sets of observed turbulence and wind profiles, combined with the analytic methods sketched in this section and numerical simulations, form a firm basis for the evaluation of astronomical sites, and for the design of interferometers and adaptive optics systems.

8 Introduction to Optical Interferometry

The angular resolution required for detection of extrasolar planets drives us to very large telescope sizes. Scaling this approach to mid-infrared wavelengths is certainly impractical; we are thus compelled to consider interferometry as a means to achieve the required resolution. This chapter introduces the basic concepts of optical and infrared interferometry, beam combination schemes, and fringe detection methods. We will then take a look at ways to transfer phase knowledge from one baseline to another, or from one wavelength to another, and discuss the limitations that instrumental and atmospheric errors

[30] The SCIDAR technique is based on auto-correlating pupil images of double stars.

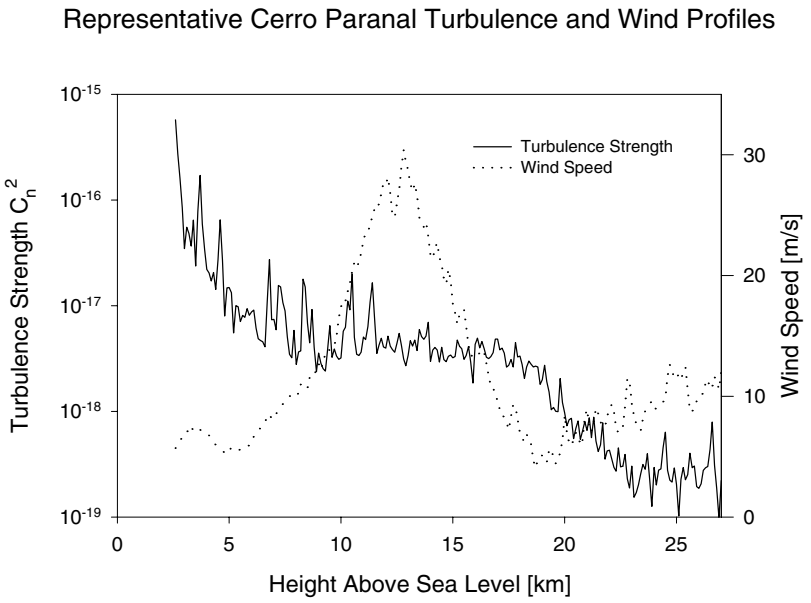

Fig. 41. Turbulence and wind profiles measured on Cerro Paranal, Chile. The turbulence is strongest close to the ground (2,635 m above sea level). The wind speed is highest at an altitude of ∼10 to 15 km. Wind shear often leads to additional layers of strong turbulence at high altitude (only weakly present in this data set)

impose on these techniques. These concepts will be applied to planet detection in the subsequent chapters. For a more detailed tutorial about optical interferometry the reader is referred to Lawson (2000); many details and references to the literature can also be found in the review by Quirrenbach (2001).

8.1 Schematic Design of an Optical Interferometer

Long-baseline interferometry is the coherent combination of light received with separate telescopes. This is shown schematically in Fig. 42. For the viewing direction indicated in this figure, light from a distant star arrives first at the telescope to the right, and a little later at the telescope to the left. The pathlength difference or *delay* D is given by the relation

$$D = \vec{B} \cdot \hat{s} \ , \tag{138}$$

where $\vec{B}$ is the *baseline vector* joining the two telescopes, and $\hat{s}$ the unit vector in the direction toward the star. D is of the order of the baseline length, i.e., up to tens or even hundreds of meters. This is much larger than the coherence length of the stellar light, which is given by $\lambda^2/\Delta\lambda$, where λ is the observing

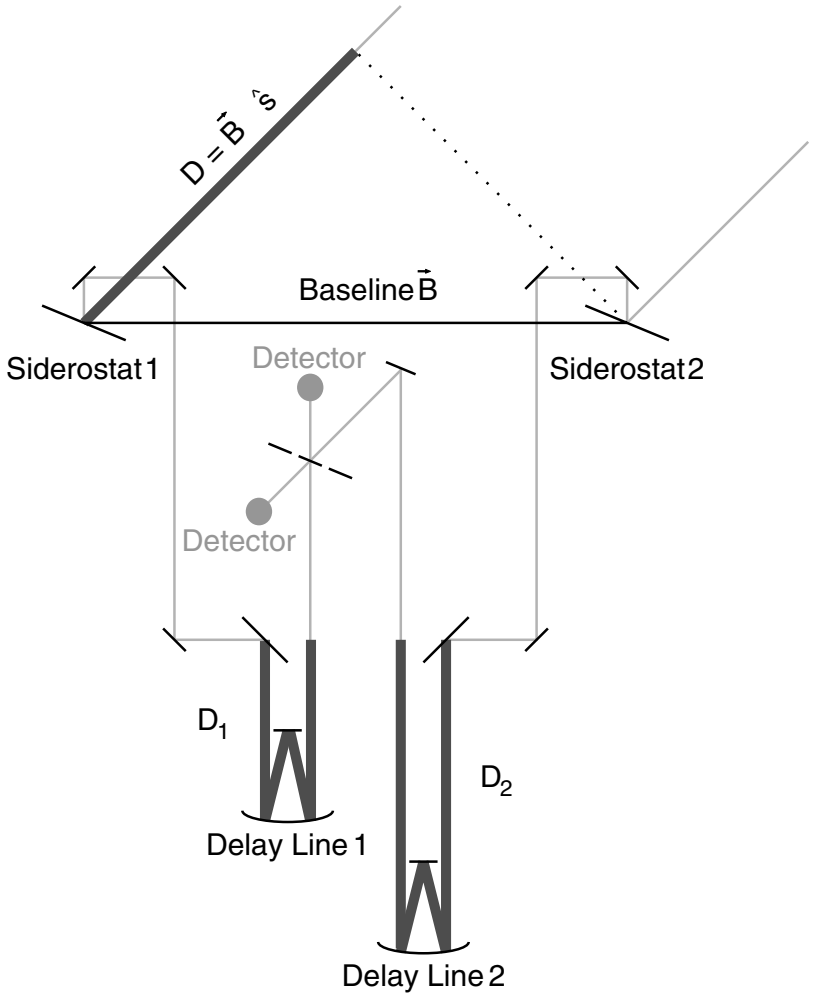

Fig. 42. Schematic drawing of the light path through a two-element interferometer. The external delay $D = \vec{B} \cdot \hat{s}$ is compensated by the two delay lines. The pathlengths D_1, D_2 through the delay lines are monitored with laser interferometers. The zero-order interference maximum occurs when the delay line positions are such that the internal delay $D_{\rm int} = D_2 - D_1$ is equal to D

wavelength and $\Delta\lambda$ the bandwidth of the filter used for the observations. To observe interference fringes, it is therefore necessary to compensate the external delay D with an opposite internal delay.

$$V = \frac{I_{\rm max} - I_{\rm min}}{I_{\rm max} + I_{\rm min}}. \tag{139}$$

8.2 Beam Combination Concepts

The various beam combination schemes that can be employed in astronomical interferometry may be classified according to several criteria (**?**): the beam étendue (single-mode or multi-mode), the beam direction (co-axial or multi-axial), the combination plane (image plane or pupil plane), and the relation between input and output pupils (Michelson or Fizeau configuration, see below). For N telescopes in an array, there are $N(N-1)/2$ baselines. The $N(N-1)/2$ visibilities can either be measured by pairwise beam combination, or by bringing the light from all telescopes together on one detector. In the latter "all-on-one" techniques the fringes from the different baselines have to be encoded either spatially (by using a non-redundant output pupil) or temporally (by using different dither frequencies for the beams from individual telescopes).

Unlike in radio astronomy, where the radiation is detected and amplified before correlation, in an optical interferometer the beam combination occurs before detection. For pairwise beam combination the light from each telescope has to be divided in $(N-1)$ beams; in "all-on-one" schemes the visibility measurement for each baseline is affected by noise contributed by the $(N-2)$ other telescopes. This means that a baseline that is part of an N-element array is always less sensitive than an equivalent two-telescope interferometer. The detailed trade-offs between different beam combination schemes depend on the predominant noise source (background, detector, photon noise), detector cost and availability, and other technical considerations. Armstrong et al. (1998) discuss the case of pupil-plane combination with temporal encoding of the fringes, for which in the photon-rich regime

$$\mathrm{SNR} \propto \left(\frac{n_{\mathrm{phot}} N}{N_{\mathrm{out}}} \right)^{1/2} \frac{V}{N_{\mathrm{corr}}} \ , \tag{140}$$

where n_{phot} is the photon rate from each telescope, N the number of array elements (equal to the number of input beams to the combiner), N_{out} the number of output beams from the combiner, and N_{corr} the number of input beams combined to produce each output beam. For pairwise combination, $N_{\mathrm{corr}} = 2$ and $N_{\mathrm{out}} = N(N-1)$. (Note that combining beams at beamsplitting surfaces produces two output beams.) For "all-on-one" combination $N_{\mathrm{corr}} = N$ and $N_{\mathrm{out}} = 2$. In both cases $\mathrm{SNR} \propto N^{-1/2}$, which demonstrates that multi-element arrays are indeed less sensitive than single-baseline instruments. Equation (140) gives a $\sim\sqrt{2}$ advantage of the "all-on-one" technique over pairwise beam combination, but the required temporal encoding of the fringes is difficult to realize technically.

Michelson and Fizeau Interferometers

In a Fizeau interferometer the output pupil is an exact replica of the input pupil, scaled only by a constant factor. This is also known as homothetic

mapping between input and output pupil. In contrast, in a Michelson interferometer there is no homothetic relation between the input and output pupils.[31] This means that the object–image relationship can no longer be described as a convolution, because the rearrangement of the apertures rearranges the high-spatial frequency part of the object spectrum in the Fourier plane (Tallon and Tallon-Bosc 1992). This has an important consequence for off-axis objects: the image position does not coincide with the white-light fringe position (see Fig. 1 in Tallon and Tallon-Bosc 1992). For a finite spectral bandwidth this means that the fringe contrast decreases with field angle and the field-of-view is limited; the maximum size of an image from a Michelson interferometer is $\sim R \equiv \lambda/\Delta\lambda$ resolution elements in diameter. This effect is known as "bandwidth smearing" in radio astronomy (see Sect. 8.4). If a Michelson interferometer is used with image plane beam combination, and the visibilities are estimated by integration over each fringe peak, the field-of-view is additionally restricted to the size of one Airy disk of the individual telescopes.

8.3 Source Coherence and Interferometer Response

The *source coherence function* is defined as

$$\gamma(\xi_1, \xi_2, \tau) = \left\langle E(\xi_1, t) E^*(\xi_2, t-\tau) \right\rangle , \tag{141}$$

where E is the radiation field, and ξ the direction cosine on the sky. For $\xi_1 = \xi_2 = \xi$, γ is the time autocorrelation function of the radiation from direction ξ; for $\tau = 0$, this is the time-averaged brightness from that direction $\left\langle |E(\xi)|^2 \right\rangle$. An extended source is *spatially incoherent* if $\gamma = 0$ for $\xi_1 \neq \xi_2$; in this case

$$\gamma(\xi_1, \xi_2, \tau) = \gamma(\xi_1, \tau) \cdot \delta(\xi_1 - \xi_2) . \tag{142}$$

An interferometer with antennae at positions u_1, u_2 measures essentially the Fourier transform of γ:

$$\Gamma(u_1, u_2, \tau) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \gamma(\xi_1, \xi_2, \tau)\, \mathrm{e}^{-2\pi i(\xi_1 u_1 - \xi_2 u_2)} d\xi_1 d\xi_2 . \tag{143}$$

In the spatially incoherent case, the interferometer response depends only on the *difference vector* of the antenna positions:

$$\Gamma(u, \tau) = \int_{-\infty}^{\infty} \gamma(\xi, \tau)\, \mathrm{e}^{-2\pi i \xi u} d\xi . \tag{144}$$

[31] I take this as the definition of Fizeau and Michelson interferometers. Sometimes these terms are also used to mean "image plane" and "pupil plane" interferometer, respectively. In my nomenclature, it is possible to build an image plane Michelson interferometer.

The interferometer output at zero delay is called *complex visibility*; the complex visibility is the Fourier transform of the source brightness distribution:

$$V = \Gamma(u,0) = \int_{-\infty}^{\infty} \left\langle \left|E(\xi)\right|^2 \right\rangle \, \mathrm{e}^{-2\pi i \xi u} d\xi \,. \tag{145}$$

Each observation on one baseline measures one Fourier component of the sky brightness distribution.

8.4 Bandwidth and Interferometric Field-of-View

For monochromatic light, the interferometer response is:

$$F = \cos\left(\frac{2\pi B}{\lambda}\sin\theta\right) = \cos\left(\frac{2\pi B\xi}{\lambda}\right) . \tag{146}$$

For a rectangular bandpass with width $\Delta\nu$, the response is:

$$\begin{aligned} F(\nu_0) &= \frac{1}{\Delta\nu}\int_{\nu_0-\Delta\nu/2}^{\nu_0+\Delta\nu/2} \cos\left(\frac{2\pi B\xi\nu}{c}\right) d\nu \\ &= \cos\left(\frac{2\pi B\xi\nu_0}{c}\right)\cdot\frac{\sin(\pi B\xi\Delta\nu/c)}{\pi B\xi\Delta\nu/c} \\ &= \cos\left(\frac{2\pi B\xi\nu_0}{c}\right)\cdot\mathrm{sinc}\left(\frac{\pi B\xi\Delta\nu}{c}\right) \,, \end{aligned} \tag{147}$$

where we have again used the function $\mathrm{sinc}(x) \equiv \sin(x)/x$. This *bandwidth smearing* limits the field-of-view of the interferometer; the maximum size of the field is of order $R = \nu/\Delta\nu$ resolution elements:

$$\xi_{\max} \approx R \cdot \lambda/B. \tag{148}$$

8.5 Fringe Detection

Delay Modulation and ABCD Detection

In each of the spectral channels, arriving photons are counted synchronously with the delay modulation in bins corresponding to $\lambda/4$. (Since the physical stroke is equal to λ only in the channel with the longest wavelength, dead time has to be added in the electronics at the end of the stroke in the other three channels.) From the four bin counts A, B, C, and D (see in Fig. 43, the square of the visibility V^2 can be estimated using

$$V^2 = \frac{\pi^2}{2}\cdot\frac{\left\langle X^2+Y^2-N\right\rangle}{\left\langle N-N_{\mathrm{dark}}\right\rangle^2} \,, \tag{149}$$

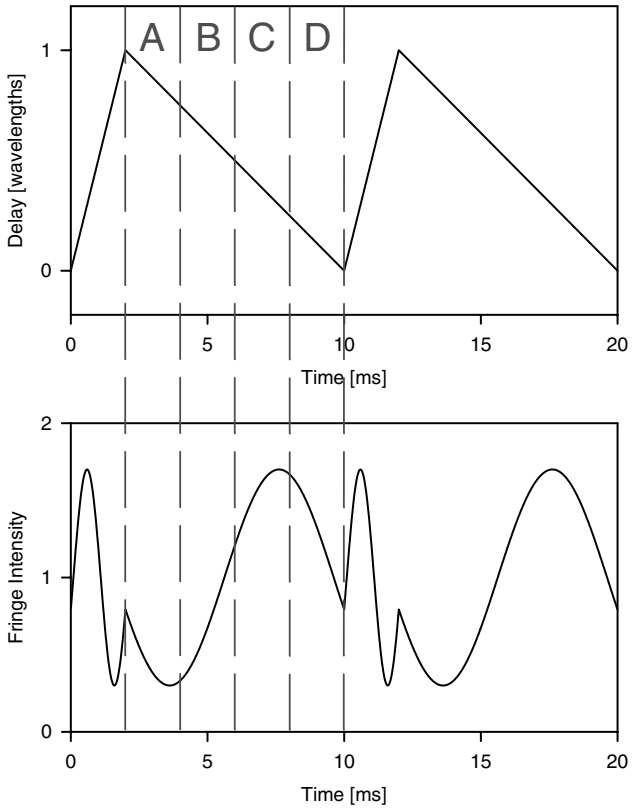

Fig. 43. ABCD fringe scanning scheme. The delay is modulated by a sawtooth pattern with amplitude one wavelength *(top)*. The time for the fringe scan is divided in four intervals of equal length, A, B, C, and D. The detector readout is synchronized with these four intervals, and the intensity integrated during each interval is measured. The fringe amplitude and phase can be derived from the four bin counts (see (149) and (150))

where $X = C - A$ and $Y = D - B$ are the real and imaginary parts of the visibility, $N = A + B + C + D$ is the total number of photons counted, and $N_{\rm dark}$ is the background count rate determined separately on blank sky. This estimator for V^2 is not biased by photon noise (Shao et al. 1988). The visibility phase is estimated from

$$\phi = \arctan\left(\frac{Y}{X}\right) - \frac{\pi}{4}\,. \tag{150}$$

For delay modulation and synchronous detection of the photon count rate in n bins per wavelength of modulation, (149) can be generalized to

$$V^2 = \left(\frac{4\pi^2}{n^2 \sin^2(\pi/n)}\right) \frac{\left\langle X^2 + Y^2 - \sigma_N^2 \right\rangle}{\left\langle N - N_{\rm dark} \right\rangle^2}\,, \tag{151}$$

where X and Y are again the real and imaginary parts of the visibility constructed from the bin counts, N the total number of counts in all bins, σ_N^2 the variance of N due to noise (Benson et al. 1998).

Coherent and Incoherent Visibility Integration

The data are averaged using a combination of coherent and incoherent integrations.[32] By choosing a coherent integration time T, an observation of total duration $M \cdot T$ is divided into M intervals, which are averaged incoherently. The variance of the V^2-estimator (149) is then given by

$$\sigma^2 = \frac{\pi^4}{4MN^2} + \frac{\pi^2 V^2}{MN} \ , \tag{152}$$

where N is the number of photons detected per coherent integration time (Colavita 1985). The signal-to-noise-ratio of V^2 is therefore:

$$\mathrm{SNR}(V^2) = \frac{2}{\pi^2} \cdot \frac{\sqrt{M} N V^2}{\sqrt{1 + \frac{4}{\pi^2} N V^2}} \, . \tag{153}$$

If $NV^2 \gg 1$, the second term in (152) dominates, and the variance depends only on the total number of photons detected, MN. If, however, $NV^2 \ll 1$, the first term is the dominant one, and the variance for a given total duration of the observation (i.e., constant total number of photons MN) decreases with increasing coherent integration time, $\sigma^2 \propto N^{-1} \propto T^{-1}$; this implies that the signal-to-noise ratio of V^2 is $\propto T^{1/2}$ (for constant $M \cdot T$). We will call the two cases the "photon-rich" and "photon-starved" regimes, although NV^2, and not N, is the critical quantity.

The important results captured in (152) and (153) have a simple intuitive interpretation. If the coherent integration time is sufficiently long, we get a good estimate of the amplitude *and phase* of the complex visibility. We can then stop the coherent integration, write out V^2 for a data sample, and average over these samples later without losing sensitivity. This is the photon-rich regime. If we are forced to stop the coherent integration (e.g., because of variations in the atmospheric or instrumental phase) before we get a meaningful phase measurement, we can still estimate V^2 for each data sample, but averaging over these estimates gives the poorer signal-to-noise characteristic of the photon-starved regime.

[32] Coherent integration means that we sort each photon arriving during the integration time in one of the bins A, B, C, D, and use (149) to get an estimate of V^2. Incoherent integration means that we average over many estimates of V^2. The intuitive meaning is that the coherent integration is used to estimate both amplitude *and phase* of the visibility, whereas the incoherent integration averages over the *modulus* of the visibility.

While these considerations show that it is advantageous to choose T large enough to get into the photon-rich regime, values larger than a fraction of the atmospheric coherence time will lead to serious phase changes and therefore to unacceptable degradation of the visibility. In the MkIII "standard" data reduction for measurements of stellar diameters and binary stars, $T = 4\,\mathrm{ms}$ is adopted, which gives a coherence loss of a few percent for seeing conditions typical for Mt. Wilson.

Visibility Calibration

The different interferometers use somewhat different observing strategies and calibration procedures, but they are all based on measurements of stars with known diameters. In the case of the MkIII, several calibrator stars were normally included in the observing list for each night. They were used to determine the "system visibility" V^2_{sys}, i.e., the value of V^2 observed for unresolved stars, as a function of seeing, zenith angle, time, and angle of incidence on the siderostat mirrors. For the seeing calibration, a seeing index S is calculated for each observation from the residual delay that the fringe tracker was unable to remove (Mozurkewich et al. 1991). After removing the relatively strong dependence of V^2 on S, calibration with respect to the other variables normally gave only a slight further improvement. (This situation is changed for phase-referenced data, where an additional strong decrease of V^2 with zenith angle has to be taken into account, see Sect. 8.6). The raw values of V^2 determined from (149) are then divided by V^2_{sys} to obtain calibrated data V^2_{cal} for further analysis. Both the internal noise, with contributions from photon noise and from short-term fluctuations, and the calibration uncertainty contribute to the error of V^2_{cal}. The two terms can be added in quadrature to obtain formal error bars.

Fringe Tracking Trade-Offs

The optimization of the fringe-tracking sensitivity is critically important for every interferometer. An important consideration in this regard concerns the trade-off between white-light fringe tracking and techniques based on dispersed fringes (le Poole and Quirrenbach 2002). White-light fringe tracking gives the highest sensitivity, because it allows for the simplest optical design (and thus for the highest optical throughput), and because it uses all available light with the smallest possible number of pixel-reads. On the other hand, white-light fringe tracking is also most sensitive to fringe mis-identification and fringe jumps. The best way to overcome this is by considering spectrally dispersed fringes. A compromise between these two conflicting requirements is to send part of the light to the white-light tracker, and part to a dispersed fringe sensor to locate the central fringe at a lower sampling rate. It is obvious that this works well only if fringe jumps are sufficiently rare, which in turn sets a limit on the minimum necessary signal-to-noise ratio and thus on

the required brightness of the reference star. Detectors with intrinsic spectral resolution would be very advantageous to solve this dilemma; they could combine the advantages of white-light fringe tracking (efficiency and sensitivity) with a simultaneous capability to measure the group delay, and thus to ensure proper identification of the central fringe. Detector types that are currently under development include Superconducting Tunneling Junctions (STJs) and Superconducting Edge Sensors (Perryman and Peacock 2000; Romani et al. 1999; Bruijn et al. 2000). A fringe tracker based on an adaptation of these detector types for near-infrared observations with a spectral resolving power of $R \approx 20$ would produce an almost optimum solution for fringe tracking on faint sources.

Single-Mode Fibers and Modal Filtering

Single-mode optical fibers can be used for many of the functions required in an optical interferometer: beam transport, beam combination (in **X** couplers), modulation of the optical path difference (by physically stretching a fiber, Shaklan 1990), polarization control, and modal filtering. The last capability is particularly attractive; single-mode fibers can eliminate the decrease in fringe visibility caused by atmospheric turbulence and thus alleviate the calibration difficulties of ground-based interferometers. The coupling efficiency into single-mode fibers depends on the wavefront shape; it is roughly equal to the Strehl ratio of the input beam (Shaklan and Roddier 1988). The output of the fiber is a perfectly flat wavefront. Introducing a single-mode fiber in each of the interferometer arms thus converts atmospheric wavefront aberrations into intensity fluctuations. Splitting off some of the light from each telescope before beam combination in a **Y** coupler allows monitoring of I_1 and I_2 (Coudé du Foresto et al. 1997). The corrected interferogram

$$I_{\rm cor} \equiv \frac{I_{\rm out} - I_1 - I_2}{2\sqrt{I_1 I_2}} = \Gamma(u, 0) \tag{154}$$

is then independent of atmospheric wavefront degradation. By definition, the étendue of a single-mode fiber is λ^2, i.e., the field-of-view is limited to one Airy disk.

Phase-Referenced Visibility Averaging

The wide-band tracking channel in the MkIII Interferometer provides a phase reference, which can be used to extend the coherent integration time T beyond the limit imposed by the atmospheric turbulence. This method provides a means of obtaining substantially better signal-to-noise in the photon-starved regime, or even to make a transition into the photon-rich regime. The phase-referenced quantities X_r, Y_r, V_r, and ϕ_r are defined by

$$X_r + iY_r = V_r \, {\rm e}^{i\phi_r} = V_s \, {\rm e}^{i(\phi_s - \frac{\lambda_t}{\lambda_s}\phi_t)} \ , \tag{155}$$

where λ_s, V_s, ϕ_s are the wavelength, visibility, and phase in the signal channel, and λ_t, ϕ_t the wavelength and phase in the tracking channel. In practice, V_r^2 is computed from (149) using X_r and Y_r instead of X and Y; this procedure retains the advantage of using an unbiased estimator.

Equation (155) assumes that the atmospheric phase at λ_s is given by $(\lambda_t/\lambda_s)\phi_t$. If this were the case exactly, there would be no coherence losses, and the integration time could be arbitrarily long. A number of systematic effects (discussed in more detail in Sect. 8.6, see also Quirrenbach et al. 1994) can lead to a decorrelation of the phases between the signal and tracking channels, however. They introduce additional phase noise, which reduces the system visibility and limits the maximum integration time. The dependence of the system visibility on seeing and zenith angle is also made steeper, which increases the uncertainty of the calibration. In practice, therefore, phase-referenced averaging involves trading off some calibration accuracy for the gain in signal-to-noise.

8.6 Phase Decorrelation Mechanisms

Phase Errors and Coherence Losses

We will now discuss a number of mechanisms that lead to phase errors and therefore to coherence losses and to a reduction of the phase-referenced visibility. These effects can be broadly divided into two classes, namely those mechanisms that are due to errors in the determination of the phase in the reference channel, and those that are due to differential atmospheric propagation effects. While some of the former processes are instrument-dependent and can be reduced (or even avoided) by improved interferometer and fringe-detector designs, the latter class sets fundamental limits to the application of phase-referencing methods from the ground. We will again use phase-referenced visibility averaging with the MkIII Interferometer to give some specific numerical examples (see also Quirrenbach et al. 1994).

If the variance of the referenced phase ϕ_r associated with a decorrelation mechanism is $\sigma^2_{\phi,r}$, it will reduce V_r^2 by a factor η, which can be computed from

$$\eta = e^{-\sigma^2_{\phi,r}} \,. \tag{156}$$

For assessing the individual mechanisms, it is not only important to compare the numerical values of the associated phase variances, but also to note their dependencies on observing conditions (e.g. seeing, zenith angle) and particularly on stellar parameters (e.g. colors). While the standard calibration procedure will correct for a uniform reduction of V^2, and to some extent for variations with observing conditions, effects that differ from star to star can introduce systematic errors that are difficult to detect. A priori limits on these effects are therefore necessary for practical applications of phase-referenced visibility averaging.

Photon Noise in the Tracking Channel

The finite number of photons detected during each coherent integration interval (4 ms in the MkIII case) sets a fundamental limit to the precision of the reference phase determination. The variance of ϕ_r due to photon noise in the tracking channel is

$$\sigma^2_{\phi,r} = \left(\frac{\lambda_t}{\lambda_s}\right)^2 \sigma^2_{\phi,t,\mathrm{phot}} = \left(\frac{\lambda_t}{\lambda_s}\right)^2 \cdot \frac{2}{N_t V_t^2} \ , \tag{157}$$

where N_t and V_t are the number of the photons counted and the visibility in the tracking channel. $\sigma^2_{\phi,r}$ depends on the brightness and color of the star, and even on the baseline length (through V_t^2). However, for the fringe tracker to work reliably under average seeing conditions, $N_t V_t^2 \approx 70$ is needed for the 4 ms sampling interval, giving $\eta \approx 0.98$ for $\lambda_t = 700\,\mathrm{nm}$, $\lambda_s = 800\,\mathrm{nm}$, and $\eta \approx 0.95$ for $\lambda_t = 700\,\mathrm{nm}$, $\lambda_s = 500\,\mathrm{nm}$. Thus the visibility reduction is slight even for stars that are close to the sensitivity limit of the fringe tracker, and negligible for stars that are substantially brighter. It is also possible to introduce the signal-to-noise in the tracking channel as an additional independent variable in the calibration process, if very high accuracy is required.

Color and Visibility Dependence of the Effective Tracking Wavelength

To achieve high sensitivity (and to keep the errors due to photon noise small), the bandpass in the fringe-tracking channel should be made as wide as possible. The effective wavelength to be used in (155) is then given by

$$\lambda_t = \frac{\int d\lambda\, \lambda\, W_t(\lambda) N(\lambda) V(\lambda)}{\int d\lambda\, W_t(\lambda) N(\lambda) V(\lambda)} \ , \tag{158}$$

where $N(\lambda)$ is the number of photons emitted by a star as a function of wavelength, $V(\lambda)$ the visibility, and $W_t(\lambda)$ the combined response of atmosphere, instrument, and detector. If the wavelength used in (155) differs from the true effective wavelength by $\delta\lambda_t$, the resultant variance of the reference phase is

$$\sigma^2_{\phi,r} = \left(\frac{\delta\lambda_t}{\lambda_s}\right)^2 \cdot \left\langle \phi_t^2 \right\rangle \ . \tag{159}$$

As evident from (158), the true effective wavelength depends on stellar colors and diameters, and on the baseline length. If for simplicity one uses $\lambda_t = 700\,\mathrm{nm}$ for all stars in the data reduction, $\delta\lambda_t \lesssim 25\,\mathrm{nm}$ for the parameters of the MkIII Interferometer. With the additional assumption that the residual atmospheric phase r.m.s. not tracked by the fringe tracker $\sqrt{\langle\phi_t^2\rangle} \lesssim 2\,\mathrm{rad}$, $\eta \gtrsim 0.99$ is derived from (159).

Stroke Mismatch

In pathlength modulation schemes like that used by the MkIII, any difference between the stroke of the 500 Hz pathlength modulation and the wavelength λ will also lead to errors in the phase estimation, since then the bins A, B, C, and D do not correspond exactly to $\lambda/4$. (This correspondence is assumed implicitly in (150).) For each channel, the gating of the electronic counters for A, B, C, and D has to be set by the on-line control system to match one quarter of the nominal wavelength. In this way, an effective stroke s is created for each channel. Defining

$$\varepsilon = \frac{2\pi}{\lambda} \cdot (s - \lambda) \quad \text{and} \quad \delta = \frac{\cos \varepsilon/4}{1 + \sin \varepsilon/4} \ , \tag{160}$$

it has been shown by Colavita (1985) that

$$\tan \phi_{\text{est}} = \delta \cdot \tan \phi_{\text{true}} \ , \tag{161}$$

where ϕ_{est} is the phase estimated from (150), and ϕ_{true} is the true phase. For a complete treatment of the effect of the stroke mismatch, these equations have to be integrated over λ, with a suitable weighting function representing the bandpass of the tracking channel. To first order, however, it can be assumed that the phase error is given by (160) and (161), evaluated at $\lambda = \lambda_t$. For $s_t - \lambda_t \leq 25\,\text{nm}$, a phase error $\phi_{\text{est}} - \phi_{\text{true}} \leq 2^\circ$ is then obtained. Errors of this order can be safely ignored for most visibility averaging applications, but may be important for phase-referenced imaging and spectroscopy.

Fringe Jumps

An ideal fringe tracker would follow the atmospheric pathlength fluctuations to a fraction of λ_t, and ϕ_t would always be well within the interval $(-\pi, \pi]$. In practice, however, temporary excursions from the central fringe that are larger than $\lambda/2$ may occur, and the phase has to be "unwrapped" by the phase-referencing algorithm. This is done by imposing the requirement that the phase in successive data segments (4 ms intervals for the MkIII) should be continuous. While this process normally works well, occasional misidentifications are possible. It is obvious from (155) that a 360° error in ϕ_t will lead to a phase jump in ϕ_r.

If the average number of these jumps during the coherent integration time T is small, the coherence loss is not dramatic. This requirement sets an upper limit to T. Since the probability of unwrapping errors depends only on the seeing and on the signal-to-noise in the tracking channel, it can be accounted for in the calibration procedure. In a series of tests with the MkIII, it turned out that the degradation of the phase-referenced visibility V_r due to fringe jumps was not serious for integration times up to 2 s, for average seeing conditions on Mt. Wilson.

Dispersion

While (155) assumes that the atmospheric pathlength fluctuations are independent of wavelength, they are actually larger in the blue spectral range than in the red, because of dispersion. The two-color dispersion coefficient D is defined by

$$D = \frac{n(\lambda_t) - 1}{n(\lambda_s) - n(\lambda_t)} \ , \tag{162}$$

where $n(\lambda)$ is the refractive index of air at λ. Typical values for $\lambda_t = 700\,\mathrm{nm}$ and $\lambda_s = 450$, 500, 550, and 800 nm are $D = 59$, 87, 137, and -364, respectively. If the total "unwrapped" phase in the tracking channel is denoted Φ_t, a phase error $(\lambda_t/\lambda_s)(\Phi_t/D)$ is introduced by the dispersion. Since the largest phase excursions occur on long time scales, this sets a limit to the coherence time. For Kolmogorov turbulence, the coherence time $t_{0,r}$ of ϕ_r is given by

$$t_{0,r} = |D|^{6/5}\, t_{0,s} \ , \tag{163}$$

where $t_{0,s}$ is the atmospheric coherence time in the data channel (Colavita 1994). Under average conditions on Mt. Wilson, $t_{0,s}$ is of order 6 to 8 ms at 500 nm. For integration times up to about 2 s, the coherence losses due to dispersion are therefore tolerable for visibility averaging, and they can be taken into account by the calibration procedure.

It is obviously possible to deal with dispersion explicitly by using

$$\tilde{\phi_r} = \phi_s - \frac{\lambda_t}{\lambda_s}\phi_t - \frac{\lambda_t}{\lambda_s} \cdot \frac{\Phi_t}{D} \tag{164}$$

instead of ϕ_r as defined in (155). While this approach can reduce the phase errors by a factor $\sim$10, a residual effect due to water vapor fluctuations remains, because their dispersion is different from the values applicable to dry air.

Anisoplanatism

If the reference phase is measured on a star at an angular separation θ from the target object, there will be some decorrelation because the light from the two sources passes through different turbulence cells, as discussed in Sect. 7.7. In interferometric applications, the independent contributions from the two arms of the interferometer have to be taken into account. Under the assumption of a Kolmogorov turbulence spectrum, the interferometric isoplanatic angle therefore is

$$\theta_i = 2^{-3/5}\theta_0 = \left[5.82\, k^2 (\sec z)^{8/3} \int_0^\infty dh\, C_N^2(h) h^{5/3}\right]^{-3/5} \ , \tag{165}$$

where $k = 2\pi/\lambda$ is the wavenumber (assumed here to be equal for the target and reference channels), and z the zenith angle. While this expression holds for

small apertures, a somewhat more optimistic estimate is obtained for larger apertures (Colavita 1994). Typical values for θ_i are of order a few arcseconds, much larger than the interferometric field of view of a Michelson interferometer. In applications where the reference phase is measured on the object of scientific interest itself, anisoplanatism does not occur at all. However, it is the most severe limitation for dual-star interferometry (see Sect. 9.5). Because of the usual $\theta_i \propto \lambda^{6/5}$ scaling finding reference stars for dual-star interferometry is much easier at longer wavelengths.

Differential Refraction

An effect somewhat similar to anisoplanatism occurs even when the angular separation between the target and the reference is zero. If $\lambda_s \neq \lambda_t$, the beams at the two wavelengths follow different paths through the atmosphere at non-zero zenith angles, due to differential refraction. For a Kolmogorov turbulence spectrum, the corresponding phase variance is

$$\sigma^2_{\phi,r} = 5.82\, k_s^2 \left[\frac{h_0\,(n(\lambda_t)-1)\,\mathrm{e}^{-h_1/h_0}}{D}\right]^{5/3} \tan^{5/3} z\, \sec^{8/3} z \times \int_0^\infty dh\, C_N^2(h) \left(1-\mathrm{e}^{-h/h_0}\right)^{5/3} , \quad (166)$$

where $k_s = 2\pi/\lambda_s$ is the wavenumber in the signal channel, $n(\lambda_t)$ is the atmospheric index of refraction at λ_t, D is the atmospheric dispersion between λ_s and λ_t defined by (162), h_0 is the scale height of the atmospheric density, h_1 is the elevation of the observatory site above sea level, z is the zenith angle, and $C_N^2(h)$ is the refractive index structure constant. Again, this estimate might be somewhat pessimistic, since averaging over the aperture has not been taken into account.

The phase variance due to differential refraction depends very strongly on z; while it is negligible close to the zenith, it is the dominant decorrelation mechanism at intermediate to large zenith angles for the parameters of the MkIII phase-referenced visibility averaging experiments. From (166) it is obvious that differential refraction – like anisoplanatism – is more strongly affected by high-altitude turbulence than by disturbances close to the ground. This is expected, since the beams from target and reference coincide at the telescope aperture; their separation increases with height when they are traced back through the atmosphere. To carry out quantitative calculations of differential refraction, it is therefore necessary to know the turbulence profile; in the absence of better measurements we use the model for the atmospheric turbulence as a function of height h (in m) by Hufnagel (1974),

$$C_N^2(h) = 2.7 \cdot \left(2.2 \cdot 10^{-53} h^{10} \mathrm{e}^{-h/1000} + 10^{-16} \mathrm{e}^{-h/1500}\right) . \quad (167)$$

Figure 45 shows the reduction of V_r^2 derived from a numerical integration of (166), with the Hufnagel turbulence profile. The values $h_0 = 8\,300\,\mathrm{m}$,

$h_1 = 1\,700$ m (applicable to Mt. Wilson), $\lambda_t = 700$ nm, and $\lambda_s = 450$, 500, 550, and 800 nm were used. This figure demonstrates that differential refraction leads to a much steeper dependence of the system visibility with zenith angle in the phase-referenced data than in incoherent averages. This effect is particularly important in the blue spectral range, where the dispersion is large (small values of D). Differential refraction therefore restricts the application of phase-referenced visibility averaging to moderate zenith angles, depending on the wavelength λ_s and on the seeing.

The top panel in Fig. 44 shows the MkIII system visibility for two nights (July 29 and 31, 1989) as a function of zenith angle z, for the data integrated

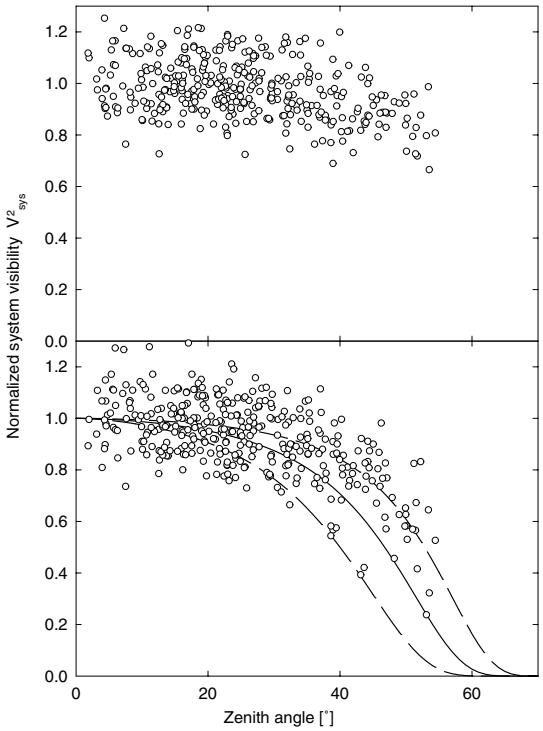

Fig. 44. *Top panel*: Observed V^2 divided by an estimate $V^2_{\rm est}$ from photometric data, for 16 stars at 500 nm. The data are plotted as a function of zenith angle z. Each night was normalized to 1 at $z = 0$. Each measurement corresponds to one 75 s observation. The standard data reduction procedure was used, which averages the 4 ms samples incoherently. *Bottom panel*: The same data as in the top panel, but processed with the phase-referenced averaging algorithm. The coherent (phase-referenced) integration time is 1,024 ms. The *solid curve* is the visibility reduction due to differential refraction predicted by the Hufnagel model atmosphere; the *dashed curves* correspond to atmospheres that have 0.5 and 2 times the C_n^2 of the Hufnagel model at all heights

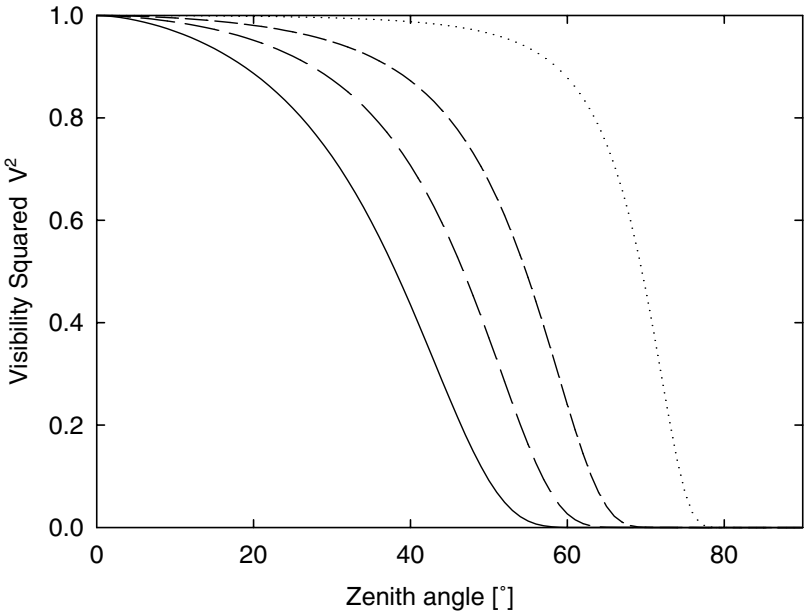

Fig. 45. Reduction of V^2 due to differential refraction as a function of zenith angle z, predicted from a Hufnagel (1974) model atmosphere. The reference wavelength $\lambda_t = 700\,\mathrm{nm}$; the wavelength in the data channels $\lambda_s = 450$, 500, 550, and 800 nm

incoherently with the standard method; it has been normalized to $V^2_{\mathrm{sys}} = 1$ at $z = 0$. It is obvious that V^2_{sys} varies only slightly with z; this variation is mostly due to the degradation of the seeing for longer pathlengths through the atmosphere. The bottom panel shows the same data, but processed with the phase-referencing algorithm, using an integration time of 1,024 ms. A strong reduction of the system visibility is now apparent at $z \gtrsim 40°$. The solid line indicates the visibility reduction due to differential refraction predicted by the Hufnagel (1974) atmosphere model. The qualitative agreement between the observations and this model demonstrates that differential refraction is indeed the dominant reason for coherence losses at intermediate to large zenith angles.

Diffraction

Finally, if $\lambda_s \neq \lambda_t$, there will be some decorrelation because of diffraction. The phase variance due to diffraction is related to the intensity scintillation variance $\sigma^2_{\ln I}$ by

$$\sigma^2_{\phi,r} = G(\lambda_t/\lambda_s)\,\sigma^2_{\ln I}(\lambda_t)\ , \tag{168}$$

with a function G(r), which can be approximated by

$$G(r) \approx \left(r^{1/2}\,(r-1)\,/2\right)^{4/3} \tag{169}$$

for $1 \leq r \lesssim 1.5$ (Colavita 1994). Observed values for $\sigma^2_{\ln I}$ on Mt. Wilson range from 0.005 to 0.05. The larger of these values gives $\sigma^2_{\phi,r} = 0.0073$, or $\eta = 0.99$ for $\lambda_t = 700\,\mathrm{nm}$, $\lambda_s = 500\,\mathrm{nm}$. Since all stars are affected equally, the calibration procedure takes into account the small coherence loss due to diffraction.

9 Astrometry with Interferometry

The principle of planet detection with astrometry is similar to that behind the Doppler technique: one infers the presence of a planet from the motion of its parent star around the common center of gravity. In the case of astrometry one observes the two components of this motion in the plane of the sky; this gives sufficient information to solve for the orbital elements without $\sin i$ ambiguity. Astrometry also has advantages for a number of specific questions, because this method is applicable to all types of stars, and more sensitive to planets with larger orbital semi-major axes. Interferometric techniques hold the promise to push astrometric precision well beyond the current state of the art; these advances are needed for planet detection. In this chapter we will therefore discuss the technical foundations of interferometric astrometry, and examine its potential for the detection and characterization of extrasolar planets.

9.1 Astrometric Signature of Low-Mass Companions

From simple geometry and Kepler's Third Law (17) it follows immediately that the astrometric signal θ of a planet with mass m_p orbiting a star with mass m_* at a distance d in a circular orbit of radius a is given by

$$\begin{aligned}\theta &= \frac{m_p}{m_*}\,\frac{a}{d} = \left(\frac{G}{4\pi^2}\right)^{1/3} \frac{m_p}{m_*^{2/3}}\,\frac{P^{2/3}}{d} \\ &= 3\,\mu\mathrm{as}\cdot\frac{m_p}{M_\oplus}\cdot\left(\frac{m_*}{M_\odot}\right)^{-2/3}\left(\frac{P}{\mathrm{yr}}\right)^{2/3}\left(\frac{d}{\mathrm{pc}}\right)^{-1}\;; \end{aligned} \tag{170}$$

This signature is shown in Table 11, for five sample planets (analogues to Earth, Jupiter, Saturn, Uranus, and a "Hot Jupiter" with $m_p = 1\,M_{\mathrm{jup}}$ and $P = 4\,\mathrm{days}$) orbiting a $1\,M_\odot$ star. From this table, the main strengths and difficulties of astrometric planet detection are readily apparent:

- The astrometric signature θ is small compared to the precision of "standard" astrometric techniques ($\lesssim 1\,\mathrm{mas}$).
- The difficulty of detecting different types of planets varies greatly, with θ ranging from $\lesssim 1\,\mu\mathrm{as}$ to $\sim 1\,\mathrm{mas}$.
- The sensitivity of astrometry for a given type of planet drops linearly with d (unlike the radial-velocity technique).

Table 11. Astrometric signature from different planet / parent star combinations

planet	orbit [AU]	star	amplitude · d [μas · pc]
Earth	1	G2	3
Jupiter	5	G2	4,800
Uranus	20	G2	880
"hot" Jupiter	0.1	G2	96
Brown Dwarf	0.1	M5	7,500
Jupiter	5	A5	2,200
Jupiter	5	M5	24,000

Note that the numbers given in the last column have to be divided by the distance to give the astrometric signature. 15 $M_{\rm jup}$ was used for the mass of the brown dwarf

- The detection bias of astrometry with orbital radius is opposite to that of the radial-velocity method, favoring planets at larger separations from their parent stars.

It should be pointed out that for circular orbits the observed astrometric signal is an ellipse with semi-major axis θ independent of the orbital inclination; the mass of the planet can therefore be derived directly from (170) if the mass of the parent star is known. The situation is a bit more complicated for non-circular orbits, but even in that case can the orbital inclination be determined from the astrometric data with techniques analogous to those used for fitting orbits of visual binaries (e.g., Binnendijk 1960).

In Table 12, the numbers from Table 11 have been converted to the maximum distance at which the planet is detectable, for different assumptions about the astrometric precision.

9.2 Upper Mass Limits and Astrometric Detection of Gl 876 B

Looking at planets that are already known from radial-velocity surveys is an obvious first application of the astrometric technique, because of its ability to determine the planet's mass without $\sin i$ ambiguity. It is clear from that this is a challenging task with a precision of slightly better than a milliarcsecond, which is currently achievable with Hipparcos data or with the Hubble Fine Guidance Sensors. Observations of Gl 876 with the latter instrument resulted in the detection of the astrometric wobble due to its companion Gl 876 b, and thus mark the first secure astrometric detection of an extrasolar planet (Benedict et al. 2002). The parameters of this system are $\phi = 0.25 \pm 0.06$ mas, $i = 84° \pm 6°$, and $m_p = 1.9 \pm 0.5\, M_{\rm jup}$.

In many cases, interesting upper limits on the companion mass can be derived from astrometric observations even if the signature is below the detection

Table 12. Maximum distance to which different planets can be observed with a ground-based interferometer in astrometric mode, with either 50 μas or 10 μas precision

precision	planet	orbit [AU]	star	max. distance [pc]
50 μas	Jupiter	5	G2	48
	Uranus	20	G2	9
	Brown Dwarf	0.1	M5	75
	Jupiter	5	A5	22
	Jupiter	5	M5	240
10 μas	Jupiter	5	$1\,M_\odot$	240
	Uranus	20	G2	44
	Jupiter	5	A5	110
	10 Earths	1	G2	1.5
1 μas	Jupiter	0.1	G2	48
	Earth	1	G2	1.5

1 μas can probably be attained only from space. It is assumed that a $4\,\sigma$ peak-to-peak variation is required for secure detection of a planet

limit. A good example is the case of ι Dra b (Frink et al. 2002). The radial-velocity observations give $P = 536$ days, $e = 0.70$, and $m_p \sin i = 8.9\,M_{\rm jup}$ for this object. The non-detection in the Hipparcos data places a 3σ upper limit of $45\,M_{\rm jup}$ on the mass of ι Dra B, and thus firmly establishes its sub-stellar nature.

9.3 Astrometric Measurements with an Interferometer

Improving the astrometric precision by one to two orders of magnitude is required to make astrometry a truly powerful and versatile tool for planet detection. Dramatic progress is indeed expected from the development of interferometric techniques, which have the potential to achieve ∼10 μas from the ground, and ∼1 μas from space.

Principles of Interferometric Astrometry

Astrometric observations by interferometry are based on measurements of the delay $D = D_{\rm int} + (\lambda/2\pi)\phi$, where $D_{\rm int} = D_2 - D_1$ is the internal delay measured by a metrology system (see Fig. 42), and ϕ the observed fringe phase. Here ϕ has to be unwrapped, i.e., not restricted to the interval $[0, 2\pi)$. In other words, one has to determine which of the sinusoidal fringes was observed. This can,

for example, be done with dispersed-fringe techniques (see Sect. 8.5). D is related to the baseline $\vec{B}$ by

$$D = \vec{B} \cdot \hat{s} = B \cos\theta \ , \tag{171}$$

where $\hat{s}$ is a unit vector in the direction toward the star, and θ the angle between $\vec{B}$ and $\hat{s}$. Each data point is thus a one-dimensional measurement of the position of the star θ, provided that the length and direction of the baseline are accurately known. The second coordinate can be measured with a separate baseline at a roughly orthogonal orientation.

In a ground-based interferometer the endpoints of the baseline (i.e., the positions of the telescopes or siderostats) can be related to the solid ground of the site and thus tied to the Earth's rotation. One can either rely on the stability of the telescope mount, or, if greater precision is needed, use a truss of laser interferometers to monitor changes of the positions of the siderostat pivot points with respect to "optical anchors" attached directly to bedrock (Armstrong et al. 1998). Repeated observations of a set of stars throughout a night can then be used to determine the baseline vector $\vec{B}$ in the rotating reference frame of the Earth. With a sufficient number of observations for each star one obtains more observables (delays) than unknowns (stellar and baseline coordinates), so that one can solve for $\vec{B}$ from a set of over-constrained equations (Thompson et al. 1986).

In space, no convenient stable platform like the Earth is available. The Space Interferometry Mission (SIM, Sect. 9.6) therefore uses two additional interferometers that look at two stars to stabilize the spacecraft attitude; these guide interferometers are essentially extremely precise star trackers (Milman and Turyshev 2000). Since the spacecraft structure is not sufficiently stiff on the sub-μm scale, an "optical truss" is again formed by laser interferometers and used to monitor the exact position of all important optical elements, including the baseline length and the relative orientation of the main and guide interferometers.

Astrometric Precision

The photon noise limit for the precision σ of an astrometric measurement is given by the expression

$$\sigma = \frac{1}{\mathrm{SNR}} \cdot \frac{\lambda}{2\pi B} \, . \tag{172}$$

Since high signal-to-noise ratios can be obtained for bright stars, σ can be orders of magnitude smaller than the resolution λ/B of the interferometer. For example, the resolution of SIM ($B = 10\,\mathrm{m}$) is about 10 mas, but the astrometric precision of measurements with SIM should approach 1 μas.

To reach this photon noise limit, it is of course necessary to control all other statistical and systematic errors very precisely. Keeping track of all instrumental contributions to the final astrometric error is quite an arduous

task; it is usually accomplished with a formal error budget. This is essentially a tree in which error sources are organized in a hierarchical structure. Each box in this error budget can be further subdivided in sub-boxes down to the noise of individual metrology detectors, surface errors of individual mirrors, or vibration spectra of the mounts of individual optical elements. Once this error tree has been established, it is possible to perform "top-down" or "bottom-up" analyses of the error budget. In a top-down error budget, one starts with a desired measurement accuracy, and divides the allowable error to individual systems, sub-systems, and so on to derive performance specifications for all components in the instrument. In a bottom-up analysis, one starts with known data or manufacturers' specifications for all components to arrive at a prediction of the final system performance. In the end, hopefully, both analyses agree, and one ends up with a plan for building an instrument that will perform to specifications. During the commissioning and operation of the instrument, the error tree is an important tool for debugging and implementation of improvements.

Radio and Millimeter Interferometry

The application of interferometric methods to astrometry is not limited to the visible and infrared wavelength ranges, of course. The technique of very long baseline interferometry (VLBI), in which telescopes on different continents are coupled coherently, has indeed produced astrometric data of such quality that the International Celestial Reference Frame has been based on radio sources. VLBI can also reach sub-milliarcsecond precision for the measurement of the photocenter position of radio stars (Lestrade 2000a). Unfortunately, the relation between the center of mass and the region from which the radio emission emanates, is frequently poorly understood, which makes the interpretation of the data difficult. Nevertheless, pilot VLBI observations of a number of dMe stars have been performed with the ultimate goal to search for planets around them (Guirado et al. 2002). The sub-millimeter array ALMA currently under construction in the Atacama desert in Chile will have sufficient sensitivity to detect the photospheres of nearby stars at 345 GHz and could thus also search for the astrometric signature of planets around them (Lestrade 2000b). However, the intrinsic faintness of stellar photospheres far in the Rayleigh-Jeans regime of their thermal emission limits the scope of such projects in comparison to the more promising astrometric programs at visible and near-infrared wavelengths.

9.4 Atmospheric Limitations of Astrometry

The Narrow-Angle Atmospheric Regime

The Earth's atmosphere imposes serious limitations on the precision that can be achieved with astrometric measurements from the ground. An obvious difficulty is atmospheric refraction (Gubler and Tytler 1998), but the effects

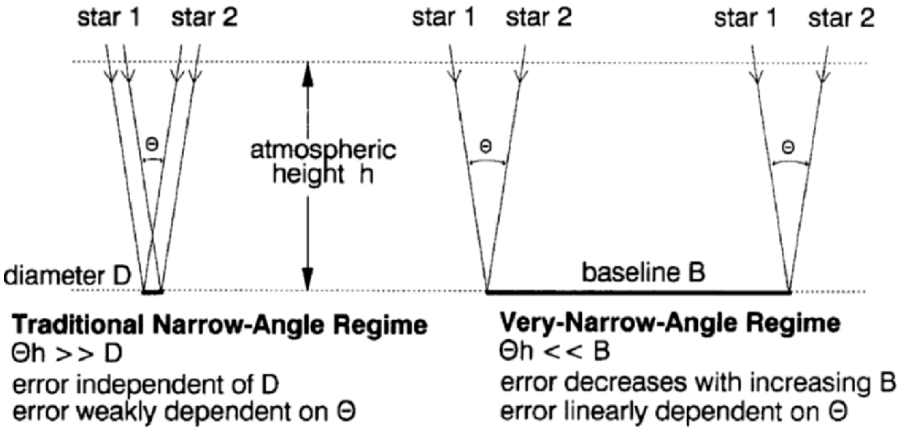

Fig. 46. Schematic of the atmospheric paths for narrow-angle astrometry with short and long baselines B (or telescope diameter D). In each panel, rays from two stars to the two telescopes at the ends of the baseline are shown. The atmosphere is represented by a single layer at height h. In this layer, the two rays originating from the same star to the two telescopes are separated by B; the two rays originating from the two stars to the same telescope are separated by θh. In the left panel the baseline B is short, $\theta h \gg B$, in the right panel the baseline is long, $\theta h \ll B$. From Shao and Colavita (1992)

of seeing are even more disturbing. In Sect. 7 we have discussed the blurring of optical images due to atmospheric turbulence. The first-order terms (frequently referred to as *tip* and *tilt*) of this blurring are global wavefront gradients, which correspond to a motion of the centroid of the stellar light in the two coordinates. Because most of the power of atmospheric turbulence is in these low-order modes (e.g. Hardy 1998), the amplitude of this image motion is similar to the width of the stellar images, i.e., $\approx \lambda/r_0 \approx 0\farcs5...1''$ (Sect. 7.4). One can obviously reduce this error by taking many exposures and thus averaging over many independent realizations of the atmospheric turbulence, but achieving a precision of a milliarcsecond or even better in this way is clearly not possible.

It helps, however, to make differential measurements over small angles on the sky, i.e., to measure the position of the target star with respect to that of a nearby reference (see Fig. 46). If the reference star is sufficiently close on the sky, the rays from the two stars are affected in almost the same way by the atmospheric turbulence, and the error in the relative position between the two stars is greatly reduced. From Fig. 46 it should be intuitively clear that the length of the baseline[33] also plays an important role.

[33] An analogous analysis can also be carried out for single telescopes, where the telescope diameter plays the role of the baseline length. It is indeed possible to perform precise narrow-angle astrometry with large telescopes (Pravdo and Shaklan 1996).

If the baseline is short ($B \ll \theta h$, where θ is the angle between the two stars, and h the effective height of the atmospheric turbulence, see the left panel of Fig. 46), all rays from star 1 pass close to each other through the atmosphere, and all rays from star 2 pass close to each other; the separation between the two ray bundles is large in comparison. This means that a localized patch of atmospheric turbulence could affect all rays from star 1, but leave those from star 2 unaffected. Therefore the image motions of the two stars are only weakly correlated, and the astrometric error is independent of the baseline length. In contrast, if the baseline is long ($B \gg \theta h$, right panel of Fig. 46), the rays from both stars to telescope 1 pass relatively close to each other, and those to telescope 2 pass close to each other. A localized patch of turbulence will thus tend to affect rays from star 1 and 2 in nearly the same way, which leads to a stronger correlation of the image motions, and therefore to an astrometric error that decreases with increasing B.

With a calculation similar to those demonstrated in Sect. 7 it can be shown that the variance σ_θ^2 of measurements of the angle θ is given by (Shao and Colavita 1992)

$$\sigma_\theta^2 \approx \frac{16\pi^2}{B^2 t} \int_0^\infty dh\, v^{-1}(h) \int_0^\infty d\kappa\, \Phi(\kappa, h)\, [1 - \cos(B\kappa)] \cdot [1 - \cos(\theta h \kappa)] \ , \quad (173)$$

provided that the integration time $t \gg \max(B, \theta h)/v$. Here $v(h)$ is the wind speed at altitude h, and $\Phi(\kappa, h)$ denotes the three-dimensional spatial power spectrum of the refractive index (see Footnote 28 on p. 132). It may at first seem surprising that stronger winds should give a smaller measurement error, but within the frozen-turbulence picture (see Sect. 7.5) a higher wind speed means that one averages faster over independent realizations of the stochastic refractive index fluctuations. Inserting a Kolmogorov power spectrum in (173) one obtains the two limiting cases

$$\sigma_\theta^2 \approx \begin{cases} 5.25\, B^{-4/3} \theta^2 t^{-1} \int_0^\infty dh\, C_N^2(h) h^2 v^{-1}(h) & \text{for } \theta \ll B/h, \quad t \gg B/v \\ 5.25\, \theta^{2/3} t^{-1} \int_0^\infty dh\, C_N^2(h) h^{2/3} v^{-1}(h) & \text{for } \theta \gg B/h, \quad t \gg \theta h/v \end{cases} \quad (174)$$

for long and short baselines, respectively. In particular we see that for sufficiently small angles θ the important scaling relations $\sigma_\theta \propto \theta$ and $\sigma_\theta \propto B^{-2/3}$ hold for the astrometric error σ_θ. This error is plotted as a function of θ for three different baseline lengths in Fig. 47; the knees in the curves mark the transition between the two limiting cases in (174). We see that for a good site such as Mauna Kea astrometric measurements with a precision of $\sim$10 μas are possible over angles of $\sim 10''$. It is also apparent from the factor h^2 under the integral in this equation that the astrometric error is dominated by the turbulence at high altitudes. The low level of high-altitude turbulence at the South Pole would therefore make an astrometric interferometer at a site on the high Antarctic plateau an attractive possibility (Lloyd et al. 2002).

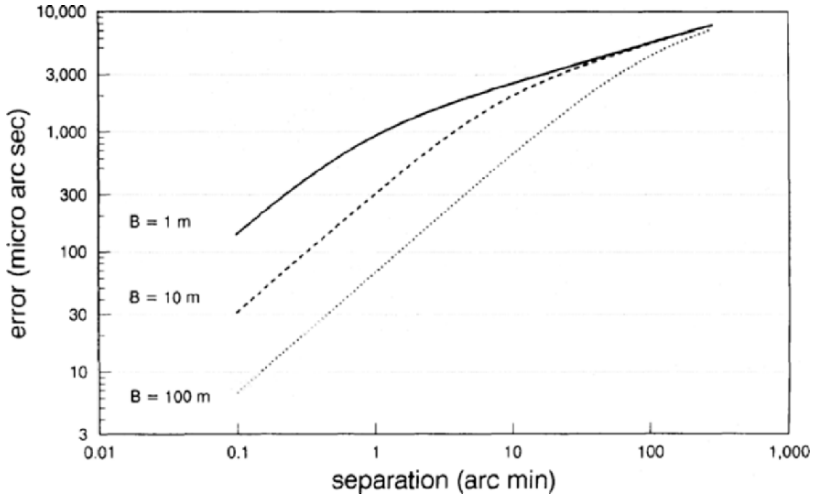

Fig. 47. Error of relative astrometric measurements for different values of the baseline length B, using a turbulence profile as measured for Mauna Kea, and assuming an integration time of 1 h. From Shao and Colavita (1992)

Influence of the Outer Scale of Atmospheric Turbulence

The calculation leading to (174), which is based on a Kolmogorov power law for the turbulence spectrum, actually gives somewhat pessimistic results. The reason is that in the Kolmogorov theory there must be an outer scale, beyond which the power in the turbulence spectrum flattens out (see Sect. 7.1). The mathematical treatment of this regime is much more involved than that of the inertial range. Two different parameterizations are in common use, namely the von Karman spectrum

$$\Phi(|\vec{\kappa}|) \propto \left(\kappa^2 + L_0^{-2}\right)^{-11/6} \tag{175}$$

and the Greenwood–Tarazano prescription

$$\Phi(|\vec{\kappa}|) \propto \left(\kappa^2 + \kappa/L_0\right)^{-11/6} . \tag{176}$$

It is obvious that in the limit of small spatial scales (large κ) both functional forms asymptotically approach the Kolmogorov spectrum $\Phi(|\vec{\kappa}|) \propto \kappa^{-11/3}$ (see Footnote 28 on p. 132).

Reliable measurements of the outer scale are very sparse, so that little is known about numerical values for L_0, much less about temporal variability of L_0 or which of the functional forms (175) or (176) is to be preferred (see Quirrenbach 2002b and references therein). Plausible typical values are in the range $L_0 \approx 10 \ldots 100\,\mathrm{m}$, but the uncertainty is very large. The best one can therefore currently do is calculate the astrometric errors for a range of outer scales – and base the design of astrometric instruments on the more conservative assumption of a Kolmogorov spectrum with infinite outer scale.

9.5 Dual-Star Interferometry

Simultaneous Observations of Two Stars

For faint objects one would like to emulate the phase calibration procedure widely used in radio astronomy in which the atmospheric phase is determined from a bright source near the target. In radio interferometry one can slew the telescope between target and reference in intervals of several minutes, but because of the short atmospheric coherence time at visible and near-infrared wavelengths, here the target and the reference have to be observed truly simultaneously. Off-source fringe tracking is therefore possible only in interferometers with a field much wider than feasible in a Michelson instrument; either a wide-field (e.g., Fizeau) setup or a *dual-star system* is required. The discussion of atmospheric limitations of ground-based interferometric astrometry (Sect. 9.4) also assumed that the target and reference are observed simultaneously, again calling for a wide-field or dual-star system.

In a dual-star interferometer, each telescope accepts two small fields and sends two separate beams through the delay lines (see Fig. 48). The delay difference between the two fields is taken out with an additional short-stroke differential delay line; an internal laser metrology system is used to monitor the delay difference (which is equal to the phase difference multiplied with $\lambda/2\pi$, of course). For astrometric observations, this delay difference ΔD is the observable of interest, because it is directly related to the coordinate difference between the target (subscript t) and reference stars (subscript r); from (171) it follows immediately that

$$\Delta D \equiv D_t - D_r = \vec{B} \cdot (\hat{s}_t - \hat{s}_r) = B(\cos\theta_t - \cos\theta_r)\,. \tag{177}$$

To get robust two-dimensional position measurements, observations of the target with respect to several references and with a number of baseline orientations are required.

Narrow-Angle Astrometry

Measurements of the delay difference between two stars give *relative* astrometric information; this means that the position information is not obtained in a global reference frame, but only with respect to the nearby comparison stars, which define a local reference frame on a small patch of sky. We have seen that this approach greatly reduces the atmospheric errors, and some instrumental requirements are also relaxed (see below). The downside is that the information that can be obtained in this way is more restricted, because the local frame may have a motion and rotation of its own. This obviously makes it impossible to measure proper motions. Moreover, all parallax ellipses have the same orientation and axial ratio, which allows only "relative parallaxes" to be measured.

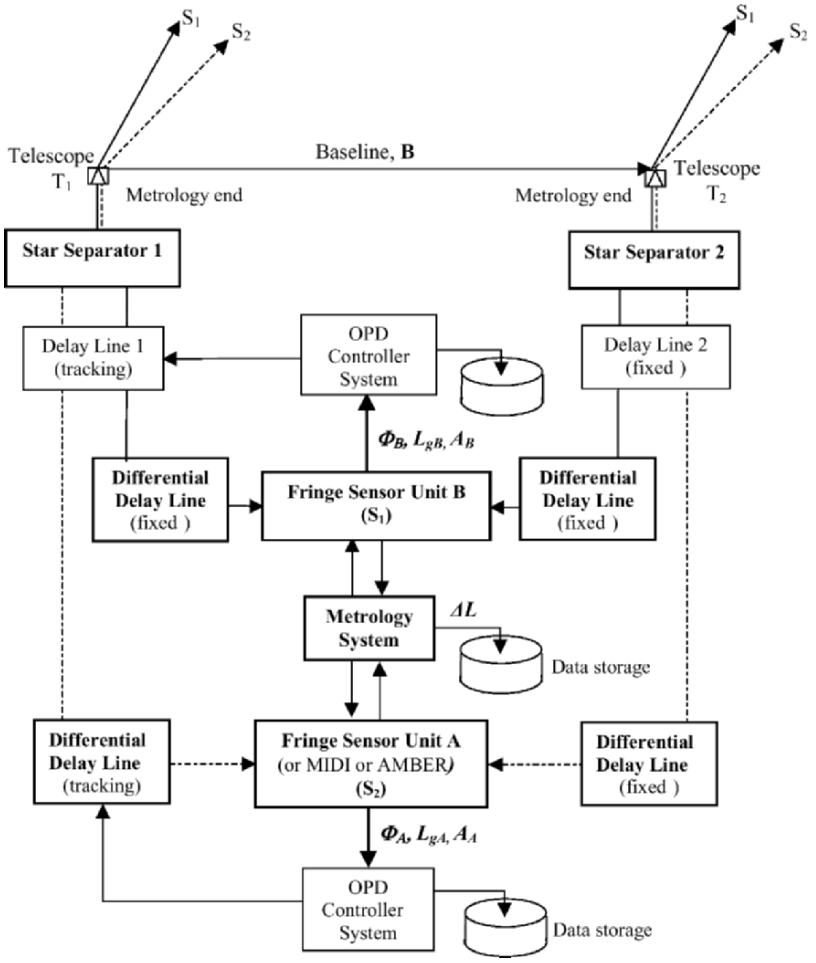

Fig. 48. Schematic setup of a dual-star interferometer. The star separators (also called dual-star modules) send the light from two stars down the main delay lines and separate differential delay lines into two beam combiners. The differential delay ΔL is measured by an internal metrology system, which measures the pathlengths backwards from the beam combiners to common metrology fiducials. The observed fringe phases $\Phi_{A,B}$, amplitudes $A_{A,B}$ and group delays $L_{gA,B}$ are recorded along with ΔL.

Narrow-angle astrometry is generally sufficient for planet detection, but there are a few caveats. First of all, if only one reference star is used, there is an ambiguity to which star an astrometric wobble has to be attributed. For example, if one chooses a distant star as reference, and this star happens to be an unrecognized binary, the resulting variation of the position difference could be mistaken for planetary signature of the much closer target (see (170)).

This can of course be avoided by using multiple reference stars, at the cost of an increase in observing time and somewhat reduced sensitivity. (Some of the signal-to-noise is used to sort out which of the references are binaries and which are not.) A somewhat more subtle effect is caused by the rotation of the local reference frame, due to uncertainties in the proper motions of the reference stars.[34] This rotation couples to the proper motion of the target star, and produces a spurious "Coriolis" acceleration, which could be mistaken for the signature of a planet in an orbit with a period longer than the time span covered by the observations. The detection of planets in long-period orbits with narrow-angle astrometry therefore requires accurate knowledge of the proper motions of the reference stars.

Anisoplanatism and Sky Coverage

The dual-star technique is being developed for interferometric astrometry and for phase-referenced visibility averaging or phase-referenced imaging. The most important problem encountered by these off-source phase-referencing techniques is anisoplanatism. The phase noise associated with anisoplanatism reduces the phase-referenced visibility dramatically if the distance to the reference source exceeds the isoplanatic angle. The need to find a reference object within the isoplanatic patch (165) is a severe limitation for off-source phase-referencing; the chances to find a suitably bright star for a randomly chosen target are typically one in a hundred or worse.[35] This is a substantial problem for interferometric observations of faint targets such as extragalactic objects or microlensing events.

The interferometric reference source can also be used for adaptive optics wavefront sensing, if such a system is available. In this case the whole entrance pupil of the interferometer is made fully co-phased and the sensitivity of the interferometer is essentially identical to the sensitivity of a single telescope with the same diameter. It is thus important to realize that bright objects are needed to co-phase an interferometer, but very faint sources can be observed in a limited field around these reference sources.

These principles can also be applied to astrometric observations; where we need to find astrometric reference stars near the intended target. In the case of astrometric planet searches, the target is a nearby and therefore bright star, which can be used for fringe tracking. The point is that now the astrometric

[34] For illustration purposes, assume that the local frame is defined by two stars. The star in the North has an Eastward proper motion, the star in the South a Westward one; these proper motions are not known to the observer. The reference frame in which these stars are at rest will have a counter-clockwise rotation.

[35] The probability depends strongly on the observing wavelength (because of the $\lambda^{6/5}$ scaling of the isoplanatic angle, (165)) and on the Galactic latitude. The stellar environment also plays a role, e.g. for observations in clusters or toward dark clouds.

references can be quite faint, because phase-referenced fringe tracking can be applied to them: the *astrometric target* is the *interferometric reference*. It is thus possible to measure the fringe phase on the astrometric reference stars with a co-phased instrument; the only limitation on their brightness is the photon noise contribution to the astrometric error budget. For a half-hour integration with 1.8 m apertures and an intended astrometric error of $10 \ldots 20\,\mu$as this limit is about $m_K = 16 \ldots 17$. The probability that suitable reference stars can be found will be further discussed in Sect. 9.8.

Instrumental Requirements

The fundamental instrumental requirements can be derived directly from (177), which can be written as

$$\Delta D \equiv D_t - D_r = \vec{B} \cdot (\hat{s}_t - \hat{s}_r) \equiv \vec{B} \cdot \Delta\vec{s}\,. \tag{178}$$

The propagation of systematic errors in measurements of the differential delay $\delta\Delta D$ and of the baseline vector δB to errors in the derived position difference $\delta\Delta s$ can be estimated from the total differential

$$\delta\Delta s \approx \frac{\delta\Delta D}{B} + \frac{\Delta D}{B^2}\delta B = \frac{\delta\Delta D}{B} + \Delta s\frac{\delta B}{B}\,. \tag{179}$$

This formula allows us to draw two important conclusions. First, the systematic astrometric error is inversely proportional to the baseline length. Together with the $B^{-2/3}$ scaling of the atmospheric differential delay r.m.s. (174) this clearly favors longer baselines, up to the limit where the target star gets resolved by the interferometer. The second important conclusion from (179) is that the relative error of the baseline measurement gets multiplied with Δs; this means that the requirement on the knowledge of the baseline vector is sufficiently relaxed to make calibration schemes possible that rely primarily on the stability of the telescope mount. For a $10\,\mu$as (50 prad)contribution to the error budget for a measurement over a $20''$ angle, with an interferometer with a 100 m baseline, the metrology system must measure $\delta\Delta D$ with a 5 nm precision; the baseline vector has to be known to $\delta B \approx 50\,\mu$m (Quirrenbach et al. 1998). These are very demanding specifications, but within reach of the Very Large Telescope Interferometer and the Keck Interferometer.

9.6 Interferometric Astrometry from Space

The Space Interferometry Mission (SIM) and GAIA

While many interesting astrometric projects on extrasolar planets can be carried out from the ground (Sect. 9.7), overcoming the fundamental limitations imposed by atmospheric anisoplanatism requires going to space. Orbits that leave the immediate vicinity of the Earth (e.g., L2 orbits or drift-away orbits) provide the added bonus of a quiet environment with stable heat flux

and low vibration levels. NASA's Space Interferometry Mission (SIM), to be launched in 2010, will exploit these advantages to perform a diverse astrometric observing program (NASA 1999a). SIM is essentially a single-baseline interferometer with 30 cm telescopes on a baseline of length 10 m. SIM is a pointed mission, i.e., targets can be observed whenever there is a scientific need (subject only to scheduling and Solar exclusion angle constraints), and the integration time can be matched to the desired signal-to-noise ratio. The limiting magnitude of SIM with "reasonable" observing times (of order one hour per visit) is thus $m_V \approx 19 \ldots 20$. The main scientific driver of SIM is observing extrasolar planets, but a wide range projects addressing Galactic and extragalactic astrophysics will also be carried out (e.g., Quirrenbach 2002a).

The European Space Agency is planning to launch an astrometric satellite of their own, called GAIA, in roughly the same time frame as SIM. GAIA's architecture builds on the successful Hipparcos mission (Lindegren and Perryman 1996). Unlike SIM, GAIA will be a continuously scanning survey instrument with a large field of view, which results in an enormous number of observed stars. The main thrust of GAIA will thus be in the area of Galactic structure (Gilmore et al. 2000; Perryman et al. 2001), but it can also detect extrasolar planets (Lattanzi et al. 2000; Sozzetti et al. 2001). SIM and GAIA will be complementary to each other in this area. SIM will provide at least one order of magnitude better accuracy and the ability to obtain well-sampled orbits (especially of multiple systems) with suitably timed observations, whereas GAIA will potentially discover massive planets around a larger number of stars.

SIM Observing Modes

For any fixed spacecraft orientation, SIM will be able to access stars in a field with a diameter of $\sim 15°$. Within each such "tile", a few stars will be selected before the mission to define an astrometric reference grid. A basic observing block will consist of observations of the target object(s) interleaved with observations of the grid stars in the tile; the measured delay differences will thus yield one-dimensional positions of the target(s) relative to the grid stars. The second coordinate can be obtained by rotating the spacecraft around the line of sight by an angle close to 90°. During the course of the mission, the grid stars will be visited regularly, about four to five times per year. By observing overlapping tiles, a full-sky reference grid can be constructed in the same manner as overlapping plates have been used to assemble all-sky astrometric catalogs. The inclusion of quasars in the grid will ensure that this reference system will represent an inertial frame. The expected performance of SIM in this "global astrometry" mode is 4 μas precision on the derived astrometric parameters (position, parallax) at the end of the nominal 5-year mission.

Many terms in the error budget of SIM measurements depend on the angle between the target and the astrometric references. It is thus possible to

improve the accuracy by choosing references close ($\lesssim 1^\circ$) to the object of interest. For this "narrow-angle" mode[36], an accuracy of 1 μas for each measurement is envisaged. (Note that this number has to be divided by $\sqrt{N}$ with $N = 30 \ldots 100$ for the number N of visits to obtain the "mission accuracy". This is the fair comparison with the "global" mode and with the figures usually quoted for other missions.) These narrow-angle measurements will not be in an inertial reference frame, but only with respect to the selected reference stars, as discussed in Sect. 9.5. The narrow-angle precision will thus not apply to absolute parallaxes and proper motions, but this is the relevant number for determining the astrometric wobble due to unseen companions.

From Observations to Astrometric Data

For an instrument like SIM, the path from the observables to the final astrometric data products is very complex. The following list illustrate the various steps that have to be taken, and the most important instrumental and astrophysical corrections that must be applied (NASA 1999a):

1. The spacecraft slews to a tile, and acquires fringes on each grid star and the science target(s) in turn, with the delay line positions and fringe phases yielding the white-light fringe positions as the primary observables. These delay line measurements include instrumental terms. During the observations, the guide interferometers, which are locked onto two bright stars, maintain the baseline orientation by actuating optical elements, while the spacecraft attitude control system maintains the coarse baseline orientation.
2. The baseline is reoriented by a spacecraft slew to a nearly orthogonal orientation. Step 1 is repeated for the new orientation, yielding a complete set of observations from which to construct the two-dimensional geometry of the target and local grid stars. A large amount of additional information necessary for reconstructing the astrometric positions of the target is recorded.
3. The data are now ready for a first reduction on the ground. A number of deterministic effects – for example, stellar aberration determined from an accurate spacecraft ephemeris, and large proper motion for some targets – are removed, and the delays are averaged over a single pointing for each object. Following this step, the delays are in the form of two sets of averaged positions, one for each baseline orientation.

[36] Note that the term "narrow-angle" as used here denotes the regime where angle-independent terms in the error budget dominate over terms increasing with angle. This is slightly different from the use of this term in Sect. 9.4, where it denoted a break in the scaling laws of atmospheric errors. One has to keep in mind that "narrow-angle" for SIM means $\lesssim 1^\circ$, whereas for ground-based astrometry it means $\lesssim 1'$.

4. The pairwise differences in these delays comprise the important observables in reconstructing the geometry of the reference stars and target, which are related to the set of star positions.
5. Metrology and supporting data, which determine the baseline length, and star-tracking pointing information are similarly averaged to give a first guess at the baseline $\vec{B}$.
6. The true baseline $\vec{B}$ for each orientation is found by iteratively solving the interferometric astrometry equation for the science interferometer using the grid star differential delays, a model for the grid that specifies grid star positions as a function of time in an inertial frame, and the temporal component of general relativistic effects obtained from a relativistic Solar System model. While the procedure for carrying out this computation for an ideal instrument is straightforward, a number of additional instrumental and systematic effects must be characterized, understood, and modeled out at this step.
7. The differential delays involving the science target are similarly calibrated using relativistic and other instrumental and systematic effects. They now represent a geometrically consistent set of fixed or rectilinearly moving points. In other words, we now have the differential delays that would be produced by an ideal instrument taken instantaneously from a fixed point relative to the Solar System barycenter at a particular time.
8. The calibrated differential delays are now transformed into a set of object positions (equivalently $\hat{s}$) in the local coordinate system of the reference grid objects.
9. The object positions $\hat{s}$ in the Solar System barycenter frame referenced to zero potential are now found, taking into account general relativistic effects arising from the particular distribution of masses in the Solar System at the mean time of the observations.
10. A set of measurements covering the whole sky multiple times is used to solve for the standard astrometric parameters (position, parallax, proper motion) of the grid stars. This information is fed back into Step 6.
11. The residuals of the target star positions after fitting the standard astrometric can be analyzed for the presence of planetary companions.

It should thus be clear that attaining the best possible sensitivity to low-mass planets requires a very detailed understanding of a plethora of instrumental and astrophysical effects. Misinterpreting systematic drifts, position- or color-dependent errors, confusion effects, correlations between parallax, proper motion and the parameters of fits to the residuals, or any other subtleties may well lead to errors in the derived orbital elements of the detected planets, or worse, to false planet detections. On the other hand, SIM will acquire a very large set of individual measurements that can be analyzed in a systematic and consistent way. If many null results are obtained (i.e., stars that do *not* show any residuals indicative of planets), one can be confident of planets discovered around the others.

9.7 Astrometric Planet Surveys

Questions that can be addressed

- Mass determination for planets detected in radial velocity surveys (without the $\sin i$ factor). The RV method gives only a lower limit to the mass, because the inclination of the orbit with respect to the line-of-sight remains unknown. Astrometry can resolve this ambiguity, because it measures two components of the orbital motion, from which the inclination can be derived.
- Confirmation of hints for long-period planets in RV surveys. Many of the stars with detected short-period planets also show long-term trends in the velocity residuals (Fischer et al. 2001, see Sect. 3.2). These are indicative of additional long-period planets, whose presence can be confirmed astrometrically.
- Inventory of planets around stars of all masses. The RV technique works well only for stars with a sufficient number of narrow spectral lines, i.e., fairly old stars with $m_* \lesssim 1.2\,M_\odot$. Astrometry can detect planets around more massive stars and complete a census of gas and ice giants around stars of all masses.
- Detection of gas giants around pre-main-sequence stars, signatures of planet formation. Astrometry can detect giant planets around young stars, and thus probe the time of planet formation and migration. Observations of pre-main-sequence stars of different ages can provide a critical test of the formation mechanism of gas giants. Whereas gas accretion on $\sim$10 $\mathrm{M}_\oplus$ cores requires $\sim$10 Myr, formation by disk instabilities would proceed rapidly and thus produce an astrometric signature even at very young stellar ages (Boss 1998b).
- Detection of multiple systems with masses decreasing from the inside out. Whereas the astrometric signal increases linearly with the semi-major axis a of the planetary orbit, the RV signal scales with $1/\sqrt{a}$. This leads to opposite detection biases for the two methods. Systems in which the masses increase with a (e.g., υ And, Butler et al. 1999) are easily detected by the RV technique because the planets' signatures are of similar amplitudes. Conversely, systems with masses decreasing with a are more easily detected astrometrically.
- Determine whether multiple systems are coplanar or not. Many of the known extrasolar planets have highly eccentric orbits. A plausible origin of these eccentricities is strong gravitational interaction between two or several massive planets (Lin and Ida 1997; Papaloizou and Terquem 2001). This could also lead to orbits that are not aligned with the equatorial plane of the star, and to non-coplanar orbits in multiple systems.
- Search for massive terrestrial planets in the Solar neighborhood. NASA's Space Interferometry Mission (SIM) has a 1 μas precision goal in its "narrow-angle" mode (over 1°). This opens the exciting perspective of looking for rocky planets down to a limit of a few Earth masses.

9.8 Astrometric References

Reference Stars for Ground-Based Narrow-Angle Astrometry

The discussion of optical/infrared interferometry in Sect. 9.5 has tacitly assumed that there are reference stars against which the motion of the target star can be measured. Identifying suitable references, however, is an important and non-trivial task by itself. It is possible to get a feeling for the difficulty of this task by cross-correlating a catalog of potential targets (e.g. the Hipparcos catalog, European Space Agency 1997, or the Gliese catalog) with a sufficiently deep catalog of possible reference stars (USNO-A1.0, Monet et al. 1996). Quirrenbach (2000a) did this exercise for potential targets for the VLT Interferometer, with the following results:

- The Hipparcos catalog contains 4,760 stars with $\delta \leq +20^\circ$ and parallax $\pi \geq 20\,\mathrm{mas}$.
- For 1,762 of these stars, 3 reference stars within $50''$ are found.
- For 734 of these, 3 reference stars within $30''$ are found.
- The Hipparcos catalog contains 1,018 stars that have $\delta \leq +20^\circ$ and $\pi \geq 40\,\mathrm{mas}$, and 130 stars with $\delta \leq +20^\circ$ and $\pi \geq 100\,\mathrm{mas}$.
- The Gliese catalog contains 2,381 stars with $\delta \leq +20^\circ$.
- The proportion of these samples for which potential references within $30''$ or $50''$ are found is about the same as for the full Hipparcos sample.

These numbers indicate that a large number of target stars are available for ground-based astrometric projects.

There are a few caveats about the results from these catalog cross-correlations, however, which may lead to an overestimate of the number of potential reference stars. For example, some of the "reference stars" may actually be galaxies that were not recognized as such during the compilation of the USNO catalog. In some cases, a target star with high proper motion is found as a potential "reference" for itself because the USNO catalog lists the stellar positions at the epoch of the underlying Schmidt plates. (It would be necessary to look up the plate epoch for each target star, compute its position at that epoch, and compare it to the positions of the "references" found by the cross-correlation. This is even more important when reference stars within $10''$ from the target are sought.) On the other hand, the available catalogs may be incomplete in the vicinity of very bright stars; one has to take CCD images in the field of each potential target to make sure that one does not miss any available references (Creech-Eakman et al. 1999).

These cautionary remarks notwithstanding, it is apparent from the above numbers that the requirement of finding nearby reference stars reduces the number of potential targets considerably, especially for the highest precision when the search radius is only $10''$. The problem is even somewhat exacerbated by the high proper motion of many of the nearby target stars: a star that has three references today may move far enough away from them during the

course of a several decade-long program that the astrometric accuracy gets severely degraded. It is obviously possible to increase the number of potential targets by requiring only one or two references, but then the risk gets high that the references turn out to be members of multiple systems, which can lead to unusable data, or worse, to false "planet" detections if the motion of the reference is ascribed to the target.

It is also possible to search for planets in double stars (Quirrenbach 2000a). The Washington Double Star Catalog (Worley and Douglass 1997) contains 745 F, G, and K main sequence stars with $\delta \leq 20^\circ$ and $V \leq 10$ in pairs with separation between $5''$ and $20''$; 23 of these are G main sequence stars with $V \leq 7.5$. Most of these are members of wide physical binaries, and searching for planets in these systems is scientifically interesting (see Sect. 3.4) and technically somewhat less challenging than searches around single stars. The downside of this approach is the difficulty of determining to which of the two system components any detected planet belongs.

The bottom line is that the availability of nearby astrometric references is an important criterion for the selection of suitable targets for ground-based astrometric observations. Optimizing the sensitivity of the faint star channel is clearly very important, because this enhances the chances of finding astrometric references. The general considerations in this regard are similar to the optimum design of a fringe tracker (Sect. 8.5). In addition, one should use as many photons as possible; simultaneous operation in the H and K bands is therefore highly advantageous.

Reference Stars for the Space Interferometry Mission

The success of SIM also depends critically on the selection of suitable grid and reference stars. About $2,000 \ldots 3,000$ stars distributed evenly over the sky are needed for the grid. These stars must be astrometrically stable (i.e., they must not show any motion other than parallax and linear proper motion) on the level of a few μas. For each narrow-angle target, at least three reference stars stable to better than 1 μas must be available within a $\sim 1^\circ$ diameter field. Finding stars that meet these stability requirements is by no means a trivial task. They have to be identified and characterized before the launch of SIM, because the potential presence of even a fairly small number of unstable references can only be compensated by a dramatic increase in redundancy of the observations (Frink et al. 2001), which is very costly in terms of SIM observing time.

A key criterion for selecting suitable grid and reference stars for SIM is their distance d, because the wobble induced by planets or other unseen companions scales with $1/d$ (170). The best class of reference stars are therefore K giants, which are numerous even at high galactic latitudes, and intrinsically bright. Samples of candidate SIM reference stars can either be selected from existing astrometric catalogs (Frink et al. 2000a,b), or identified in a

specialized survey (Patterson et al. 1999; Rhee et al. 2001). The case for distant K giants as good grid stars rests on a three-fold argument (Frink et al. 2001):

- The wobble due to planetary companions is sufficiently small.
- We know from the radial-velocity (RV) surveys that brown dwarfs are rare as companions to G dwarfs (the "brown dwarf desert", see Sect. 3.2 or e.g. Marcy and Butler 2000), which are the progenitors of K giants. Brown-dwarf companions to K giants will therefore also be rare.
- Stellar companions can be detected efficiently before the launch of SIM by an RV survey. This is a non-trivial statement as photospheric activity could corrupt precise RV measurements. A survey with the Hamilton Echelle Spectrograph at Lick Observatory has shown, however, that many K giants are sufficiently stable; about 2/3 of all K giants are drawn from a distribution with a mean of $\sim 20\,\mathrm{m\,s^{-1}}$ (see Fig. 49). This allows the detection of most stellar companions with only two or three RV data points.

It should thus be possible to define the SIM grid, and to gain confidence in its integrity, well before launch. The selection of narrow-angle reference stars is a more difficult problem because of the increased accuracy requirement and more limited search area for suitable stars. Observational programs aimed at identifying candidate reference stars for high-priority narrow-angle targets

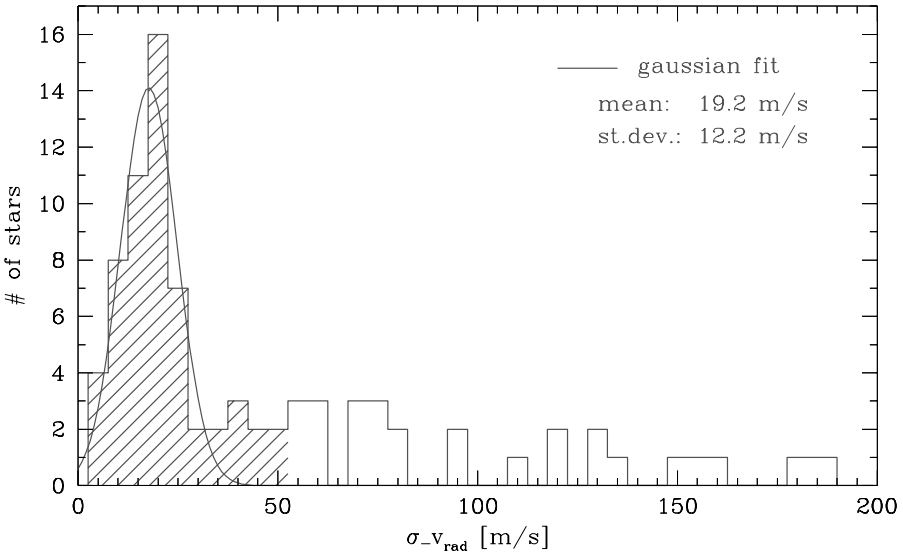

Fig. 49. Histogram of radial velocity scatter (i.e., dispersion of repeated RV measurements) observed in a sample of K giants (updated from Frink et al. 2001). About 2/3 of the observed stars have radial velocity scatter of $19.2 \pm 12.2\,\mathrm{m\,s^{-1}}$ and are good candidates for the SIM grid. Stars showing larger RV scatter could have companions and would not be included in the grid

have recently been started. Without this preparation it would not be possible to carry out an efficient planet detection program with SIM.

9.9 The Differential-Phase Method

Another interesting application of phase referencing, which is related to astrometry, consists of making differential phase measurements between different wavelengths (Akeson and Swain 1999; Quirrenbach 2000a). The near-infrared spectra of giant extrasolar planets should be characterized by extremely deep absorption bands of water and methane (e.g. Burrows et al. 1997a, see also Sect. 6.6). This opens the possibility of using wavelength-dependent astrometric information for the detection of extrasolar planets, and even to obtain their spectra.

The photocenter of a star–planet system is slightly different between two wavelengths, one of which falls in a region free of molecular bands, where the planet is relatively bright, and the other inside a band, where the planet is much fainter (see Fig. 50). Actually, the shift of the photocenter is proportional to the planet/star brightness ratio and can thus be used as a proxy for the planet spectrum (Quirrenbach 2000a). The shift of the photocenter gives rise to a corresponding wavelength dependence of the interferometer phase, which can be measured if the signal-to-noise ratio is sufficient and systematic effects are kept small. In the case of "hot Jupiters" like 51 Peg b, which is quite favorable because the planet is close to the star and therefore hot and bright, the expected effect on the interferometer phase is $\approx 0.5\,\mathrm{mrad}$ on the longest baselines of the VLTI (see Fig. 51). To measure such a small phase difference, a signal-to-noise ratio of $\approx 3,000$ is needed, and it remains to be seen whether

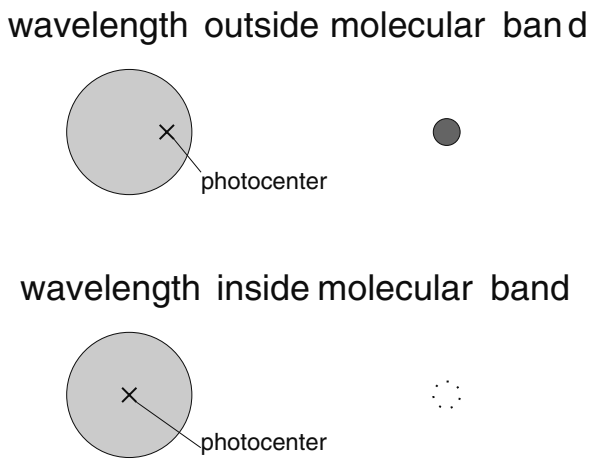

Fig. 50. The shift of the star–planet photocenter with wavelength gives rise to an interferometric phase shift that can be exploited to obtain a spectrum of the planet

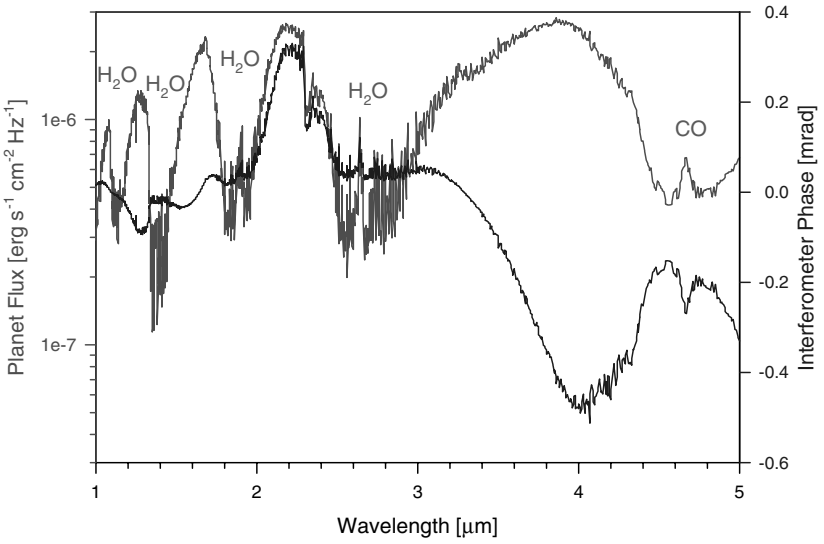

Fig. 51. Model spectrum of the planet 51 Peg b (red, from Sudarsky et al. 2003), and interferometer phase predicted for a 100 m baseline aligned with the star–planet separation vector (blue). The planetary spectrum is dominated by absorption bands of water and carbon monoxide. At short wavelengths ($\leq 2\,\mu$m) it is very difficult to detect the planet because of the very high contrast between the star and the planet. The phase changes significantly across the K band near 2.2 μm due to water bands shortward and longward of the window in the Earth's atmosphere that defines this observing band. For the specific baseline length chosen, the phase goes through zero near 3.3 μm

the systematic instrumental and atmospheric effects can be overcome at this level. For this technique, the dispersion in the air in the delay lines is a serious difficulty (Daigne and Lestrade 1999; Meisner and le Poole 2002). This problem can be overcome either by the use of a evacuated delay lines, or by making double differential measurements with respect to a nearby reference star

$$\phi_{dd} \equiv \phi_t(\lambda_1) - \phi_t(\lambda_2) - [\phi_r(\lambda_1) - \phi_r(\lambda_2)] \ , \tag{180}$$

where the subscripts t and r stand for "target" and "reference", respectively (le Poole and Quirrenbach 2002). If the reference star has no companion, its photocenter position is independent of wavelength, and the term $[\phi_r(\lambda_1) - \phi_r(\lambda_2)]$ is equal to the phase offset caused by dispersion, which is thus subtracted from the phase difference of the target in (180). But even with this double differential technique the spectroscopy of extrasolar planets will be a very challenging project.

10 Nulling Interferometry

The principal problem of direct planet detection is the large contrast between the planet and the parent star. Bracewell (1978) proposed to overcome this difficulty by using an interferometer to suppress the starlight. The key to the success of this method is the creation of an *achromatic null*, which ensures that the light arriving on axis interferes destructively at all wavelengths within the observing bandpass. Nulling interferometry in the mid-IR from the ground is a promising approach to the detection of exozodiacal dust disks. From space, the contrast and signal-to-noise ratio can be made sufficient for low-resolution spectroscopy of Earth-like planets. This is one of the leading architectures that have been proposed for the DARWIN and Terrestrial Planet Finder missions.

10.1 Principles of Nulling

Starlight Rejection in a Michelson Interferometer

For monochromatic light, the output intensity of a standard Michelson interferometer with an ideal 50% beam combiner varies with the phase ϕ as

$$I_{\rm out} = I_{\rm in}(1 + V\cos\phi) \ , \tag{181}$$

where V is the fringe visibility and $I_{\rm in} \equiv I_1 + I_2$ the sum of the intensities in the two interferometer input arms. This means that the intensity oscillates between $I_{\rm min} = (1-V)I_{\rm in}$ and $I_{\rm max} = (1+V)I_{\rm in}$ when the delay D is varied, in agreement with (139). If $V = 1$, we can set the delay line such that $\phi = 2\pi D/\lambda = 180^\circ$ and get completely destructive interference, $I_{\rm out} = 0$. For a non-zero bandwidth, we have to integrate the right-hand side of (181) over frequency; for $V = 1$ and a rectangular bandpass we thus get

$$\begin{aligned} I_{\rm out} &= \frac{1}{\Delta\nu} I_{\rm in} \int_{\nu_0-\Delta\nu/2}^{\nu_0+\Delta\nu/2} \left[1 + \cos\left(2\pi D\nu/c\right)\right] d\nu \\ &= I_{\rm in}\left[1 + \cos\left(\frac{2\pi D\nu_0}{c}\right) \cdot \mathrm{sinc}\left(\frac{\pi D\Delta\nu}{c}\right)\right] \ , \end{aligned} \tag{182}$$

in analogy to (147). If $\Delta\nu/\nu_0 \approx \Delta\lambda/\lambda \ll 1$, the cosine term in (182) varies much faster than the sinc term; the minima of the right-hand side therefore occur close to the delays for which the cosine term assumes the value -1. The condition for the first and deepest of these minima is $D = c/(2\nu_0)$, and we obtain

$$\frac{I_{\rm min}}{I_{\rm max}} \approx \frac{1}{2}\left[1 - \mathrm{sinc}\left(\frac{\pi}{2}\frac{\Delta\lambda}{\lambda}\right)\right] \ . \tag{183}$$

For bandwidths of 10%, 30%, and 50%, the depth of the first minimum is 0.2%, 2%, and 5%, respectively. It is thus possible to reject most of the starlight

by simply offsetting the delay line to the first interference minimum, but rejection factors of more than a few hundred are possible only with a very small bandwidth.[37] An achromatic method to generate destructive interference is therefore needed.

The Principle of Achromatic Nulling

If we can introduce a phase shift in the interferometer, which is exactly π rad at all wavelengths, the output signal from an interferometer is changed to

$$I_{\rm out} = I_{\rm in}(1 - V\cos\phi)\,. \tag{184}$$

Now the intensity is zero if $V = 1$ and $\phi = 0$, i.e., the light from an on-axis point source is completely rejected at zero delay. This is the principle of *achromatic nulling*. A nulling interferometer can be used to detect extrasolar planets in the following way (see Fig. 52): The parent star is placed on the interferometer line-of-sight, and a fringe tracker assures that $\phi = 0$, so that no light from the star is received. The planet, on the other hand, is located at an off-axis angle which is comparable to the interferometer resolution λ/B. The light from the planet therefore has a significantly non-zero phase, leading to a detectable interferometer output according to (184). The nulling interferometer thus acts as an ideal coronographic mask, with complete rejection of the starlight and only moderate attenuation of the planetary signal.

In practice, the null is never perfect, of course, because there are always wavefront corrugations, phase fluctuations, and internal contrast losses, which reduce the visibility. This means that $I_{\rm out} \neq 0$ even in the absence of planets. It is therefore necessary to modulate the signal from the planet, for instance by rotating the interferometer around its axis, which leads to a periodic modulation of the projected baseline length and therefore of ϕ and $I_{\rm out}$. In this way it is possible to separate the constant term due to the starlight leak (or a uniform face-on dust disk) from the AC term due to the planet (see Fig. 53).

To characterize the quality of a nulling instrument and to quantify the leakage of unwanted photons, we introduce the *null depth*

$$N \equiv I_{\rm min}/I_{\rm max}\,. \tag{185}$$

For low-resolution spectroscopy of Earth-like planets with space-based interferometers an extremely deep null ($N \lesssim 10^{-6}$) is required; for the detection of exozodiacal dust disks $N \approx 10^{-3} \ldots 10^{-4}$ is sufficient.

[37] This conclusion is strictly valid only for single-baseline nulling interferometers. Mieremet and Braat (2002a) have shown that interferometric arrays with multiple telescopes can produce a deep broad-band null if appropriate delays are introduced in the interferometer arms.

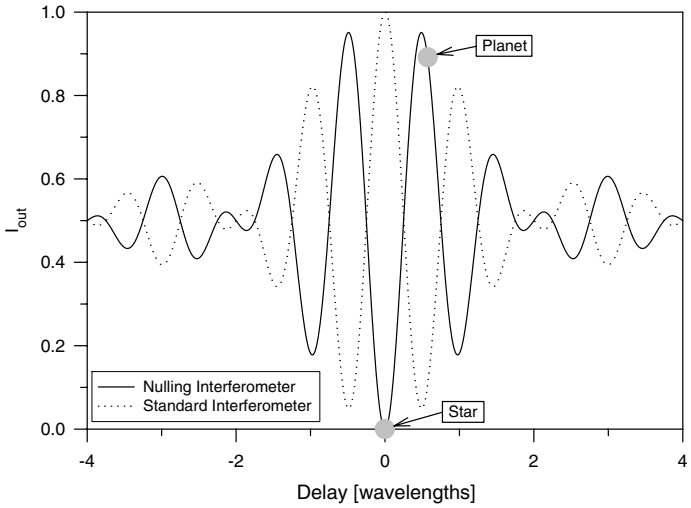

Fig. 52. Fringe pattern for a standard and a nulling interferometer. The fractional bandwidth $\Delta\lambda/\lambda = 0.5$. The pattern of the nulling interferometer has a central zero; the depth of the first minima of the standard interferometer are 5%. Fringe tracking ensures that the delay for the star is zero. The resolution of the interferometer is matched to the star–planet separation such that the planet is close to a transmission maximum

Symmetry and Stability Requirements

To produce a deep null, the two arms of the interferometer must be made symmetric with very high precision (apart from the π phase shift, of course). Any of the following imperfections can ruin the performance of the nuller: residual phase fluctuations, differences between the two arms in dispersion or polarization properties, a rotation between the two beams, and a mismatch between the two intensities (e.g., Serabyn 2000; Wallner et al. 2001). We denote the residual phase (i.e., the difference between the actual phase and the "best compromise" for all wavelengths within the bandpass and the two polarization states) by $\Delta\phi$, the rms phase difference due to dispersion mismatch, averaged over the bandpass, by $\Delta\phi_\lambda$, the phase difference between the two polarization states by $\Delta\phi_{s-p}$, the relative rotation angle between the two beams by α, and the normalized intensity mismatch by $\delta I/I \approx \delta \ln I$. A fairly straightforward analysis shows that the null depth is then given by (Serabyn 2000)

$$N = \frac{1}{4}\left[(\Delta\phi)^2 + (\Delta\phi_\lambda)^2 + \frac{1}{4}(\Delta\phi_{s-p})^2 + \alpha^2 + (\delta \ln I)^2\right] . \qquad (186)$$

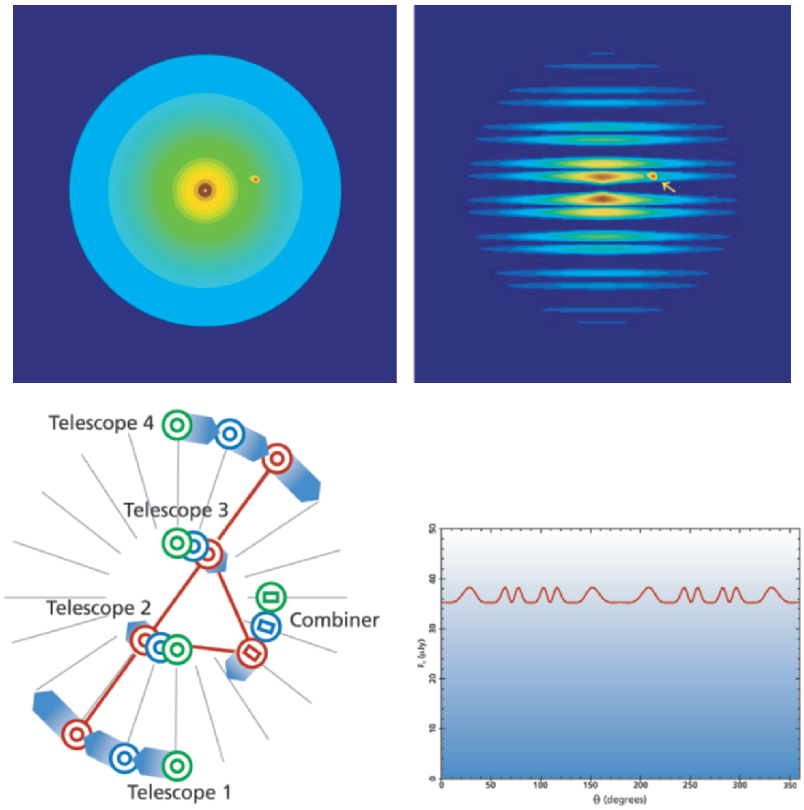

Fig. 53. *Top left*: model face-on planetary system with one planet and exozodiacal dust. *Top right*: the same system, multiplied with the response function of a linear four-element nulling interferometer. *Bottom left*: positions of the four telescopes and the beam combiner during rotation of the interferometer. *Bottom right*: output of the nuller for one full rotation. From NASA (1999b)

For illustration, we show how N can be calculated in the case that intensity mismatch is the only imperfection in the interferometer. $I_{\rm min}$ and $I_{\rm max}$ are then given by

$$\begin{aligned} I_{\rm min} &= (I+\delta I)+(I-\delta I)-2\sqrt{(I+\delta I)(I-\delta I)} \\ &= 2\left(I-\sqrt{I^2-(\delta I)^2}\right) \\ &= 2I\left(1-\sqrt{1-\left(\frac{\delta I}{I}\right)^2}\right) \approx \frac{(\delta I)^2}{I}\ , \end{aligned} \tag{187}$$

and

$$I_{\rm max} = (I+\delta I)+(I-\delta I)+2\sqrt{(I+\delta I)(I-\delta I)} \approx 4I\ , \tag{188}$$

from which we get

$$N \equiv I_{\min}/I_{\max} \approx \frac{1}{4}\left(\frac{\delta I}{I}\right)^2 = \frac{1}{4}(\delta \ln I)^2 \ , \tag{189}$$

in agreement with the more general expression given in (186). To achieve a good null depth it is obviously necessary to control the optical path difference between the two interferometer arms, and to balance their intensities. If the mean values for $\Delta\phi$ and $\delta \ln I$ are kept at zero, the time-averaged null depth can be written as

$$\overline{N} = \frac{1}{4}\left[\sigma_\phi^2 + (\Delta\phi_\lambda)^2 + \frac{1}{4}(\Delta\phi_{s-p})^2 + \alpha^2 + \sigma_{\ln I}^2\right] \ , \tag{190}$$

where σ_ϕ^2 is the residual phase variance and $\sigma_{\ln I}^2$ the variance of the residual intensity fluctuations. If we assume that the purely instrumental effects (dispersion, polarization, beam rotation) are relatively stable, the fluctuations of the null depth around its mean value are dominated by the fluctuations of the phase and intensity. It can then be shown that the variations of the null depth are given by (Serabyn 2000)

$$\sigma_N^2 = \frac{1}{8}\left(\sigma_\phi^4 + \sigma_{\ln I}^4\right) \ . \tag{191}$$

If a certain depth of the null is desired, (190) and (191) can be used to construct an error budget for the various contributions to $\overline{N}$. A slight complication arises from the fact that the phase and intensity variations determine not only the mean null depth, but also the level of the fluctuations around this mean value. For example, if the null depth is dominated by phase fluctuations, a $+2\,\sigma$ excursion of N from its mean value means that the instantaneous null depth is

$$N_{2\sigma} = \overline{N} + 2\sigma_N = \left(\frac{1}{4} + \frac{2}{\sqrt{8}}\right)\sigma_\phi^2 \approx \sigma_\phi^2 \, . \tag{192}$$

This means that specifying an error budget for $+2\,\sigma$ fluctuations requires placing a four times more stringent requirement on the phase variance than constructing an error budget for the mean null depth. The same argument obviously also applies to intensity variations. In the error budget one can allocate a maximum value N_i to each one of the contributing effects; the sum of these values must be smaller than the desired null depth. We thus get the set of requirements

$$\sigma_\phi < \sqrt{N_1} \tag{193}$$

$$\sigma_{\ln I} < \sqrt{N_2} \tag{194}$$

$$\alpha < 2\sqrt{N_3} \tag{195}$$

$$\Delta\phi_\lambda < 2\sqrt{N_4} \tag{196}$$

$$\Delta\phi_{s-p} < 4\sqrt{N_5} \tag{197}$$

$$\sum_{i=1}^{5} N_i < \overline{N}\,. \tag{198}$$

Wavefront Aberrations and Modal Filtering

In the previous section (186) and (190) we have implicitly assumed that at each wavelength and for each polarization there is a unique phase difference between the two interferometer arms. This is not necessarily the case, however, as aberrations and atmospheric turbulence distort the wavefronts and thus create phase variations across each of the apertures. Denoting the phase variances across the two pupils with $\sigma^2_{w_{1,2}}$, we can write in analogy to (190)

$$\overline{N} = \frac{1}{4}\left(\sigma^2_{w_1} + \sigma^2_{w_2}\right) = \frac{1}{2}\sigma^2_w\ , \tag{199}$$

where the second equality holds if the contributions from the two telescopes are equal. We will see in Sect. 10.3 that (199) places a requirement on the wavefront quality that is impossible to achieve even with state-of-the-art adaptive optics systems, if a null depth of $\sim 10^{-3}$ is desired at 10 µm in a ground-based interferometer. Similarly, it appears infeasible to produce optics that would allow a space interferometer to obtain a 10^{-6} null in this way.

It is thus necessary to introduce a modal filter in the interferometer, as described in Sect. 8.5 (Clark and Roychoudhuri 1979; Mennesson et al. 2002). Since the coupling efficiency into a single-mode fiber is roughly given by the Strehl number of the input beam, we can use (123) to calculate the intensity fluctuations

$$\delta I \equiv \frac{1}{2}(I_1 - I_2) \approx \frac{1}{2}\left(e^{-\sigma^2_{w_1}} - e^{-\sigma^2_{w_2}}\right) I_0 \approx \frac{1}{2}\left(\sigma^2_{w_2} - \sigma^2_{w_1}\right) I_0\,. \tag{200}$$

The second approximation is valid only if $\sigma^2_w \ll 1$, which is the case for nulling interferometers in space. Inserting this result in (186) gives

$$N = \frac{1}{4}\left(\frac{\delta I}{I}\right)^2 = \frac{1}{16}\left(\sigma^2_{w_2} - \sigma^2_{w_1}\right)^2\ . \tag{201}$$

It is thus the *difference* between the wavefront qualities in the two beams that matters for the null depth, but in most practical circumstances this difference is directly related to the wavefront quality itself. It is also worth pointing out that (201) implies that $N \propto \lambda^{-4}$, which is a very steep scaling with the observing wavelength.

Pointing Requirements

If the telescopes are not pointed exactly at the target star, the phase across the pupil is not constant, and we expect an associated null leakage. If there is a pointing error $\delta\theta$ in the x-direction, the phase varies across the pupil according to

$$\phi(x) = \frac{2\pi x}{\lambda}\delta\theta = \frac{2\pi r\cos\psi}{\lambda}\delta\theta \;, \tag{202}$$

where we have introduced polar coordinates (r, ψ). The phase variance is thus

$$\begin{aligned}\sigma_w^2 &= \frac{1}{\pi R^2}\int_0^{2\pi} d\psi \int_0^R r dr \phi^2 \\ &= \frac{1}{\pi R^2}\frac{4\pi^2(\delta\theta)^2}{\lambda^2}\int_0^{2\pi} d\psi \int_0^R dr\, r^3 \cos^2\psi \\ &= \left(\frac{\pi D}{2\lambda}\right)^2 (\delta\theta)^2 \,.\end{aligned} \tag{203}$$

If there is no modal filtering, the resulting null depth can be computed by inserting (203) in (199)

$$\overline{N} = \frac{1}{8}\left(\frac{\pi D\sigma_\theta}{\lambda}\right)^2 \;. \tag{204}$$

In the case of modal filtering we have to insert (203) in (201) and obtain

$$\overline{N} = \frac{1}{256}\left(\frac{\pi D}{\lambda}\right)^4 \left\langle \left[(\delta\theta_1)^2 - (\delta\theta_2)^2\right]^2 \right\rangle \;. \tag{205}$$

To simplify this expression we note that for Gaussian variables χ with zero mean and standard deviation σ_χ

$$\left\langle \chi^4 \right\rangle = 3\sigma_\chi^4 \,. \tag{206}$$

Therefore

$$\left\langle \left[(\delta\theta_1)^2 - (\delta\theta_2)^2\right]^2 \right\rangle = \left\langle (\delta\theta_1)^4 \right\rangle - 2\left\langle (\delta\theta_1)^2(\delta\theta_2)^2 \right\rangle + \left\langle (\delta\theta_2)^4 \right\rangle = 4\sigma_\theta^4 \,. \tag{207}$$

Inserting this result in (205) we finally obtain the desired expression for the null depth

$$\overline{N} = \frac{1}{64}\left(\frac{\pi D\sigma_\theta}{\lambda}\right)^4 \;. \tag{208}$$

To achieve a deep null it is therefore necessary to stabilize the telescope pointing at a rather small fraction of the width of an Airy disk, which is of order λ/D.

Leakage from the Stellar Disk

Up to this point we have computed limitations of the null depth due to a variety of instrumental effects, under the assumption that we want to reject light from a point source. This assumption is violated if the star is partially resolved by the interferometer; in that case the light from the stellar limb arriving at an off-axis angle is not completely rejected even by a perfect nulling device. To first order (ignoring limb darkening) the star can be modeled as a uniform disk of angular diameter $\theta_{\rm dia}$ and normalized intensity

$$I(\mathbf{\Omega}) = \left(\frac{\pi\theta_{\rm dia}^2}{4}\right)^{-1} . \tag{209}$$

The fringe phase for light emanating from a point on the stellar surface with polar coordinates $(\theta_r,\, \psi)$ is

$$\phi = \frac{2\pi B \sin(\theta_r \cos\psi)}{\lambda} \approx \frac{2\pi B\, \theta_r \cos\psi}{\lambda} . \tag{210}$$

The null depth can easily be calculated by integrating ϕ^2 over the stellar surface, which results in

$$\begin{aligned} N &= \frac{1}{4}\int d\mathbf{\Omega}\, \phi^2(\theta_r,\psi) I(\mathbf{\Omega}) \\ &= \frac{1}{4}\int_0^{\theta_{\rm dia}/2}\int_0^{2\pi} \theta_r d\theta_r d\psi\, \phi^2(\theta_r,\psi) I(\mathbf{\Omega}) \\ &= \frac{1}{\pi}\left(\frac{2\pi B}{\theta_{\rm dia}\lambda}\right)^2 \int_0^{\theta_{\rm dia}/2}\int_0^{2\pi} \theta_r d\theta_r d\psi\, \theta_r^2 \cos^2\psi \\ &= \frac{1}{\pi}\left(\frac{2\pi B}{\theta_{\rm dia}\lambda}\right)^2 \int_0^{2\pi} d\psi \cos^2\psi \int_0^{\theta_{\rm dia}/2} d\theta_r \theta_r^3 \\ &= \frac{\pi^2}{16}\left(\frac{B\theta_{\rm dia}}{\lambda}\right)^2 . \end{aligned} \tag{211}$$

This expression relates the null depth to the ratio of the angular diameter and the resolution of the interferometer λ/B. To get a quick overview of the importance of this stellar leak for stars of different types, it is convenient to derive a scaling relation with stellar magnitude and effective temperature (Quirrenbach 2000b). For observations in the infrared ($\lambda \gtrsim 2.2\,\mu$m) we can use the Rayleigh–Jeans approximation of blackbody radiation

$$\mathcal{F} \propto \theta_{\rm dia}^2 \cdot T_{\rm eff} \tag{212}$$

to rewrite (211) as

$$N = 0.26 \left(\frac{B}{100\,{\rm m}}\right)^2 \left(\frac{\lambda}{2.2\,\mu{\rm m}}\right)^{-2} \left(\frac{T_{\rm eff}}{10,000\,{\rm K}}\right)^{-1} \cdot 10^{-0.4\,m_{\rm K}} , \tag{213}$$

where m_K is the K-band magnitude of the star. The leakage from the stellar disk is tolerable for ground-based nulling if the baseline is not too long, but it may limit the null depth for the very brightest stars. For space interferometers with a requirement of $N \approx 10^{-6}$ the stellar leakage seems to be an insurmountable obstacle, and indeed it is necessary to employ more complicated nulling schemes involving multiple baselines. These will be discussed in Sect. 10.4.

10.2 Implementation of Achromatic Phase Shifts

There are two fundamentally different approaches to implement achromatic phase shifts. The first is the introduction of a "geometric" phase shift, in most cases by 180°. Techniques based on this principle are truly achromatic in that they manipulate the phase, not the delay. The alternative is the use of dispersive elements to create "pseudo-achromatic" phase shifts; the goal here is find a combination of different materials that approximates a constant phase shift over a broad band. Both methods have shown great potential, and they will now be discussed in turn.

Geometrical Field Reversal

The basic principle of creating a field flip by purely geometrical means is illustrated in Fig. 54. The beams in the two arms of the interferometer are sent

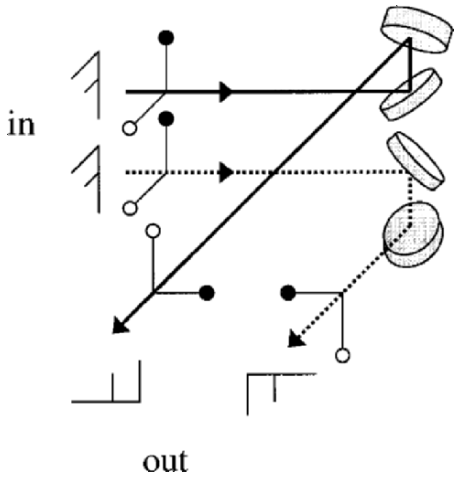

Fig. 54. Effect of a mirror-reflected pair of right-angle periscopes. Each beam encounters two mirrors at the locations of the 90° folds. Both the apertures and the fields undergo a relative rotation of 180°, as shown schematically by the clock hands and the letters "F". Each polarization component undergoes one s-plane and one p-plane reflection. From Serabyn and Colavita (2001)

through right-angle periscopes that are mirror-images of each other. This leads to an inversion of the apertures and of the relative direction of the electric fields in the two beams, which is equivalent to a 180° phase shift. Since this phase shift is achieved by reflections, it is strictly achromatic. In every other respect the field inverter is fully symmetric, e.g., each polarization component undergoes one s-plane and one p-plane reflection at a 45° incidence angle. This symmetry facilitates the task of meeting the stringent tolerances discussed in Sect. 10.1.

After passage through the pair of periscopes, the light from the two interferometer arms is sent to a pupil plane beam combiner. It is important again to make the combiner as symmetric as possible, to keep δI and $\Delta\phi_{s-p}$ as small as possible. A single pass through the beam splitter as drawn in Fig. 42 is not suitable, because it is impossible to manufacture a beam splitter for which the reflectivity and transmissivity are exactly equal to each other over a large wavelength range and for both polarizations. It is thus necessary to design a beam combiner in which each beam undergoes one reflection and one transmission before emerging at the nulled output. A further subtlety arises from the point that beam splitters are multi-layer dielectric films on a transparent substrate. In the presence of slight internal absorption, the reflectivity of the beam splitter is different between the front and back sides, whereas the transmissivity does not depend on the direction.[38] Figure 55 shows an example of a fully symmetric beam combiner, based on a classical Mach-Zehnder interferometer. Note that each of the beams emanating at the balanced outputs (shown as solid heave arrows) undergoes two reflections at a mirror, one reflection at the front side of the beam splitter (r), and one transmission through the beam splitter. The transmissions are in different directions (front side first (t), and back side first (t'), respectively), but this doesn't matter since $t = t'$. The combination of the field inverter (Fig. 54) and the beam combiner (Fig. 55) is thus fully symmetric by design; the null depth is only limited by imperfections of the coatings, and the quality and alignment of the optical elements.

A variation on the theme of geometric field reversal is the *rotational shearing interferometer*. Here the functions of field reversal and beam combination are not separated, but are both done in a modified Michelson interferometer (Serabyn 1999; Serabyn et al. 1999). The field reversal can for example be performed by rooftop (i.e., V-shaped) mirrors in the two interferometer arms. One of the rooftops is oriented vertically, the other horizontally, so that the two beams are rotated by 180° with respect to each other (see Fig. 56).[39] Although this arrangement is not fully symmetric (one of the beams undergoes

[38] This is the "Left-and-Right Incidence Theorem", see e.g. Knittl (1976).

[39] Take two identical copies of a written page. Flip one of them around a horizontal axis, the other around a vertical axis. The two pages now have a relative rotation of 180°.

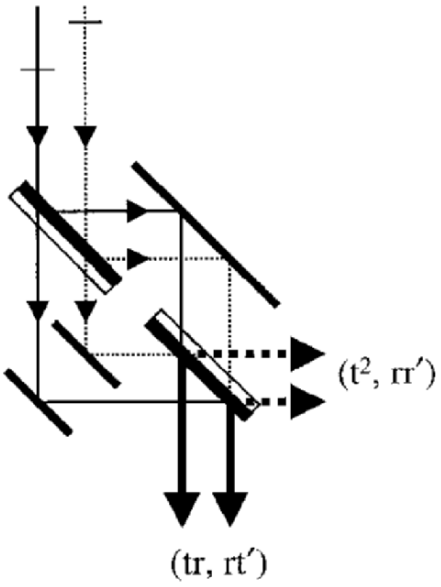

Fig. 55. Symmetric beam combiner derived from a classical Mach-Zehnder interferometer. At zero optical path difference, constructive interference occurs at the balanced outputs (shown as solid heavy arrows). In conjunction with a prior field flip these balanced outputs become nulled outputs at zero OPD. The pair of markers on the input beams indicate the wavefront offset needed for pathlength matching at the outputs. From Serabyn and Colavita (2001)

a front-side (r) reflection, the other back-side (r') reflection), it has been used to demonstrate a $\overline{N} = 7 \cdot 10^{-5}$ null in the laboratory with broadband visible light (Wallace et al. 2000).

Asymmetric Passage through Focus and Use of Berry's Phase

An alternative way of introducing a geometric phase shift of 180° is the introduction of a focus in one arm of the interferometer (Gay and Rabbia 1996). This method should in principle work well, but because of its intrinsic asymmetry it is probably more difficult to achieve a very deep null in this way than by symmetric field inversion.

Yet another geometric method is based on an effect known as Pancharatnam's phase or Berry's phase (Pancharatnam 1956). The essence of this phenomenon is that converting two beams from identical initial polarization states to identical final polarization states via different intermediate states (with a suitable arrangement of polarizers, quarter-wave plates etc.) in general results

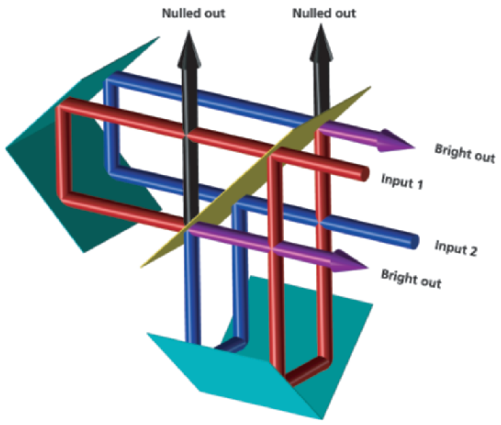

Fig. 56. Schematic layout of the beam paths in a rooftop-based rotational shearing interferometer. As a result of the double-pass beam splitter, the two input beams to be combined (red and blue) yield four output beams, two of which (black) are nulled at zero optical path difference and two of which add constructively (violet). From Serabyn (1999)

in a phase shift between the two beams.[40] Nulling with Pancharatnam's phase has been demonstrated experimentally in the laboratory; the most important technical challenge is obtaining fully achromatic wave plates (Baba et al. 2001).

Phase Shift with Dispersive Elements

If we had a plane-parallel plate of thickness s, made of a material with strictly linear dispersion, $n(\lambda) = n_0 + a\lambda$, we could use it to create an achromatic phase shift. Inserting the plate in one of the interferometer arms, and compensating for the fixed delay $D_0 \equiv (n_0 - 1) \cdot s$, the resulting phase would be

$$\phi = \frac{2\pi}{\lambda} \cdot a\lambda \cdot s = 2\pi a s \ , \tag{214}$$

which is obviously independent of λ. Unfortunately real optical materials do not have linear dispersion; it is therefore not possible to obtain a truly achromatic null in this way. For a plate with refractive index $n(\lambda)$, the general expression for the phase is

$$\phi(\lambda) = \frac{2\pi}{\lambda} \cdot \Big[D + \big(n(\lambda) - 1\big) \cdot s \Big] \ , \tag{215}$$

[40] The acquired phase difference is equal to $-1/2$ times the area enclosed by the two paths of the polarization state on the Poincaré sphere (for definition see e.g. Born and Wolf 1997). Pancharatnam's phase is closely analogous to the well-known Aharonov–Bohm effect, according to which two electron beams acquire a relative phase proportional to the magnetic flux they enclose (Berry 1987).

where D is some extra pathlength introduced by the delay line. If we pick two wavelengths λ_1, λ_2 and a desired phase shift $\tilde{\phi}$, (215) becomes a system of two equations with two free variables D and s, which can thus be chosen such that $\phi(\lambda_1) = \phi(\lambda_2) = \tilde{\phi}$.[41] The residual phase error $\phi(\lambda) - \tilde{\phi}$ at $\lambda \neq \lambda_1, \lambda_2$ is then due to the second order curvature of the dispersion $n(\lambda)$. This concept can easily be generalized to an arbitrary number N of plates made of different materials, for which

$$\phi(\lambda) = \frac{2\pi}{\lambda} \cdot \left[D + \sum_{i=1}^{N} \big(n_i(\lambda) - 1\big) \cdot s_i \right] , \tag{216}$$

which allows for perfect phase adjustment at $N + 1$ distinct wavelengths, $\phi(\lambda_j) = \tilde{\phi}$ for $j = 1 \ldots N + 1$. One can choose the λ_j optimally spaced across the bandpass of interest, and thus get $\phi(\lambda) \approx \tilde{\phi}$ with very small errors, because now the dispersion can be balanced up to order N in the Taylor expansion of the $n_i(\lambda)$. Many different glasses are available at visible wavelengths to implement this approach, but the choices in the mid-infrared are much more limited. Nevertheless, just two materials (ZnSe and ZnS) are sufficient to achieve a nearly achromatic phase shift consistent with a 10^{-5} null across a $7\,\mu\text{m} \ldots 19\,\mu\text{m}$ bandpass (Morgan et al. 2000). A nulling interferometer working in the $8\,\mu\text{m} \ldots 13\,\mu\text{m}$ band has been tested on the Multiple Mirror Telescope (Hinz et al. 1998). In this instrument a single ZnSe plate was used for the phase shift. This would ideally have allowed to reach a null depth $N \approx 10^{-4}$, but the observed null depth was limited to only 0.06 by atmospheric turbulence.

One obvious difficulty with the plane-parallel plate setup are manufacturing tolerances; one needs plates of precisely prescribed thickness. It is possible, however, to adjust the effective thickness by introducing a slight tilt of the plate (Mieremet et al. 2000).[42] The delay $D(\lambda)$ introduced by a tilted plate can be calculated by referring to Fig. 57; we get

$$\begin{aligned} D(\lambda) &= n(\lambda)\,\overline{AB} + \overline{BC} - \overline{AD} \\ &= \frac{n(\lambda)s}{\cos\beta} + s\sin\alpha(\tan\alpha - \tan\beta) - \frac{s}{\cos\alpha} \\ &= \frac{n(\lambda)s}{\cos\beta}\left(1 - \sin^2\beta\,\right) + \frac{s}{\cos\alpha}\left(\sin^2\alpha - 1\right) \\ &= s\sqrt{n^2(\lambda) - n^2(\lambda)\sin^2\beta} - s\cos\alpha \\ &= \left(\sqrt{n^2(\lambda) - \sin^2\alpha} - \cos\alpha\right) s\,. \end{aligned} \tag{217}$$

[41] Note that we can make D or s negative by introducing the extra delay or the plate in the other arm of the interferometer. For symmetry reasons, one should in any case put a plate in each arm; in that case s is the thickness difference between these plates.

[42] Note that there is a typo in (10) of Mieremet et al. (2000). It should read $W = n_1 \cos(\alpha - \beta) d_1 / \cos(\alpha)$.

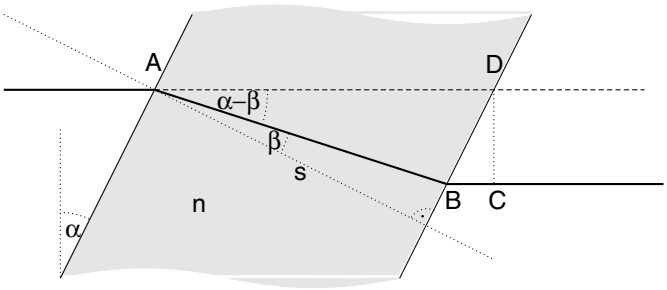

Fig. 57. Light path through a tilted plate with refractive index n. The angles α and β are related by Snell's Law $\sin\alpha = n\sin\beta$

Expansion of this expression in powers of the tilt angle α gives for $\alpha \ll 1$

$$D = (n-1)\left(1+\frac{1}{2}\frac{\alpha^2}{n}\right)s \quad , \quad \frac{dD}{d\lambda} = \left(1+\frac{1}{2}\frac{\alpha^2}{n^2}\right)s\,\frac{dn}{d\lambda}\,; \tag{218}$$

this means that tilting the plate affects not only its effective thickness but also the effective dispersion.

The phase shifter can also be implemented with a pair of prisms in place of the plane-parallel plate (Bokhove et al. 2002). The pathlength through a pair of prisms (see Fig. 58) is

$$\begin{aligned}
D(\lambda) &= n(\lambda)\,\left(\overline{AB}+\overline{CD}\right)+\overline{BC} \\
&= n(\lambda)\,(s-h\tan\alpha)+g\frac{\cos\alpha}{\cos\beta} \\
&= n(\lambda)\left(s-g\frac{\cos\alpha}{\cos\beta}\sin(\beta-\alpha)\tan\alpha\right)+g\frac{\cos\alpha}{\cos\beta} \\
&= n(\lambda)\left(s+g\frac{\sin\alpha}{\cos\beta}(\sin\alpha\cos\beta-\cos\alpha\sin\beta)\right)+g\frac{\cos\alpha}{\cos\beta} \\
&= n(\lambda)\left(s+g\sin^2\alpha\right)-g\frac{\cos\alpha\sin^2\beta}{\cos\beta}+g\frac{\cos\alpha}{\cos\beta} \\
&= n(\lambda)\left(s+g\sin^2\alpha\right)+g\cos\alpha\cos\beta \\
&= n(\lambda)\left(s+g\sin^2\alpha\right)+g\cos\alpha\sqrt{1-n^2(\lambda)\sin^2\alpha}\,.
\end{aligned} \tag{219}$$

The pair of prisms is thus equivalent to two plane-parallel plates; one with thickness s and index $n(\lambda)$, and one with thickness g and index

$$n_e \equiv n(\lambda)\sin^2\alpha+\cos\alpha\sqrt{1-n^2(\lambda)\sin^2\alpha}\,. \tag{220}$$

This is a remarkable result, because we now have effectively twice as many materials at our disposal to flatten the phase across the bandpass, a significant

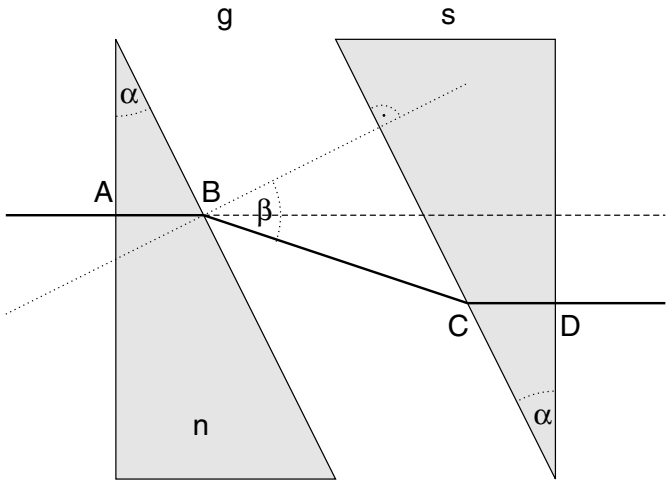

Fig. 58. Light path through a pair of prisms with refractive index n. The angles α and β are related by Snell's Law $n \sin \alpha = \sin \beta$

advantage especially in the mid-infrared, where only very few optical materials are available. Dispersive elements appear thus equally suited for nulling interferometry as the geometric methods discussed above, because they can generate phase shifts that are close enough to achromatic. In addition, dispersive elements can produce any desired phase shift (not only π rad), which is required in some classes of multi-element nulling arrays (see Sect. 10.4). An additional degree of freedom can sometimes be used to minimize achromatic effects, namely the possibility to use a (nominal) phase shift of 2π instead of 0 (Mieremet and Braat 2002b). For example, a three-telescope configuration with relative phases of (0, π, 0) is equivalent to one with phases of (0, π, 2π), but the achromatic errors are different between the two cases.

10.3 Nulling Interferometry from the Ground

Strehl Fluctuations, and Tip-Tilt Accuracy

We will now consider a simple Bracewell nulling interferometer on the ground, and we assume that the two telescopes are equipped with adaptive optics systems, which correct the beams from the two telescopes with a Strehl ratio S. From (199) and (123) we obtain

$$\overline{N} = \frac{1}{2}\sigma_w^2 \approx -\frac{1}{2}\ln S \approx \frac{1}{2}(1-S) \ , \tag{221}$$

which means that the AO systems have to provide $S = 0.998$ to achieve a 10^{-3} null. This is hardly feasible even in the thermal infrared, which means

that modal filtering of the wavefront is necessary. We will therefore analyze performance requirements for the AO system to reach a specified null depth under the assumption that the nulling beam combiner contains a single-mode modal filter. Using (123), (190), and (200), we see that we can then write the average null depth as

$$\overline{N} = \frac{\left\langle (S_1 - S_2)^2 \right\rangle}{16 \left\langle S^2 \right\rangle} \ , \tag{222}$$

where S_1 and S_2 are the Strehl ratios produced by the two AO systems, and $\left\langle S^2 \right\rangle$ the mean squared Strehl ratio (assumed to be equal for the two AO systems). This means that the null depth depends primarily on the stability of the wavefront correction, not so much on the actual value of $\left\langle S^2 \right\rangle$. In practice, the stability and the quality of AO correction are closely linked to each other, however. For strictly stationary Kolmogorov turbulence it is possible to calculate the expected variability of the AO performance (Yura and Fried 1998), but in practice the non-stationary nature of turbulence is probably the limiting factor.[43] Slow variations of the seeing produce equal variations of the Strehl ratio in both telescopes, but fast variations must be uncorrelated over distances of ∼100 m. These fast Strehl fluctuations are difficult to measure with most AO systems, but it is plausible to assume that an AO system producing a mean Strehl ratio of 0.5 actually provides an instantaneous Strehl ratio fluctuating between 0.3 and 0.7. We can therefore estimate $\sqrt{\langle (S_1 - S_2)^2 \rangle} \approx 0.2$ and get $\overline{N} \approx 0.01$ from (222). This is not very satisfactory, but fortunately the null depth is $\propto \lambda^{-4}$ as pointed out above. If we can build an AO system that achieves the above performance in the K band (2.2 µm), it will provide a mean Strehl ratio of 0.97 and a null depth of $2 \cdot 10^{-5}$ at 10 µm (see Table 13).

The requirements on the quality of the tip-tilt correction follow immediately from (208). The angle tracker has to follow not only the atmospheric image motion, but also telescope vibrations induced by wind shake. The performance should be independent of the observing wavelength. If the residual tip-tilt fluctuation are of order 10 mas (which is realistic for the 10 m Keck telescopes), we get $\overline{N} \approx 3.6 \cdot 10^{-3}$ in the K band, about a factor of three smaller than the effect of Strehl ratio fluctuations. For the 8 m telescopes of the VLTI, $\overline{N} \approx 1.5 \cdot 10^{-3}$. Since again $\overline{N} \propto \lambda^{-4}$, the ratio between the effects of AO compensation and tip-tilt correction is the same for all wavelengths. We conclude that on an 8 m to 10 m telescope a tip-tilt servo loop that provides 10 mas rms angle tracking is well-matched to an AO system that delivers a Strehl ratio of 0.5 in the K band.

[43] In practice the seeing is observed to vary on all accessible time scales. "Lucky moments" are interrupted by spells of bad seeing, which may last only seconds. Real-time seeing monitors typically average over a few minutes to get representative figures, yet they frequently report variations of r_0 by sizeable amounts within less than one hour.

Table 13. Expected null depth for the near-infrared bands

band	wavelength	$\langle S \rangle$	$\overline{N}_{\rm AO}$	$\overline{N}_{\rm OPD}$	$\overline{N}^{0}_{\rm scint}$	$\overline{N}^{8{\rm m}}_{\rm scint}$
K	2.2 µm	0.50	$1.0 \cdot 10^{-2}$	$3.8 \cdot 10^{-3}$	$4.4 \cdot 10^{-3}$	$7.4 \cdot 10^{-7}$
L	3.6 µm	0.77	$1.4 \cdot 10^{-3}$	$1.4 \cdot 10^{-3}$	$2.5 \cdot 10^{-3}$	$7.4 \cdot 10^{-7}$
M	5.0 µm	0.87	$3.9 \cdot 10^{-4}$	$7.4 \cdot 10^{-4}$	$1.7 \cdot 10^{-3}$	$7.4 \cdot 10^{-7}$
N	10.0 µm	0.97	$2.4 \cdot 10^{-5}$	$1.9 \cdot 10^{-4}$	$7.5 \cdot 10^{-4}$	$7.4 \cdot 10^{-7}$
Q	20.0 µm	0.99	$1.5 \cdot 10^{-6}$	$4.6 \cdot 10^{-5}$	$3.3 \cdot 10^{-4}$	$7.4 \cdot 10^{-7}$

The following assumptions have been made (for details see text). AO system: mean Strehl at K-band = 0.5, r.m.s. Strehl imbalance at K-band = 0.2. Atmosphere: f_G = 21.35 Hz at 500 nm, one layer at 10 km with r_0 = 0.2 m at 500 nm for scintillation calculation. Fringe tracker: 2 ms loop lag

Fringe Tracking

Nulling interferometry requires precise fringe tracking. The null depth due to residual uncorrected high-frequency fluctuations of the optical path difference can be calculated

$$\overline{N} = \frac{1}{4}\sigma_R^2 = \frac{1}{2}\kappa \left(\frac{f_G}{f_S}\right)^{5/3} . \tag{223}$$

A careful modeling of the dynamical behavior of the control loop is needed to determine the value of κ. To get a feeling for the numerical values involved, consider the case of a "pure delay" of 2 ms, so that $\kappa = 28.4$, and f_S = 500 Hz. Let's further assume that at 500 nm the Greenwood frequency $f_{G,500}$ = 21.35 Hz; this corresponds to a wind velocity $v = 10\,{\rm m\,s^{-1}}$ and Fried parameter $r_{0,500} = 20$ cm. Scaling f_G to 2.2 µm and inserting into (223), we obtain a K-band null depth of $3.8 \cdot 10^{-3}$. We see that the speed of the control loop must be much faster than the Greenwood frequency to obtain a good null depth; lags due to detector readout or processing time are particularly damaging because of the large value of κ associated with them. The wavelength scaling of high-frequency fringe tracking residuals is $\overline{N} \propto \lambda^{-2}$.

Photon noise in the fringe tracker is also a source of phase noise, which becomes important for faint stars. The null depth due to photon noise from (157) and (190) is

$$\overline{N} = \frac{1}{4}\left(\frac{\lambda_t}{\lambda_n}\right)^2 \sigma^2_{\phi,t,{\rm phot}} = \left(\frac{\lambda_t}{\lambda_n}\right)^2 \cdot \frac{1}{2N_t V_t^2} , \tag{224}$$

where λ_t is the wavelength at which the fringe tracking is performed, and λ_n the wavelength at which the nuller operates. The factor $(\lambda_t/\lambda_n)^2$ in (224) favors fringe tracking at a short wavelength. This means that one has to be careful about additional errors due to dispersion between λ_t and λ_n. One

possibility is using two nested control loops. This scheme uses a fast servo at λ_t, which tracks the rapid atmospheric phase fluctuations, and an "outer" loop with a sensor that measures the residual phase at λ_n, and determines tracking offsets between the two wavelengths with a much slower update rate.

Another implication of (224) is the degradation of the quality of the null when the star is partially resolved at the tracking wavelength. If we require $V_t^2 \gtrsim 0.03$, the stellar diameter has to be $\lesssim \lambda_t/B$. The numerical value for a 50 m baseline at K band is 10 mas, which is exceeded only by a relatively small number of bright cool giants and Mira stars. On much longer baselines or at much shorter λ_t, however, many of the nearby main-sequence stars will have low visibilities. It thus appears reasonable to perform the fringe tracking at K band. In this band a 0 mag star observed with two 8 m telescopes with a 10% total efficiency generates $\sim 2 \cdot 10^{10}$ photons per second. Scaling this value to $K = 12$ gives 320 photons per millisecond. If the AO systems achieve a Strehl ratio of 0.5, half of these photons are rejected by the modal filter. The corresponding K-band null depth from (224) is $3.1 \cdot 10^{-3}$; it scales with $\overline{N} \propto \lambda^{-2}$. Consequently, the nulling performance should not depend much on the stellar brightness down to $K \approx 12$, if the fringe sensor remains photon-noise limited down to that magnitude.

Scintillation Effects

While the atmospheric phase fluctuations will be corrected to a large extent by an adaptive optics system, it may be difficult to implement an intensity control that works on similar time scales. In that case atmospheric scintillation creates a rapidly variable imbalance between the two beams. The consequent limitation on the null depth can be calculated combining (110) and (134)

$$\sigma^2_{\ln I} = 1.14 \left(\frac{\sqrt{\lambda h \sec z}}{r_0(h)} \right)^{5/3} \tag{225}$$

and by inserting it in (190); this gives[44]

$$\overline{N} = 0.14 \left(\frac{\sqrt{\lambda h \sec z}}{r_0(h)} \right)^{5/3} . \tag{226}$$

A layer at 10 km above the observatory with $r_0 = 20$ cm at 500 nm will thus produce a K-band null depth of $4.4 \cdot 10^{-3}$. The improvement with wavelength, $\overline{N} \propto \lambda^{-7/6}$, is the slowest of all effects considered.

[44] A little care has to be taken about factors of 2 here. A fluctuation δI in one beam increases the average intensity I; the two beams thus have intensities $I + \delta I/2$ and $I - \delta I/2$. So the variance in (190) is 1/4 of the variance in (225). Multiplying this by 2 for the contributions from the two telescopes we get the numerical factor in (226).

In the derivation of (226) we have made use of (134), which is valid for the intensity fluctuations at any given point of the wavefront. This means that (226) is applicable if the interferometer consists of two very small telescopes, or if a classical co-axial beam combiner is used.[45] If, on the other hand, a single-mode fiber is used in the beam combiner, the contributions from different parts of the pupil are mixed, and it is the intensity averaged over the telescope aperture that matters for the null depth. The scintillation pattern varies spatially on the Fresnel scale $r_F \equiv \sqrt{\lambda h \sec z}$. For telescopes much larger than r_F most of the fluctuations represent high-order modes that are rejected by the modal filter; only variations of the average intensity couple efficiently into a single-mode fiber. According to Ryan (2002), aperture averaging can be approximately described by multiplying the right-hand side of (134) or (225) by a factor

$$A \approx \left[1 + 1.1 \left(\frac{D^2}{\lambda h \sec z}\right)^{7/6}\right]^{-1} . \tag{227}$$

This leads to a null depth of $7.4 \cdot 10^{-7}$ independent of λ for $D = 8\,\mathrm{m}$, so that scintillation should be negligible for nulling with the VLTI UTs or the Keck 10 m telescopes.

Subtraction of the Thermal Background

Infrared nulling observations from the ground have to cope with the thermal background radiation of the atmosphere and instrument. This can in principle be done by the standard chopping and nodding techniques that are widely used in infrared astronomy (e.g., McLean 1997). This is technically not easy, however, because it means that the chopping devices (usually the telescope secondaries), adaptive optics systems, and the fringe tracker have to be synchronized. The adaptive optics and fringe tracking loops have to be opened during the off-source part of the chopping cycle; at the beginning of the on-source part, the adaptive optics loops have to be closed, the fringes have to be re-acquired and the servo has to find the zero optical path difference position before data taking can resume. Since canceling the atmospheric fluctuations requires chopping frequencies of several Hz, this all has to be done very quickly to keep the resulting overhead tolerable.

A more efficient way of dealing with the thermal background is creating an internal modulation of the interferometric signal (see Fig. 59). For this purpose the apertures of the two telescopes are split in halves, so that four beams are sent to the beam combination laboratory. The right halves and the

[45] Recall that in co-axial beam combiner a 50% beam splitter is used to superpose the pupils from the two telescopes. Light from each point of the first pupil therefore interferes only with light from the corresponding point of the second pupil.

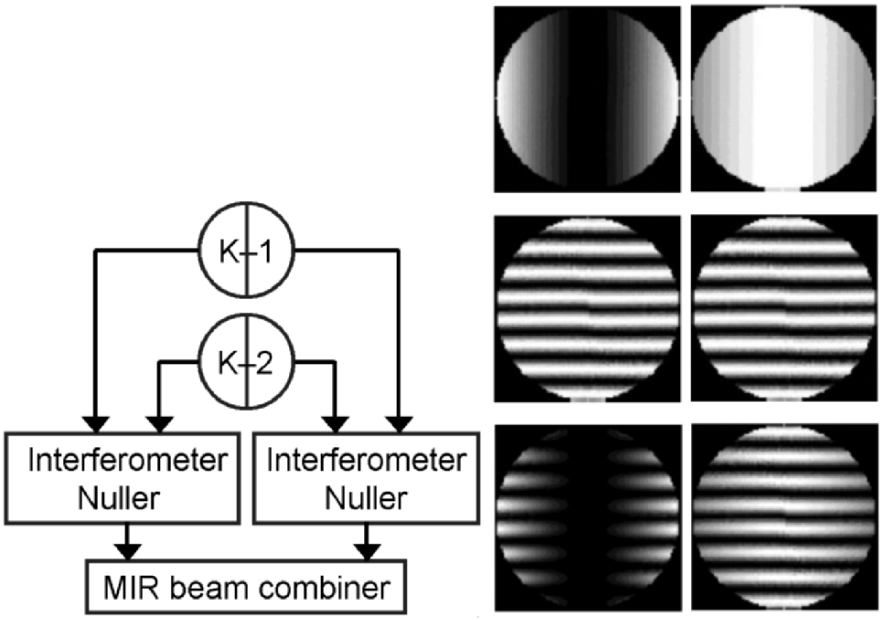

Fig. 59. Schematic setup (left) and transmission maps of the Keck Interferometer nulling experiment. Each of the two telescope apertures is split into two halves; the right halves and the left halves form two nulling interferometers. The outputs of these nulling interferometers are combined on a third nuller. Modulating the optical path difference of this aperture nuller by $\lambda/2$ alternates between the transmission maps shown in the top panels. Multiplying by the transmission maps of the two nulling interferometers (center panels) gives the resulting transmission maps shown in the bottom panels. A point source on the optical axis is rejected at all times, while the light from an extended source (e.g., dust disk) or off-axis companion produces a modulated signal on top of the constant thermal background

left halves form two nulling interferometers with baselines B, which reject the light of an on-axis star. The angular scale of these nulls is therefore λ/B. The two outputs are combined on a third nuller, which interferes the light from the sub-aperture pairs and therefore creates a null with angular scale λ/D. Modulating the optical path difference of this aperture nuller by $\lambda/2$ creates the desired modulated signal. Whereas the light from an on-axis point source is rejected at all times by the λ/B nuller, and the incoherent background is not affected by the modulation, any source at a separation between λ/B and λ/D (or of corresponding size) will alternatingly be passed an rejected. For stars at $d = 20\,\mathrm{pc}$, these two numbers correspond to 0.4 AU and 4 AU for an interferometer with two 10 m telescopes on a 100 m baseline, operated at 10 µm. This setup is therefore well-suited for the detection of dust disks and planets around nearby stars.

10.4 Design of Nulling Arrays

Quadratic and Higher-Order Nulling

We have seen above (213) that the non-zero diameter of the stellar disk causes a leak, which severely limits the null depth that can be obtained on bright stars. The reason is that the null depth is proportional to the square of the off-axis angle (see (186) and (210)). The only way around this problem is creating a higher-order null, which is possible if the light from more than two telescopes is combined with proper phase shifts. Let us consider an array consisting of n telescopes with diameters D_k, located at positions with polar coordinates (L_k, δ_k). If a phase shift ϕ_k is introduced in the beam of telescope k before beam combination, the complex amplitude for a point source with polar coordinates (θ, ψ) is given by

$$A_{\text{out}} = \sum_{k=1}^{n} D_k \cdot e^{2\pi i (L_k \theta/\lambda) \cos(\delta_k - \psi)} \cdot e^{i\phi_k} \,. \tag{228}$$

We are interested in the scaling of A_{out} for small off-axis angles; therefore we expand the exponential into a Taylor series in θ. With the abbreviation $x_k \equiv 2\pi(L_k\theta/\lambda)$ we obtain

$$e^{ix_k \cos(\delta_k - \psi)} = 1 + ix_k \cos(\delta_k - \psi) - \tfrac{1}{2} x_k^2 \cos^2(\delta_k - \psi) + \mathcal{O}(x_k^3) \,. \tag{229}$$

Back-substituting this expansion into (228) we see that the condition for on-axis nulling ($A_{\text{out}} = 0$ for $\theta = 0$) is

$$\sum_{k=1}^{n} D_k \cdot e^{i\phi_k} = 0 \,. \tag{230}$$

The complex amplitude is then $\propto \theta$ and the intensity leaking through the nuller $\propto \theta^2$. Setting the second term of the expansion (228) and (229) to zero we get

$$\sum_{k=1}^{n} D_k \cdot x_k \cdot \cos(\delta_k - \psi) \cdot e^{i\phi_k} = 0 \,. \tag{231}$$

If (230) and (231) are satisfied simultaneously for all values of ψ, the intensity varies $\propto \theta^4$. This result can easily be generalized by adding equations from higher orders of the expansion; the additional condition to achieve a θ^6 null is

$$\sum_{k=1}^{n} D_k \cdot x_k^2 \cdot \cos^2(\delta_k - \psi) \cdot e^{i\phi_k} = 0 \,. \tag{232}$$

Using the standard formula for the cosine of a difference, we see that requiring (231) to be satisfied for all values of ψ is equivalent to the two conditions

$$\sum_{k=1}^{n} D_k \cdot x_k \cdot \cos\delta_k \cdot e^{i\phi_k} = 0, \tag{233}$$

$$\sum_{k=1}^{n} D_k \cdot x_k \cdot \sin\delta_k \cdot e^{i\phi_k} = 0\,. \tag{234}$$

Similarly, (232) is equivalent to

$$\sum_{k=1}^{n} D_k \cdot x_k^2 \cdot \cos^2\delta_k \cdot e^{i\phi_k} = 0, \tag{235}$$

$$\sum_{k=1}^{n} D_k \cdot x_k^2 \cdot \sin^2\delta_k \cdot e^{i\phi_k} = 0, \tag{236}$$

$$\sum_{k=1}^{n} D_k \cdot x_k^2 \cdot \sin 2\delta_k \cdot e^{i\phi_k} = 0\,. \tag{237}$$

Equations (230), (233), and (235) provide a systematic framework for the design of arrays that perform quadratic, fourth-order, or sixth-order nulling. An example for a linear array that generates a sixth-order null is the OASES concept (Angel and Woolf 1997). The transmission of this configuration is compared to that of a two-element Bracewell interferometer in Fig. 60.

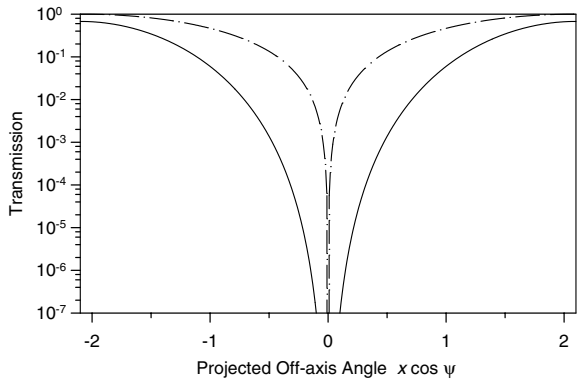

Fig. 60. Transmission of the linear OASES interferometer concept consisting of four telescopes with diameters (1, 2, 2, 1) located at positions $(-2, -1, 1, 2)$, and combined with phases (0, π, 0, π). The horizontal axis is the component of the off-axis angle parallel to the array. The θ^6 null of OASES (full line) is compared to the θ^2 null of a two-telescope Bracewell interferometer 3/8 as long, for which the first maximum in the transmission occurs at the same off-axis angle (*dash-dotted line*)

External and Internal Modulation

The next important question to consider is the way of extracting useful information from the nulling array. A pupil plane interferometer does not produce images, but rather a photon count from a single-pixel detector as a function of time. The output of the beam combiner for light arriving from the direction (θ, ψ) is given by (228); the total amplitude received can therefore be computed by multiplying this expression with the electric field at (θ, ψ), and integrating over the plane of the sky. This process is illustrated in the top panels of Fig. 53; the observed photon count is the total integrated intensity contained in the top right panel. This single number obviously does not provide us with the information we are seeking, namely the position and brightness of the planet(s) associated with the target star. The simplest way of increasing the amount of information from the nulling interferometer is to rotate it around the line-of-sight toward the star, as illustrated in the bottom panels of Fig. 53. In the course of the rotation the planet is traversed by regions of high and low response, which leads to a modulation of the observed intensity. The information about the position and brightness of the planet is encoded in this signal, which is plotted as a function of time for a full rotation in the bottom right panel.

The rotational modulation method has a significant difficulty, namely the requirement that the instrumental response has to remain stable over a full rotation period, which is typically of order several hours. Small drifts of the detector sensitivity or background, of the telescope pointing, or other components of the interferometer can mask the presence of a planet or lead to spurious detections. It is therefore highly desirable to modulate the output at a higher rate ($\approx 1\,\mathrm{Hz}$). In a "standard" Michelson stellar interferometer one can modulate the delay and measure the visibility with a lock-in detection scheme as described in Sect. 8.5. This is generally not possible for a nulling interferometer, because delay scanning changes the ϕ_k and thus violates the nulling conditions. It is obvious from (230), however, that four identical telescopes combined with phases (0, π, α, and $\alpha + \pi$) will produce a central null irrespective of the value of α. One can, for example, quickly alternate between $\alpha = \pi/2$ and $\alpha = -\pi/2$, which generally changes the interferometer response pattern on the sky (see Fig. 61). This technique of modulating the planetary signal is called "internal chopping".

We can think of this configuration as constructed from two Bracewell pairs with phases (0, π) and (α, $\alpha+\pi$), respectively. This concept can be generalized to arbitrary nulling configurations. If the outputs from N arrays, each of which produces a θ^μ null, are combined with relative phases $\alpha_1, \ldots, \alpha_N$, the output produces again a θ^μ null, and the α_i can be varied to produce a modulation of the signals from off-axis sources. This recipe can be used to construct complicated nulling arrays with desirable properties from simpler building blocks. We should also note that compressing or stretching an array in one direction changes the width of the null in that direction, but preserves the order of the null.

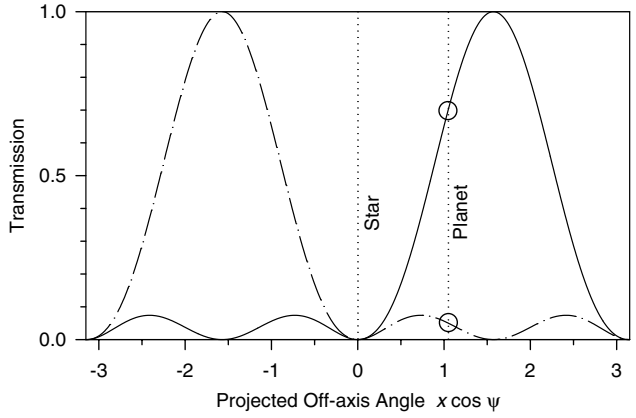

Fig. 61. Transmission of a linear double-Bracewell interferometer consisting of four telescopes of identical diameter with unit spacing. The horizontal axis is the component of the off-axis angle parallel to the array. The *full curve* is the transmission for phases $(0, -\pi/2, \pi, \pi/2)$, the *dash-dotted curve* for $(0, \pi/2, \pi, -\pi/2)$. Chopping between these states modulates the planetary signal between the two values indicated by *circles*

Nulling Array Geometries

Various array geometries have been proposed for the direct detection of extrasolar planets in the mid-infrared, each one with certain technical advantages and disadvantages. If all telescopes are mounted on a single long structure, linear configurations are certainly easier to assemble and maneuver. On the other hand, two-dimensional arrays in which each telescope is mounted on its own free-flying spacecraft have more versatility. In that case, an "ideal" nulling array should perhaps fulfill the following criteria:

- All telescopes are at equal distance from the beam combiner spacecraft. If this is not the case, the beam transport is more complicated, and diffraction effects will be different between the interferometer arms.
- All telescopes and the beam combiner are in one plane (perpendicular to the line of sight to the target star). This allows a thermal design that minimizes radiative coupling between the illuminated and cold parts of the spacecraft.
- The null is of fourth or sixth order. We have seen above that second-order starlight suppression is not sufficient.
- The only phase shift needed is π. While it is possible to introduce shifts by other amounts with dispersive elements, they are more difficult to control precisely.
- All telescopes have the same diameter. This is certainly cheaper than launching telescopes with different sizes.

- Internal phase modulation is possible. This relaxes the required stability of the optics and detectors.
- The transmission function should not have any rotational symmetries, which lead to ambiguities in the position of any detected planets. (For example, linear configurations with mirror-symmetric transmission functions have a 180° degree ambiguity.) The determination of planet orbits from multiple observations is much easier if the positions are unambiguous, especially in multi-planet systems.

As we will see, it is possible to design a arrays that have all of these properties. The price to be paid is in the number of telescopes and in the complexity of the beam combination, when a high-order null and chopping capability are required. Since the more complicated arrays are constructed from simpler building blocks, it is useful to look at the various arrangements, progressing from the elementary to the more complex. A summary is also given in Table 14. All concepts assume that the whole interferometer is oriented perpendicular to the line-of-sight to the target star, i.e., there has to be a spacecraft maneuver to reorient the array for each observed star. The array can then be rotated around the line-of-sight during the observation, either by a full 360° or a smaller angle, as shown in Fig. 53.

Bracewell Pair.

This is the most basic nulling interferometer, consisting of two telescopes with equal size and a π phase shift between them. The resulting θ^2 null is insufficient for the direct detection of Earth-like planets.

Table 14. Configurations for space-borne nulling interferometers

name	N_{tel}	configuration	order	chopping	ambiguities
Bracewell	2	Single Baseline	θ^2	no	yes
Double Bracewell	4	Linear 1:1:1:1	θ^2	yes	no
OASES	4	Linear 1:2:2:1	θ^6	no	yes
Angel's Cross	4	Cross-shaped	θ^4	no	yes
DAC	3	Linear 1:$\sqrt{2}$:1	θ^4	no	yes
Double DAC	4	Linear 1:$\sqrt{3}$:$\sqrt{3}$:1	θ^4	yes	no
Mariotti 3-DAC	6	Triangular	θ^4	yes	no
DARWIN 5-Telescope	5	Compressed Pentagon	θ^4	no	no
Robin Laurance	6	Hexagon	θ^4	yes	no

For explanations of the individual configurations see text. Note that variants of many of the concepts exist; only the main version for each one is included in the table

Double Bracewell Interferometer.

This is the simplest nulling interferometer that allows for internal chopping. It consists of four identical telescopes in a linear arrangement with equal spacing. The first and third telescope form one Bracewell pair, the second and fourth the other. Switching the relative phase between the two Bracewell pairs phases between $\pi/2$ and $-\pi/2$ is illustrated in Fig. 61. The double Bracewell interferometer inherits the θ^2 null from the single Bracewell pair.

OASES.

The OASES concept is a linear array that provides θ^6 nulling (Angel and Woolf 1997). It consists of two Bracewell pairs; one of them has twice the baseline but only half the aperture diameter of the other. The phase between the two Bracewell pairs is π. The OASES configuration can thus be described by $D_k = (1, 2, 2, 1)$, $L_k = (-2, -1, 1, 2)$, and $\phi_k = (0, \pi, 0, \pi)$.

Angel's Cross.

This is a two-dimensional configuration, in which four telescopes of equal diameter are placed at the corners of a rhombus, so that their arrangement resembles a cross with pairwise equal bar lengths (Angel 1990). The telescopes located opposite of each other on the same bar have the same phase; the phase difference between the two bars is π. Angel's cross produces a θ^4 null.

Degenerate Angel's Cross (DAC).

This configuration is derived from Angel's cross by collapsing it along one of the bars, i.e., the two telescopes of one of the bars are replaced by a single telescope with twice the aperture area. Taking a linear configuration with equal spacings, telescope diameters $(1, \sqrt{2}, 1)$, and phases $(0, \pi, 0)$, and dividing the light from the central telescope in two equal beams, we get an array of four telescopes with equal diameters at positions $(-1, 0, 0, 1)$ and with phases $(0, \pi, \pi, 0)$. This is just the projection of an Angel's cross on one coordinate axis.[46] The order of the null in the DAC configuration is also θ^4.

Double Degenerate Angel's Cross.

If we construct two DACs, we can combine their outputs with a time-variable phase and thus obtain an array with a θ^4 null and internal chopping. A double

[46] There seems to be a paradox here. According to our formalism (230), one should expect diameters of $(1, 2, 1)$, not $(1, \sqrt{2}, 1)$. Actually, both versions are "correct". Consistency with our formalism requires that if we choose $(1, 2, 1)$, we must send the light from the central telescope in one entrance of the beam combiner; if we choose $(1, \sqrt{2}, 1)$ we must divide the light from the central telescope and send the two beams into two different entrances. It is not too difficult to convince oneself that the light in the nulled output is the same in both cases; the extra photons in the $(1, 2, 1)$ case end up in the non-nulled outputs.

DAC can either be realized with six independent telescopes (Woolf and Angel 1997), or with a linear array of four equally spaced telescopes with diameters $(1, \sqrt{3}, \sqrt{3}, 1)$. The latter approach is based on the idea of overlaying separate nulling arrays such that they share one or more telescopes. The aperture area of a shared telescopes must be equal to the sum of the areas of the constituent telescopes at that position. In the case of the double DAC, splitting the light from the large telescopes with a 2:1 intensity ratio forms two arrays that are equivalent to diameters $(1, \sqrt{2}, 1, 0)$, and $(0, 1, \sqrt{2}, 1)$, respectively; these are the two constituent DACs.

Mariotti 3-DAC.

This configuration consists of three DACs, which form the three sides of an equilateral triangle. Each of the telescopes at the vertices of the triangle is shared between two DACs, so that all telescopes are of equal size. This fact, together with the θ^4 null, chopping capability, redundancy, and two-dimensional coverage of the uv plane make the Mariotti 3-DAC quite attractive for a space nulling array.

DARWIN 5-Telescope Configuration.

A five-telescope solution to (230) and (233) is given by $x_k = \text{const.}$, $\delta_k = 2(k-1)\pi/5$, and $\phi_k = 4(k-1)\pi/5$, i.e., the telescopes are located on a regular pentagon, and the phase difference between neighboring telescopes is 144° (Mennesson and Mariotti 1997). The DARWIN 5-telescope configuration is derived from this solution by compressing the array along one axis by a factor ∼2. Upon rotation of the array (internal chopping is not possible), the transmission function provides fairly strong modulation of planetary signals but only weak modulation of the light from a symmetric exozodiacal disk. Unlike most other non-chopping configurations, the DARWIN 5-telescope concept does not suffer from ambiguities of the planet position; the five-fold symmetry of the regular pentagon is broken by the compression along one axis.

Robin Laurance Interferometers.

This is a class of nulling arrays that consists of building blocks, which can be called generalized Angel's crosses (GACs); a GAC is here defined simply as a nulling interferometer that satisfies (230) and (233) and thus produces a θ^4 null. One can easily verify that assigning telescope diameters $D_k = (3, 2, 0, 1, 0, 2)$ and phases $\phi_k = (0, \pi, \times, 0, \times, \pi)$ to the six vertices of a regular hexagon defines a GAC. Three such GACs rotated by 120° with respect to each other and overlayed on the hexagon as shown in Fig. 62 constitute a Robin Laurance interferometer (Karlsson and Mennesson 2000). This configuration satisfies all of the requirements on an "ideal" nulling array listed above. The most significant drawbacks are its complexity and the need to divide the light from three of the telescopes asymmetrically with a 4:4:1 intensity ratio. Many similar overlays of GACs on regular hexagons and on pentagons exist. It is

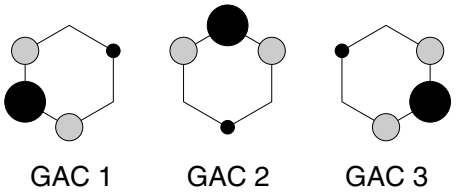

Fig. 62. Robin Laurance interferometer. Three generalized Angel's crosses (GACs) are overlayed on a hexagonal telescope array. The designation of this specific configuration RL3(3,2,0,1,0,2) gives the number of GACs, and the amplitudes contributing to each GAC, starting with the largest and counting along the hexagon's vertices. Within each GAC, the large and small telescope are combined with phase 0 (black), the two intermediate-size telescopes with phase π (light gray). The contributions from the three GACs add up to 9 units for each telescope aperture area

therefore possible to devise layouts that are technically simpler (for example solutions in which all beams have the same intensities), and to optimize various aspects of their performance (Karlsson and Mennesson 2000).

Location of the Beam Combiner

In the preceding section the various array geometries have been described in terms of the location and size of the telescopes, and of the applied phase shifts. It is also necessary, of course, to provide for a beam combiner. In the case of the Robin Laurance configuration, for example, the beam combiner should clearly be located at the center of the hexagon. The pathlength from each telescope is then equal, and the beam relay to the central hub comparatively simple. If there is no location from which all telescopes have the same distance (e.g., in the OASES and Mariotti 3-DAC configurations), the beam can be passed through several spacecraft rather than sent directly to the beam combiner. An example is shown in the bottom left panel of Fig. 53. Sending the beams from telescopes 1 and 3 to the combiner via telescope 2 ensures equality of the paths. It is thus normally necessary to provide a separate beam combiner spacecraft in addition to those carrying the telescopes.

Image Reconstruction

Like any other interferometer, a nulling array produces a non-intuitive output when a complicated object is observed. Whereas in a "standard" two-element interferometer the signal and the source structure are related by a Fourier transform (145), we can derive an analogous generalized equation for a nulling array by multiplying (228) with the electric field distribution $E(\theta, \psi)$ on the sky, and integrating over θ and ψ. Because more than two telescopes contribute to the observed signal, there is no simple Fourier relation between source

structure and data, but general requirements on the sampling of the *uv* plane, and the use of image reconstruction algorithms, apply in a similar way.

Figure 63 shows an example of these principles applied to a simulated observation of a planetary system with a linear nulling interferometer, which

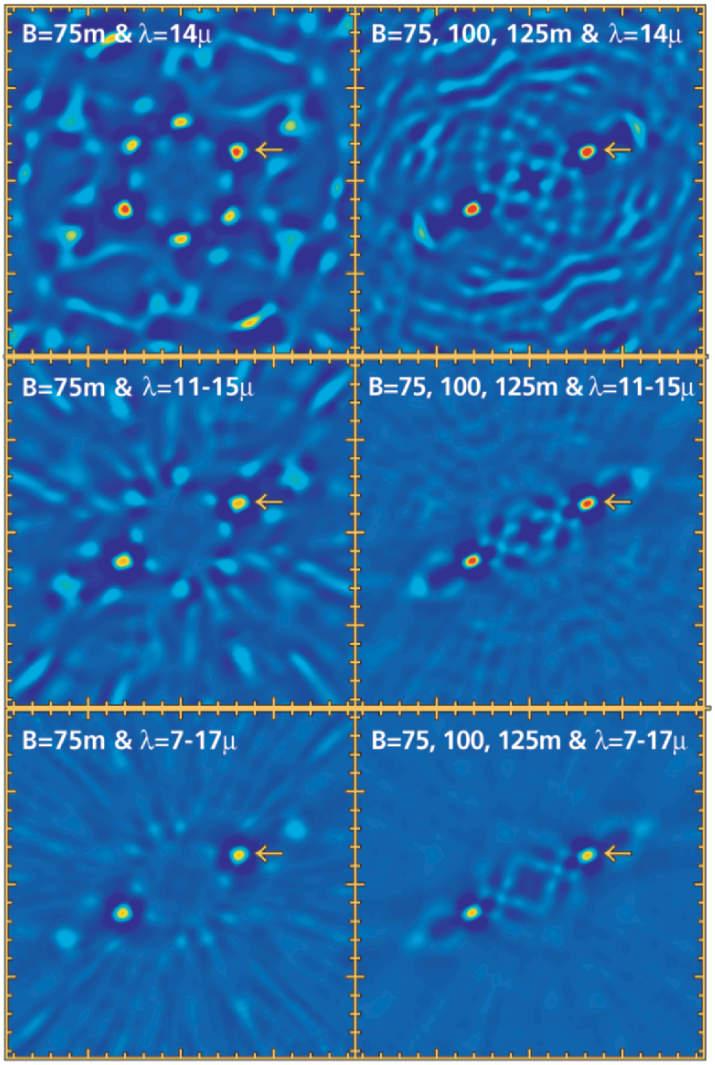

Fig. 63. Reconstruction of an observation of a terrestrial planet at a distance of 10 pc with a linear nulling interferometer. The level of exozodiacal emission was assumed to be equal to that in the Solar System; the inclination was taken to be $i = 30°$. The panels show how the reconstructed image improves as more baselines (left to right) and more wavelengths (top to bottom) are added. From NASA (1999b)

was rotated around the viewing direction to synthesize a reasonable uv plane coverage as illustrated in Fig. 53. The left-hand panels assume that the telescopes of the interferometer are fixed with respect to each other, whereas the right-hand panels assume that the size of the array can be varied. The top panels use only monochromatic light for the reconstruction, whereas the center and bottom panels take advantage of wavelength synthesis. These images demonstrate clearly how the quality of the reconstruction improves as more baselines and more wavelengths are added. A few artifacts remain, however, even in the best case. Most significantly, there is a mirror image of the planet due to the symmetric transmission pattern of the array, and an aliasing artifact at twice the orbital distance. These effects can be avoided by asymmetric array configurations or asymmetric internal chopping techniques. It is clear that a careful choice of the array geometry and of the image reconstruction technique are necessary to optimize the ability of a nulling interferometer to perform meaningful observations of systems with multiple planets.

11 Appendix: Useful Definitions and Results from Fourier Theory

For reference, this appendix lists a few useful results from Fourier theory without proofs. In the notation adopted, $g \Longleftrightarrow G$ means "G is the Fourier transform of g", and it is understood that small and capital letters designate Fourier transforms pairs, i.e., $g \Longleftrightarrow G$ and $h \Longleftrightarrow H$. H^* is the complex conjugate of H. Introductions into Fourier theory and more details can be found in many textbooks (e.g. Bracewell 1965). The results are frequently formulated for the one-dimensional Fourier pair time and frequency ($t \longleftrightarrow f$), but they can equally be applied to the three-dimensional variables position and spatial frequency ($x \longleftrightarrow \kappa$).

The *convolution* $g * h$ and *correlation* $\mathrm{Corr}(g, h)$ of two functions g and h are defined by:

$$g * h \equiv \int_{-\infty}^{\infty} d\tau \, g(t - \tau) h(\tau) \tag{238}$$

and

$$\mathrm{Corr}(g, h) \equiv \int_{-\infty}^{\infty} d\tau \, g(t + \tau) h(\tau) \, . \tag{239}$$

A special case of the latter is the correlation of a function with itself, the *covariance*:

$$B_g \equiv \mathrm{Corr}(g, g) \, . \tag{240}$$

For complex functions, the *coherence function* is defined by:

$$B_g \equiv \mathrm{Corr}(g, g^*)\,. \tag{241}$$

The customary use of the same symbol B for covariance and coherence function is somewhat unfortunate, but should not be too confusing. The power spectral density $\Phi(f)$ is defined as

$$\Phi(f) \equiv \big|G(f)\big|^2\,. \tag{242}$$

The famous *Convolution Theorem* and *Correlation Theorem* are:

$$g * h \Longleftrightarrow G(f)H(f) \tag{243}$$

and

$$\mathrm{Corr}(g, h) \Longleftrightarrow G(f)H^*(f)\,. \tag{244}$$

The special case of the Correlation Theorem for the covariance is the *Wiener-Khinchin Theorem*:

$$B_g = \mathrm{Corr}(g, g) \Longleftrightarrow \big|G(f)\big|^2 = \Phi(f)\,. \tag{245}$$

The *structure function* D_g of a function g is defined by:

$$D_g(t_1, t_2) \equiv \Big\langle \big|g(t_1) - g(t_2)\big|^2 \Big\rangle\,. \tag{246}$$

If g describes a homogeneous and isotropic random process, D_g depends only on $t = |t_1 - t_2|$. By expanding the square in (246), we see that in this case

$$D_g(t) = 2\big(B_g(0) - B_g(t)\big)\,. \tag{247}$$

Finally, *Parseval's Theorem* states that the total power in a time series is the same as the total power in the corresponding spectrum:

$$\mathrm{TotalPower} \equiv \int_{-\infty}^{\infty} dt\,\big|g(t)\big|^2 = \int_{-\infty}^{\infty} df\,\big|G(f)\big|^2\,. \tag{248}$$

Acknowledgment

I thank Laurance Doyle and Oswald Wallner for careful reading of parts of the manuscript.

References

Aikawa Y, Herbst E. 1999. Molecular evolution in protoplanetary disks. Two-dimensional distributions and column densities of gaseous molecules. *ApJ* 351:233–46

Akeson RL, Swain MR. 1999. Differential phase mode with the Keck Interferometer. In *Working on the fringe,* ed. S Unwin, R Stachnik, pp. 89–94. ASP Conf. Ser. Vol. 194. San Francisco, CA

Alard C, Lupton R. 1998. A method for optimal image subtraction. *ApJ* 503:325–31

Albrow MD, An J, Beaulieu JP, Caldwell JAR, DePoy DL, et al. 2001. Limits on the abundance of Galactic planets from 5 years of PLANET observations. *ApJ* 556:L113–16

Albrow MD, An J, Beaulieu JP, Caldwell JAR, DePoy DL, et al. 2002. A short, non-planetary, microlensing anomaly: observations and lightcurve analysis of MACHO 99-BLG-47. *ApJ* 572:1031–1040

Albrow MD, Beaulieu JP, Birch P, Caldwell JAR, Kane S, et al. 1998. The 1995 pilot campaign of PLANET: searching for microlensing anomalies through precise, rapid, round-the-clock monitoring. *ApJ* 509:687–702

Albrow MD, Beaulieu JP, Caldwell JAR, DePoy DL, Dominik M, et al. 2000a. Limits on stellar and planetary companions in microlensing event OGLE-1998-BUL-14. *ApJ* 535:176–89

Albrow MD, Beaulieu JP, Caldwell JAR, Dominik M, Gaudi BS, et al. 2000b. Detection of rotation in a binary microlens: PLANET photometry of MACHO 97-BLG-41. *ApJ* 534:894–906

Alcock C. 2000. The dark halo of the Milky Way. *Science* 287:74–79

Alcock C, Akerlof CW, Allsman RA, Axelrod TS, Bennett DP, et al. 1993. Possible gravitational microlensing of a star in the Large Magellanic Cloud. *Nature* 365:621–23

Alcock C, Allen WH, Allsman RA, Alves D, Axelrod TS, et al. 1997. MACHO alert 95–30: first real-time observation of extended source effects in graviational microlensing. *ApJ* 491:436–50

Alcock C, Allsman RA, Alves D, Axelrod TS, Baines D, et al. 2000a. Binary microlensing events from the MACHO project. *ApJ* 541:270–97

Alcock C, Allsman RA, Alves DR, Axelrod TS, Becker AC, et al. 2000b. The MACHO Project: microlensing optical depth toward the Galactic bulge from difference image analysis. *ApJ* 541:734–66

Allard F, Hauschildt PH, Alexander DR, Starrfield S. 1997. Model atmospheres of very low mass stars and brown dwarfs. *ARAA* 35:137–77

André P. 1994. Observations of protostars and protostellar stages. In *The cold Universe,* ed. T Montmerle, CJ Lada, IF Mirabel, J Trân Thanh Vân, pp. 179–92. Editions Frontieres, Gif-sur-Yvette, France

Angel JRP. 1990. Use of a 16 m telescope to detect Earthlike planets. In *The Next Generation Space Telescope,* ed. PY Bely, CJ Burrows, GD Illingworth, pp. 81–94. Baltimore, MD

Angel JRP, Woolf NJ. 1997. An imaging nulling interferometer to study extrasolar planets. *ApJ* 475:373–79

Armitage PJ. 2000. Suppression of giant planet formation in stellar clusters. *A&A* 362:968–72

Armitage PJ. 2003. A reduced efficiency of terrestrial planet formation following giant planet migration. *ApJ* 582:L47–50

Armstrong JT, Mozurkewich D, Rickard LJ, Hutter DJ, Benson JA, et al. 1998. The Navy Prototype Optical Interferometer. *ApJ* 496:550–71

Artymowicz P. 1997. Beta Pictoris: an early Solar System? *Ann Rev Earth Planet Sci* 25:175–219

Arzoumanian Z, Joshi K, Rasio FA, Thorsett SE. 1996. Orbital parameters of the PSR 1620−26 triple system. In *Pulsars: problems and progress,* ed. S Johnston, MA Walker, M Bailes, pp. 525–30. IAU Coll. 160, ASP Conf. Ser. Vol. 105. San Francisco, CA

Ashton CE, Lewis GF. 2001. Gravitational microlensing of planets: the influence of planetary phase and caustic orientation. *MNRAS* 325: 305–11

Aubourg E, Bareyre P, Br'ehin S, Gros M, Lachieze-Rey M, et al. 1993. Evidence for gravitational microlensing by dark objects in the Galactic halo. *Nature* 365:623–25

Aumann HH, Gillett FC, Beichman CA, de John T, Houck JR, et al. 1984. Discovery of a shell around Alpha Lyrae. *ApJ* 278:L23–27

Baba N, Murakami N, Ishigaki T. 2001. Nulling interferometry by use of geometric phase. *Opt Lett* 26:1167–69

Backman DE, Paresce F. 1993. Main-sequence stars with circumstellar solid material: the Vega phenomenon. In *Protostars and planets III,* ed. EH Levy, JI Lunine, pp. 1253–1304. University of Arizona Press.

Baglin A, Auvergne M, Barge P, Buey JT, Catala C, et al. 2002. COROT: asteroseismology and planet finding. In *Proceedings of the first Eddington workshop on stellar structure and habitable planet finding,* ed. F Favata, IW Roxburgh, D Galadi, pp. 17–24. ESA SP-485, Noordwijk: ESA Publications Division

Bailes M, Lyne AG, Shemar SL. 1991. A planet orbiting the neutron star PSR1829−10. *Nature* 352:311–13

Bally J, O'Dell CR, McCaughrean MJ. 2000. Disks, microjets, windblown bubbles, and outflows in the Orion nebula. *AJ* 119:2919–59

Baraffe I, Chabrier G, Barman TS, Allard F, Hauschildt PH. 2003. Evolutionary models for cool brown dwarfs and extrasolar giant planets. The case of HD 209458. *A&A* 402:701–712

Baranne A. 1999. Spectrographs for the measurement of radial velocities. In *Precise stellar radial velocities,* ed. JB Hearnshaw, CD Scarfe, pp. 1–12. IAU Coll. 170, ASP Conf. Ser. Vol. 185. San Francisco, CA

Baranne A, Queloz D, Mayor M, Adrianzyk G, Knispel G, et al. 1996. ELODIE: A spectrograph for accurate radial velocity measurements. *A&AS* 119:373–90

Barbieri M, Gratton RG. 2002. Galactic orbits of stars with planets. *A&A* 384:879–83

Barbieri M, Marzari F, Scholl H. 2002. Formation of terrestrial planets in close binary systems: the case of α Centauri A. *A&A* 396:219–24

Barman TS, Hauschildt PH, Allard F. 2001. Irradiated planets. *ApJ* 556:885–95

Barman TS, Hauschildt PH, Schweitzer A, Stanch, PC, Baron E, Allard F. 2002. Non-LTE effects of Na I in the atmosphere of HD 209458 b. *ApJ* 569:L51–54

Barnes JW, O'Brien DP. 2002. Stability of satellites around close-in extrasolar giant planets. *ApJ* 575:1087–93

Barnes SA. 2001. An assessment of the rotation rates of the host stars of extrasolar planets. *ApJ* 561:1095–1106

Bastian TS, Dulk GA, Leblanc Y. 2000. A search for radio emission from extrasolar planets. *ApJ* 545:1058–63

Batten AH. 1973. *Binary and multiple systems of stars.* Oxford: Pergamon Press, 278 pp.

Beckwith SVW, Sargent AI. 1993. The occurence and properties of disks around young stars. In *Protostars and planets III,* ed. EH Levy, JI Lunine, pp. 521–41. University of Arizona Press.

Béjar VJS, Martín EL, Zapatero Osorio MR, Rebolo R, Barrado y Navascués D, et al. 2001. The substellar mass function in σ Orionis. *ApJ* 556:830–36

Benedict GF, McArthur BE, Forveille T, Delfosse X, Nelan E, et al. 2002. A mass for the extrasolar planet Gl 876 b determined from Hubble Space Telescope Fine Guidance Sensor 3 astrometry and high-precision radial velocities. *ApJ* 581:L115–18

Bennett DP, Alcock C, Allsman RA, Axelrod TS, Cook KH, et al. 1995. Recent developments in gravitational microlensing and the latest MACHO results: microlensing towards the Galactic bulge. In *Dark matter,* ed. S Holt, DP Bennett, pp. 77–90. AIP Conf. Proc. Vol. 336, New York

Bennett DP, Rhie SH. 1996. Detecting Earth-mass planets with gravitational microlensing. *ApJ* 472:660–64

Bennett DP, Rhie SH. 2000. The Galactic Exoplanet Survey Telescope: a proposed space-based microlensing survey for terrestrial extrasolar planets. In *Disks, planetesimals, and planets,* ed. F Garzón, C Eiroa, D de Winter, TJ Mahoney, pp. 542–49. ASP Conf. Ser. Vol. 219, San Francisco, CA

Bennett DP, Rhie SH, Becker AC, Butler N, Dann J, et al. 1999. Discovery of a planet orbiting a binary star system from gravitational microlensing. *Nature* 402:57–59

Benson JA, Mozurkewich D, Jefferies SM. 1998. Active optical fringe tracking at the NPOI. In *Astronomical interferometry,* ed. RD Reasenberg, pp. 493–96. SPIE Vol. 3350. Bellingham, WA

Berry MV. 1987. The adiabatic phase and Pancharatnam's phase for polarized light. *J Mod Opt* 34:1401–07

Bertout C. 1989. T Tauri stars: wild as dust. *ARAA* 27:351–95

Binnendijk L. 1960. *Properties of double stars.* Philadelphia: University of Pennsylvania Press.

Boden AF, Shao M, Van Buren D. 1998. Astrometric observation of MACHO gravitational microlensing. *ApJ* 502:538–49

Bodenheimer P, Hubickyj O, Lissauer JJ. 2000. Models of the in situ formation of detected extrasolar giant planets. *Icarus* 143:2–14

Bodenheimer P, Lin DNC, Mardling RA. 2001. On the tidal inflation of short-period extrasolar planets. *ApJ* 548:466–72

Bodenheimer P, Pollack JB. 1986. Calculations of the accretion and evolution of giant planets: the effects of solid cores. *Icarus* 67:391–408

Bokhove H, Kappelhof P, Vink R. 2002. Achromatic phase shifting: the dispersive approach. In *Proceedings of GENIE-DARWIN Workshop - Hunting for Planets*, ESA SP-522, p13

Bolatto AD, Falco EE. 1994. The detectability of planetary companions of compact Galactic objects from their effects on microlensed light curves of distant stars. *ApJ* 436:112–16

Bond IA, Abe F, Dodd RJ, Hearnshaw JB, Honda M, et al. 2001. Real-time difference imaging analysis of MOA Galactic bulge observations during 2000. *MNRAS* 327:868–80

Bond IA, Abe F, Dodd RJ, Hearnshaw JB, Kilmartin PM, et al. 2002a. Improving the prospects for detecting extrasolar planets in gravitational microlensing events in 2002. *MNRAS* 331:L19–23

Bond IA, Rattenbury NJ, Skuljan J, Abe F, Dodd RJ, et al. 2002b. Study by MOA of extra-solar planets in gravitational microlensing events of high magnification. *MNRAS* 333:71–83

Bonnell IA, Smith KW, Davies MB, Horne K. 2001. Planetary dynamics in stellar clusters. *MNRAS* 322:859–65

Born M, Wolf E. 1997. *Principles of optics, 6th edition.* Cambridge: Cambridge University Press, 808 pp.

Borucki WJ, Caldwell D, Koch DG, Webster LD, Jenkins JM, et al. 2001. The Vulcan photometer: a dedicated photometer for extrasolar planet searches. *PASP* 113:439–51

Borucki WJ, Koch DG, Dunham EW, Jenkins JM. 1997. The Kepler mission: a mission to determine the frequency of inner planets near the habitable zone of a wide range of stars. In *Planets beyond the Solar System and the next generation of space missions,* ed. D Soderblom, pp. 153–73. ASP Conf. Ser. Vol. 119. San Francisco, CA

Borucki WJ, Summers AL. 1984. The photometric method of detecting other planetary systems. *Icarus* 58:121–34

Boss A. 1998a. *Looking for Earths.* New York: Wiley, 240 pp.

Boss AP. 1998b. Astrometric signatures of giant-planet formation. *Nature* 393:141–43

Boss A. 2002. Stellar metallicity and the formation of extrasolar gas giant planets. *ApJ* 567:L149–53

Boss AP, Butler RP, Hubbart WB, Ianna PA, Kürster M, et al. 2003. Working group on extrasolar planets. *Rep Astron* XXVA:144–46

Bouchy F, Pepe F, Queloz D. 2001. Fundamental photon noise limit to radial velocity measurements. *A&A* 374:733–39

Bracewell R. 1965. *The Fourier transform and its applications.* New York: McGraw-Hill, 381 pp.

Bracewell RN. 1978. Detecting nonsolar planets by spinning infrared interferometer. *Nature* 274:780–81

Brown TM. 2001. Transmission spectra as diagnostics of extrasolar giant planet atmospheres. *ApJ* 553:1006–26

Brown TM, Charbonneau D. 2000. The STARE project: a transit search for hot Jupiters. In *Disks, planetesimals and planets,* ed. F Garzón, C Eiroa, D de Winter, TJ Mahoney, pp. 584–89. ASP Conf. Ser. Vol. 219. San Francisco, CA

Brown TM, Charbonneau D, Gilliland RL, Noyes RW, Burrows A. 2001. Hubble Space Telescope time-series photometry of the transiting planet of HD 209458. *ApJ* 552:699–709

Brown TM, Libbrecht KG, Charbonneau D. 2002. A search for CO absorption in the transmission spectrum of HD 209458 b. *PASP* 114:826–32

Bruijn MP, Tiest WB, Hoevers HF, van der Kuur J, Mels WA, de Korte PA. 2000. Toward a cryogenic imaging array of transition edge X-ray microcalorimeters. In *X-Ray optics, instruments, and missions III,* ed. JE Trümper, B Aschenbach, pp. 145–53. SPIE Vol. 4012. Bellingham, WA

Bryden G. 2001. Ejected planets near young stars. In *Young stars near Earth: progress and prospects,* ed. R Jayawardhana, TP Greene, pp. 328–33. ASP Conf. Ser. Vol. 244. San Francisco, CA

Bundy KA, Marcy GW. 2000. A search for transit effects in spectra of 51 Pegasi and HD 209458. *PASP* 112:1421–25

Burrows C, Stapelfeldt K, Watson A, et al. 1996. Hubble space telescope observation of the disk and jet of HH 30. *ApJ* 473:437

Burrows A, Guillot T, Hubbard WB, Marley MS, Saumon D. 2000. On the radii of close-in giant planets. *ApJ* 534:L97–100

Burrows A, Marley M, Hubbard WB, Lunine JI, Guillot T, et al. 1997a. A nongray theory of extrasolar giant planets and brown dwarfs. *ApJ* 491:856–75

Burrows A, Sudarsky D, Lunine JI. 2003. Beyond the T dwarfs: Theoretical spectra, colors, and detectability of the coolest brown dwarfs. *ApJ* 596:587–596

Burrows C, Stapelfeldt K, Watson A, et al. 2006, APJ 473,437. "Hubble Space Telegre objevation of the Diskance jet of HH30"

Buscher DF, Armstrong JT, Hummel CA, Quirrenbach A, Mozurkewich D, et al. 1995. Interferometric seeing measurements on Mt. Wilson: power spectra and outer scales. *Appl Opt* 34:1081–96

Butler RP, Marcy GW. 1996. A planet orbiting 47 Ursae Majoris. *ApJ* 464:L153–56

Butler RP, Marcy GW, Fischer DA, Brown TM, Contos AR, et al. 1999. Evidence for multiple companions to υ Andromedae. *ApJ* 526:916–27

Butler RP, Marcy GW, Vogt SS, Fischer DA, Henry GW, et al. 2003. Seven new Keck planets orbiting G and K dwarfs. *ApJ* 582:455–66

Butler RP, Marcy GW, Vogt SS, Tinney CG, Jones HRA, et al. 2002. On the double planet system around HD 83443. *ApJ* 579:565–72

Butler RP, Marcy GW, Williams E, Hauser H, Shirts, P. 1997. Three new "51 Pegasi-type" planets. *ApJ* 474:L115–18

Butler RP, Marcy GW, Williams E, McCarthy C, Dosanjh P, Vogt SS. 1996. Attaining Doppler precision of $3\,\mathrm{m\,s^{-1}}$. *PASP* 108:500–09

Butler RP, Tinney CG, Marcy GW, Jones HRA, Penny AJ, Apps K. 2001. Two new planets from the Anglo-Australian planet search. *ApJ* 555:410–17

Butler RP, Vogt SS, Marcy GW, Fischer DA, Henry GW, Apps K. 2000. Planetary companions to the metal-rich stars BD $-10°3166$ and HD 52265. *ApJ* 545:504–11

Campbell B, Walker GAH. 1979. Precision radial velocities with an absorption cell. *PASP* 91:540–45

Campbell B, Walker GAH, Yang S. 1988. A search for substellar companions to Solar-type stars. *ApJ* 331:902–21

Charbonneau D, Brown TM, Latham DW, Mayor M. 2000. Detection of planetary transits across a Sun-like star. *ApJ* 529:L45–48

Charbonneau D, Brown TM, Noyes RW, Gilliland RL. 2002. Detection of an extrasolar planet atmosphere. *ApJ* 568:377–384

Charbonneau D, Jha S, Noyes RW. 1998. Spectral line distortions in the presence of a close-in planet. *ApJ* 507:L153–56

Charbonneau D, Noyes RW, Korzennik SG, Nisenson P, Jha S., et al. 1999. An upper limit on the reflected light from the planet orbiting the star τ Bootis. *ApJ* 522:L145–48

Chiang EI, Murray N. 2002. Eccentricity excitation and apsidal resonance capture in the planetary system υ Andromedae. *ApJ* 576:473–77

Chiang EI, Tabachnik S, Tremaine S. 2001. Apsidal alignment in υ Andromedae. *AJ* 122:1607–15

Christensen-Dalsgaard J. 2002. MONS on the Danish Rømer satellite: measuring oscillations in nearby stars. In *Proceedings of the first Eddington workshop on stellar structure and habitable planet finding,* ed. F Favata, IW Roxburgh, D Galadi, pp. 25–34. ESA SP-485, Noordwijk: ESA Publications Division

Chwolson, O. 1924. Uber eine mögliche Form fiktiver Doppelsterne. *Astron Nachrichten* 221:329

Clampin M, Krist JE, Ardila DR, Golimowski DA, Hartig GF, et al. 2003. HST/ACS coronagraphic imaging of the circumstellar disk around HD 141569 A. *AJ* 126:385–392

Claret A, Díaz-Cordovés J, Giménez A. 1995. Linear and non-linear limb-darkening coefficients for the photometric bands *R I J H K*. *A&AS* 114:247–52

Clark GL, Roychoudhuri C. 1979. Interferometry through single-mode optical fibers. In *Interferometry*, ed. GW Hopkins, pp. 196–203. SPIE Vol. 192. Bellingham, WA

Cocconi G, Morrison P. 1959. Searching for interstellar communications. *Nature* 184:844–46

Cochran WD, Hatzes AP, Butler RP, Marcy GW. 1997. The discovery of a planetary companion to 16 Cygni B. *ApJ* 483:457–63

Cochran WD, Hatzes AP, Paulson DB. 2000. The McDonald Observatory planetary search program: past, present, and future. In *Planetary Systems in the Universe,* ed. A Penny, P Artymowicz, AM Lagrange, S Russel. IAU Symp. 202, ASP Conf. Ser. San Francisco, CA, in press

Cohen CJ, Hubbard EC. 1965. Libration of the close approaches of Pluto to Neptune. *AJ* 70:10–13

Colavita, MM. 1985. Atmospheric limitations of a two-color astrometric interferometer. Ph.D. thesis, M.I.T., Cambridge, MA

Colavita MM. 1994. Measurement of the atmospheric limit to narrow-angle interferometric astrometry using the Mark III stellar interferometer. *A&A* 283:1027–36

Collier Cameron A, Horne K, Penny A, James D. 1999. Probable detection of starlight reflected from the giant planet orbiting τ Bootis. *Nature* 402:751–55

Collier Cameron A, Horne K, Penny A, Leigh C. 2002. A search for starlight reflected from υ And's innermost planet. *MNRAS* 330:187–204.

Conan R, Ziad A, Borgnino J, Martin F, Tokovinin A. 2000. Measurements of the wave-front outer scale at Paranal: influence of this parameter in interferometry. In *Interferometry in optical astronomy,* ed. PJ Léna, A Quirrenbach, pp. 963–73. SPIE Vol. 4006. Bellingham, WA

Connes P. 1985. Absolute astronomical accelerometry. *Ap&SS* 110:211–55

Conway Morris S. 1998. *The crucible of creation.* New York: Oxford University Press. 272 pp.

Copi CJ, Starkman GD. 2000. The Big Occulting Steerable Satellite (BOSS). *ApJ* 532:581–92

Coudé du Foresto V, Ridgway S, Mariotti JM. 1997. Deriving object visibilities from interferograms obtained with a fiber stellar interferometer. *A&AS* 121:379–92

Crawford I. 1990. Intestellar travel: a review for astronomers. *QJRAS* 31:377–400

Crawford I. 2000. Where are they? *Sci Am* 283:28–33

Creech-Eakman MJ, Kulkarni SR, Pan XP, Shaklan SB. 1999. Photometric measurements of the fields of more than 700 nearby stars. *ApJ* 118:2483–87

Cumming A, Marcy GW, Butler RP. 1999. The Lick planet search: detectability and mass thresholds. *ApJ* 526:890–915

Cumming A, Marcy GW, Butler RP, Vogt SS. 2002. The statistics of extrasolar planets: results from the Keck survey. In *Scientific Frontiers iin Research on Extrasolar Planets* ed. D Deming and S. Seager *ASP Conf Ser 294*: 27–30

Cuntz M, Shkolnik E. 2002. Chromospheres, flares and exoplanets. *Astron Nachr* 323:387–91

Daigne G, Lestrade JF. 1999. Astrometric optical interferometry with non-evacuated delay lines. *A&AS* 138:355–63

Davies MB, Sigurdsson S. 2001. Planets in 47 Tuc. *MNRAS* 324:612–16

Davis J, Lawson PR, Booth AJ, Tango WJ, Thorvaldson ED. 1995. Atmospheric path variations for baselines up to 80 m measured with the Sydney University Stellar Interferometer. *MNRAS* 273:L53–58

Deeg HJ, Doyle LR, Kozhevnikov VP, Martín EL, Oetiker B, et al. 1998. Near-term detectability of terrestrial extrasolar planets: TEP network observations of CM Draconis. *A&A* 338:479–90

Deeg HJ, Doyle LR, Kozhevnikov VP, et al. 2000. A search for Jovian-Mass planets around CM Draconis using eclipse minima timing. *A & A* 358:L5.

Delfosse X, Forveille T, Mayor M, Perrier C, Naef D, Queloz D. 1998. The closest extrasolar planet. A giant planet around the M4 dwarf GL 876. *A&A* 338:L67–70

Delplancke F, Górski KM, Richichi A. 2001. Resolving gravitational microlensing events with long-baseline optical interferometry. Prospects for the ESO Very Large Telescope Interferometer. *A&A* 375:701–10

De Pater I, Lissauer JJ. 2001. *Planetary sciences.* Cambridge: Cambridge University Press. 528 pp.

Dermott SF, Grogan K, Durda DD, Jayaraman S, Kehoe TJJ, et al. 2001. Orbital evolution of interplanetary dust. In *Interplanetary dust,* ed. E Grün, BÅS Gustafson, SF Dermott, H Fechtig, pp. 569–639. Springer-Verlag

Derue F, Afonso C, Alard C, Albert JN, Andersen J, et al. 2001. Observation of microlensing toward the Galactic spiral arms. EROS II 3 year survey. *A&A* 373:126–38

Díaz-Cordovés J, Claret A, Giménez A. 1995. Linear and non-linear limb-darkening coefficients for LTE model atmospheres. *A&AS* 110:329–50

Dick SJ. 1982. *Plurality of worlds: the extraterrestrial life debate from Democritus to Kant.* Cambridge: Cambridge University Press. 246 pp.

Dick SJ. 1998. *Life on other worlds: the twentieth century extraterrestrial life debate.* Cambridge: Cambridge University Press. 304 pp.

Dominik C, Laureijs RJ, Jourdain de Muizon M, Habing HJ. 1998. A Vega-like disk associated with the planetary system of ρ^1 Cnc. *A&A* 329:L53–56

Doyle LR, Deeg HJ, Kozhevnikov VP, Oetiker B, Martín EL, et al. 2000. Observational limits on terrestrial-sized inner planets around the CM Draconis system using the photometric transit method with a matched-filter algorithm. *ApJ* 535:338–49

Drake FD. 1962. *Intelligent life in space.* New York: Macmillan. 128 pp.

Dravins D, Lindegren L, Mezey E, Young AT. 1998. Atmospheric intensity scintillation of stars. III. Effects for Different Telescope Apertures. *PASP* 110:610–33

Dreher JW, Cullers DK. 1997. SETI figure of merit. In *Astronomical and biochemical origins and the search for life in the Universe,* ed. CB Cosmovici, S Bowyer, D Werthimer, pp. 711–17. Bologna: Editrice Compositori

Dreizler S, Hauschildt PH, Kley W, Rauch T, Schuh SL, et al. 2003. OGLE-TR-3: a possible new transiting planet. *A&A* 402:791–99

Dreizler S, Rauch T, Hauschildt P, Schuh S, Kley W, Werner K. 2002. Spectral types of planetary host star candidates: two new transiting planets? *Astron Nachrichten Suppl* 324:23–28

Dutrey A. 1999. Latest stages of star formation and circumstellar environment of young stellar objects. In *Planets outside the Solar System: theory and observations,* ed. JM Mariotti, D Alloin, pp. 13–49. NATO ASI Vol. 532, Dordrecht: Kluwer

Dutrey A, Guilloteau S, Guelin M. 1997. Chemistry of protosolar-like nebulae: the molecular content of the DM Tau and GG Tau disks. *A&A* 317:L55–58

Dutrey A, Guilloteau S, Prato L, Simon M, Duvert G. 1998. CO study of the GM Aurigae Keplerian disk. *A&A* 338:L63–66

Dyson FW, Eddington AS, Davidson C. 1920. A Determination of the deflection of light by the Sun's gravitational field, from observations made at the total eclipse of May 29, 1919. *Phil Trans Roy Soc Series A* 220:291–333

Eddington AS. 1920. *Space, time and gravitation.* Cambridge: Cambridge University Press.

Einstein A. 1936. Lens-like action of a star by the deviation of light in the gravitational field. *Science* 84:506–07

Eisenhauer F, Quirrenbach A, Zinnecker H, Genzel R. 1998. Stellar content of the galactic starburst template NGC 3603 from adaptive optics observations. *ApJ* 498:278–92

Eisner JA, Kulkarni SR. 2001a. Sensitivity of the radial-velocity technique in detecting outer planets. *ApJ* 550:871–83

Elliot JL. 1978. Direct imaging of extrasolar planets with stationary occultations viewed by a space telescope. *Icarus* 35:156–64

Encrenaz T. 2001. The formation of planets. In *Solar and extra-solar planetary systems, ed. IP Williams, N Thomas,* pp. 76–90. Lecture Notes in Physics, Springer, Berlin

Endl M, Kürster M, Els S, Hatzes AP, Cochran WD, et al. 2002. The planet search program at the ESO Coudé Echelle Spectrometer. III. The complete long camera survey results. *A&A* 392:671–90

Erdl H, Schneider P. 1993. Classification of the multiple deflection two point-mass gravitational lens models and application of catastrophe theory in lensing. 268:453–71

European Space Agency. 1997. *The Hipparcos and Tycho catalogues.* ESA SP-1200

Evans NJ. 1999. Physical conditions in regions of star formation. *ARAA* 37:311–62

Everett ME, Howell SB. 2001. A technique for ultrahigh-precision CCD photometry. *PASP* 113:1428–35

Farrell WM, Desch MD, Zarka P. 1999. On the possibility of coherent cyclotron emission from extrasolar planets. *JGR* 104:14025–32

Favata F. 2002. The Eddington baseline mission. In *Proceedings of the first Eddington workshop on stellar structure and habitable planet finding,* ed. F Favata, IW Roxburgh, D Galadi, pp. 3–10. ESA SP-485, Noordwijk: ESA Publications Division

Ferlet R, Vidal-Madjar A, Hobbs LM. 1987. The Beta Pictoris circumstellar disk. V. Time variations of the Ca II-K line. *A&A* 185:267–70

Fischer DA, Butler RP, Marcy GW, Vogt SS, Henry GW. 2003a. A sub-Saturn mass companion to HD 3651. *ApJ*, submitted

Fischer DA, Marcy GW, Butler RP, Laughlin G, Vogt SS. 2002a. A Second Planet Orbiting 47 Ursae Majoris. *ApJ* 564:1028–34

Fischer DA, Marcy GW, Butler RP, Vogt SS, Frink S, Apps K. 2001. Planetary companions to HD 12661, HD 92788, and HD 38529 and variations in Keplerian residuals of extrasolar planets. *ApJ* 551:1107–18

Fischer DA, Marcy GW, Butler RP, Vogt SS, Henry GW, et al. 2003b. A planetary companion to HD 40979 and additional planets orbiting HD 12661 and HD 38529. *ApJ* 586:1394–408

Fischer DA, Marcy GW, Butler RP, Vogt SS, Walp B, Apps K. 2002b. Planetary companions to HD 136118, HD 50554, and HD 106252. *PASP* 114:529–35

Ford EB, Joshi KJ, Rasio FA, Zbarsky B. 2000. Theoretical implications of the PSR B1620−26 triple system and its planet. *ApJ* 528:336–50

Freundling W, Lagrange AM, Vidal-Madjar A, Ferlet R, Forveille T. 1995. Gas around β Pictoris: an upper limit on the H I content. *A&A* 301:231–35

Fried DL. 1994. Atmospheric turbulence optical effects: understanding the adaptive-optics implications. In *Adaptive optics for astronomy,* ed. DM Alloin, JM Mariotti, pp. 25–57. NATO ASI Vol. 423, Dordrecht: Kluwer

Frink S, Mitchell DS, Quirrenbach A, Fischer DA, Marcy GW, Butler RP. 2002. Discovery of a substellar companion to the K2 III giant ι Draconis. *ApJ* 576:478–84

Frink S, Quirrenbach A, Fischer D, Röser S, Schilbach E. 2000a. K giants as astrometric reference stars for the Space Interferometry Mission. In *Interferometry in optical astronomy,* ed. PJ Léna, A Quirrenbach, pp. 806–14. SPIE Vol. 4006. Bellingham, WA

Frink S, Quirrenbach A, Fischer D, Röser S, Schilbach E. 2001. A strategy for identifying the grid stars for the Space Interferometry Mission (SIM). *PASP* 113:173–87

Frink S, Quirrenbach A, Röser S, Schilbach E. 2000b. Testing Hipparcos K giants as grid stars for SIM. In *Working on the fringe,* ed. S Unwin, R Stachnik, pp. 128–33. ASP Conf. Ser. Vol. 194, San Francisco, CA

Gatewood G, Eichhorn H. 1973. An unsuccessful search for a planetary companion of Barnard's Star (BD +4°3561). *AJ* 78:769–76

Gaudi BS. 1998. Distinguishing between binary-source and planetary microlensing perturbations. *ApJ* 506:533–39

Gaudi BS. 2000. Planetary transits toward the Galactic bulge. *ApJ* 539:L59–62

Gaudi BS, Albrow MD, An J, Beaulieu JP, Caldwell JAR, et al. 2002. Microlensing constraints on the frequency of Jupiter-mass companions: analysis of 5 years of PLANET photometry. *ApJ* 566:463–99

Gaudi BS, Gould, A. 1997. Planet parameters in microlensing events. *ApJ* 486:85–99

Gaudi BS, Naber RM, Sackett PD. 1998. Microlensing by multiple planets in high-magnification events. *ApJ* 502:L33–37

Gaudi BS, Sackett PD. 2000. Detection efficiencies of microlensing data sets to stellar and planetary companions. *ApJ* 528:56–73

Gay J, Rabbia Y. 1996. Principe du coronographe interférentiel achromatique. *C R Acad Sci Paris* 322:265–71

Ge J, Erskine D, Rushford M. 2002. An externally dispersed interferometer for sensitive Doppler extrasolar planet searches. *PASP* 114:1016–28

Genzel R. 1992. In *The galactic interstellar medium.* Springer-Verlag

Gilliland RL, Brown TM, Guhathakurta P, Sarajedini A, Milone EF, et al. 2000. A Lack of planets in 47 Tucanae from a Hubble Space Telescope search. *ApJ* 545:L47–51

Gilmore G, de Boer K, Favata F, Høg E, Lattanzi M, et al. 2000. GAIA: origin and evolution of the Milky Way. In *UV, optical, and IR space telescopes and instruments,* ed. JB Breckinridge, P Jakobsen, pp. 453–72. SPIE Vol. 4013, Bellingham, WA

Gladman B. 1993. Dynamics of systems of two close planets. *Icarus* 106:247–63

Goldreich P, Soter S. 1966. Q in the Solar System. *Icarus* 5:375–89

Gonzalez G, Laws C, Tyagi S, Reddy BE. 2001b. Parent stars of extrasolar planets. VI. Abundance analyses of 20 new systems. *AJ* 121:432–52

Goodwin SP. 1997. Residual gas expulsion from young globular clusters. *MNRAS* 284:785–802

Gould A, Loeb A. 1992. Discovering planetary systems through gravitational microlensing. *ApJ* 396:104–14

Gould SJ. 1989. *Wonderful life: the Burgess Shale and the nature of history.* New York: W.W. Norton. 347 pp.

Goździewski K. 2002. Stability of the 47 UMa planetary system. *A&A* 393:997–1013

Goździewski K, Maciejewski, AJ. 2001. Dynamical analysis of the orbital parameters of the HD 82943 planetary system. *ApJ* 563:L81–85

Goździewski K, Maciejewski, AJ. 2003. The Janus head of the HD 12661 planetary system. *ApJ* 586:L153–56

Grady CA, Mora A, de Winter D. 2000a. Infall, accretion, and the spectroscopic evidence for planetesimals. In *Disks, planetesimals and planets,* ed. F Garzón, C Eiroa, D de Winter, TJ Mahoney, pp. 202–14. ASP Conf. Ser. Vol. 219. San Francisco, CA

Grady CA, Pérez MR, Talavera A, Bjorkman KS, de Winter D, et al. 1996. The β Pictoris phenomenon among Herbig Ae/Be stars. UV and optical high dispersion spectra. *A&AS* 120:157–77

Grady CA, Sitko ML, Bjorkman KS, Pérez MR, Lynch DK, et al. 1997. The star-grazing extrasolar comets in the HD 100546 system. *ApJ* 483:449–56

Grady CA, Sitko ML, Russell RW, Lynch DK, Hanner MS, et al. 2000b. Infalling planetesimals in pre-main stellar systems. In *Protostars and planets IV,* ed. V Mannings, AP Boss, SS Russell, pp. 613–38. University of Arizona Press

Graff DS, Gaudi BS. 2000. Direct detection of large close-in planets around the source stars of caustic-crossing microlensing events. *ApJ* 538:L133–36

Gratton RG, D'Antona F. 1989. HD 39853: a high velocity K5 III star with an exceptionally large Li content. *A&A* 215:66–78

Gray DF. 1998. A planetary companion for 51 Pegasi implied by absence of pulsations in the stellar spectra. *Nature* 391:153–54

Greaves JS, Holland WS, Moriarty-Schieven G, Jenness T, Dent WR, et al. 1998. A dust ring around ε Eridani: analog to the young Solar System. *ApJ* 506:L133–37

Greaves JS, Mannings V, Holland WS. 2000. The dust and gas content of a disk around the young star HR 4796 A. *Icarus* 143:155–58

Griest K, Safizadeh N. 1998. The use of high-magnification microlensing events in discovering extrasolar planets. *ApJ* 500:37–50

Griffin RF, Griffin RE. 1973. On the possibility of determining stellar radial velocities to 0.01 km/s. *MNRAS* 162:243–53

Grinin VP. 1999. Infalling material on young stars. In *Planets outside the Solar System: theory and observations,* ed. JM Mariotti, D Alloin, pp. 51–63. NATO ASI Vol. 532, Dordrecht: Kluwer

Gubler J, Tytler D. 1998. Differential atmospheric refraction and limitations on the relative astrometric accuracy of large telescopes. *PASP* 110:738–46

Guilloteau S, Dutrey A. 1998. Physical parameters of the Keplerian protoplanetary disk of DM Tauri. *A&A* 339:467–76

Guirado J, Ros E, Jones D, Alef W, Marcaide J, Preston R. 2002. Searching for low mass objects around nearby dMe radio stars. In *Proceedings of the 6th European VLBI Network Symposium,* ed E Ros, RW Porcas, AP Lobanov, JA Zensus, pp. 255–58. MPIfR, Bonn, Germany

Haisch KE, Lada EA, Lada CJ. 2001a. Circumstellar disks in the IC 348 cluster. *AJ* 121:2065–74

Haisch KE, Lada EA, Lada CJ. 2001b. Disk frequencies and lifetimes in young clusters. *ApJ* 553:L153–56

Halbwachs JL, Arenou F, Mayor M, Udry S, Queloz D. 2000. Exploring the brown dwarf desert with Hipparcos. *A&A* 355:581–94

Hamilton CM, Herbst W, Shih C, Ferro AJ. 2001. Eclipses by a circumstellar dust feature in the pre-main-sequence star KH 15D. *ApJ* 554:L201–04

Han C. 2002. Astrometric method to break the photometric degeneracy between binary-source and planetary microlensing perturbations. *ApJ* 564:1015–18

Hanner MS, Lynch DK, Russell RW. 1994. The 8–13 micron spectra of comets and the composition of silicate grains. *ApJ* 425:274–85

Hardy JW. 1998. *Adaptive optics for astronomical telescopes.* New York: Oxford University Press. 438 pp.

Harrington RS, Kallarakal VV, Dahn CC. 1983. Astrometry of the low-luminosity stars VB8 and VB10. *AJ* 88:1038–39

Hatzes AP, Cochran WD. 1994. Short-period radial velocity variations of α Bootis: Evidence for radial pulsations. *ApJ* 422:366–73

Hatzes AP, Cochran WD, McArthur B, Baliunas SL, Walker GAH, et al. 2000. Evidence for a long-period planet orbiting ε Eridani. *ApJ* 544:L145–48

Heap SR, Lindler DJ, Lanz TM, Cornett RH, Hubeny I, et al. 2000. Space Telescope Imaging Spectrograph Coronagraphic observations of β Pictoris. *ApJ* 539:435–44

Heintz WD. 1971. *Doppelsterne.* München: Wilhelm Goldmann Verlag. 186 pp.

Henry GW, Donahue RA, Baliunas SL. 2002. A false planet around HD 192263. *ApJ* 577:L111–14

Henry GW, Marcy GW, Butler RP, Vogt SS. 2000. A Transiting "51 Peg-like" Planet. *ApJ* 529:L41–44

Herbst W, Hamilton CM, Vrba FJ, Ibrahimov MA, Bailer-Jones CAL, et al. 2002. Fine structure in the circumstellar environment of a young, Solar-like star: the unique eclipses of KH 15D. *PASP* 114:1167–72

Hestroffer D. 1997. Centre to limb darkening of stars: new model and application to stellar interferometry. *A&A* 327:199–206

Hill HGM, Grady CA, Nuth JA, Hallenbeck SL, Sitko ML. 2001. Constraints on nebular dynamics and chemistry based on observations of annealed magnesium silicate grains in comets and in disks surrounding Herbig Ae/Be stars. *PNAS* 98:2182–87

Hinz PM, Angel R, Hoffman W, McCarthy DW, McGuire PC, et al. 1998. Imaging circumstellar environments with a nulling interferometer. *Nature* 395:251–53

Holland WS, Greaves JS, Zuckerman B, Webb RA, McCarthy C, et al. 1998. Submillimetre images of dusty debris around nearby stars. *Nature* 392:788–91

Holland WS, Greaves JS, Dent WRF, Wyatt MC, Zuckerman B, et al. 2003. Submillimeter observations of an asymmetric dust disk around Fomalhaut. *ApJ* 582:1141–46

Hubbard WB, Fortney JJ, Lunine JI, Burrows A, Sudarsky D, Pinto P. 2001. Theory of extrasolar giant planet transits. *ApJ* 560:413–19

Hufnagel, RE. 1974. Variations of atmospheric turbulence. In *Digest of technical papers presented at the topical meeting on optical propagation through turbulence*, Optical Society of America, p. Wa1:1–4

Hui L, Seager S. 2002. Atmospheric lensing and oblateness effects during an extrasolar planetary transit. *ApJ* 572:540–55

Ida S, Makino J. 1993. Scattering of planetesimals by a protoplanet: slowing down of runaway growth. *Icarus* 106:210–27

Ignace R. 2001. Spectral energy distribution signatures of Jovian planets around white dwarf stars. *PASP* 113:1227–31

Israelian G, Santos NC, Mayor M, Rebolo R. 2001. Evidence for planet engulfment by the star HD 82943. *Nature* 411:163–66

Janes K. 1996. Star clusters: Optimal targets for a photometric planetary search program. *JGR* 101:14853–60

Jayawardhana R, Holland WS, Kalas P, Greaves JS, Dent WRF, et al. 2002. New submillimeter limits on dust in the 55 Cancri planetary system. *ApJ* 570:L93–96

Jenkins JM, Caldwell DA, Borucki WJ. 2002. Some tests to establish confidence in planets discovered by transit photometry. *ApJ* 564:495–507

Jewitt D, Luu J. 1993. Discovery of the candidate Kuiper belt object 1993 QB1. *Nature* 362:730–32

Ji J, Li G, Liu L. 2002. The dynamical simulations of the planets orbiting GJ 876

Jiang IG, Ip WH, Yeh LC. 2003. On the fate of close-in extrasolar planets. *ApJ* 582:449–54

Johnstone D, Hollenbach D, Bally J. 1998. Photoevaporation of disks and clumps by nearby massive stars: application to disk destruction in the Orion Nebula. *ApJ* 499:758–76

Jorissen A, Mayor M, Udry S. 2001. The distribution of exoplanet masses. *A&A* 379:992–98

Joshi KJ, Rasio FA. 1997. Distant companions and planets around milliseconds pulsars. *ApJ* 479:948–59

Kalas P, Graham JR, Beckwith SVW, Jewitt DC, Lloyd JP. 2002. Discovery of reflection nebulosity around five Vega-like stars. *ApJ 567:999–1012*

Kalas P, Larwood J, Smith BA, Schultz A. 2000. Rings in the planetesimal disk of β Pictoris. *ApJ* 530:L133–37

Kamp I, van Zadelhoff GJ, van Dishoeck EF, Stark R. 2003. Line emission from circumstellar disks around A stars. *A&A* 397:1129–41

Kant I. 1755. *Allgemeine Naturgeschichte und Theorie des Himmels.* Leipzig

Karlsson A, Mennesson B. 2000. The Robin Laurance nulling interferometers. In *Interferometry in optical astronomy,* ed. PJ Léna, A Quirrenbach, pp. 871–80. SPIE Vol. 4006. Bellingham, WA

Kenyon SJ, Bromley BC. 2002. Dusty rings: signposts of recent planet formation. *ApJ* 577:L35–38

Kiseleva-Eggleton L, Bois E, Rambaux N, Dvorak R. 2002. Global dynamics and stability limits for planetary systems around HD 12661, HD 38529, HD 37124, and HD 160691. *ApJ* 578:L145–48

Kley W, D'Angelo G, Henning T. 2001. Three-dimensional simulations of a planet embedded in a protoplanetary disk. *ApJ* 547:457–64

Knacke RF, Fajardo-Acosta SB, Telesco CM, Hackwell JA, Lynch DK, Russell RW. 1993. The silicates in the disk of β Pictoris. *ApJ* 418:440–50

Knittl Z. 1976. *Optics of thin films.* New York: Wiley. 548 pp.

Koch D, Borucki W, Cullers K, Dunham E, Webster L, et al. 1996. System design of a mission to detect Earth-size planets in the inner orbits of Solar-like stars. *JGR* 101:9297–302

Koch D, Borucki W, Webster L, Dunham E, Jenkins J, et al. 1998. Kepler: a space mission to detect earth-class exoplanets. In *Space telescopes and instruments V,* ed. PY Bely, JB Breckinridge, pp. 599–607. SPIE Vol. 3356. Bellingham, WA

Koerner DW. 1997. Analogs of the early Solar System. *Orig Life and Evol Biosphere* 27:157–84

Koerner DW, Sargent AI, Beckwith SVW. 1993. A rotating gaseous disk around the T Tauri star GM Aurigae. *Icarus* 106:2–10

Konacki M, Torres G, Jha S, Sasselov DD. 2003. An extrasolar planet that transits the disk of its parent star. *Nature* 421:507–09

Konacki M, Wolszczan A. 2003. Masses and orbital inclinations of planets in the PSR B1257+12 system. *ApJ* 591:L147-L150

Kornet K, Bodenheimer P, Różyczka M. 2002. Models of the formation of the planets in the 47 UMa system. *A&A* 396:977–86

Korzennik SG, Brown TM, Fischer DA, Nisenson P, Noyes RW. 2000. A high-eccentricity low-mass companion to HD 89744. *ApJ* 533:L147–50

Kovács G, Zucker S, Mazeh T. 2002. A box-fitting algorithm in the search for periodic transits. *A&A* 391:369–77

Kozai Y. 1962. Secular perturbations of asteroids with high inclination and eccentricity. *AJ* 67:591–98

Krist JE, Stapelfeldt KR, Ménard F, Padgett DL, Burrows C.J. 2000. WFPC2 images of a face-on disk surrounding TW Hydrae. *ApJ* 538:793–800

Krist JE, Stapelfeldt KR, Watson AM. 2002. Hubble Space Telescope/ WFPC2 images of the GG Tauri circumbinary disk. *ApJ* 570:785–92

Kroupa P, Aarseth S, Hurley J. 2001. The formation of a bound star cluster: from the Orion nebula cluster to the Pleiades. *MNRAS* 321: 699–712

Kürster M, Endl M, Els S, Hatzes AP, Cochran WD, et al. 2000. An extrasolar giant planet in an Earth-like orbit. Precise radial velocities of the young star ι Horologii = HR 810. *A&A* 353:L33–36

Lacy CH. 1977. Absolute dimensions and masses of the remarkable spotted dM4e eclipsing binary flare star CM Draconis. *ApJ* 218:444–60

Lagrange AM, Backman DE, Artymowicz P. 2000. Planetary material around main-sequence stars. In *Protostars and planets IV,* ed. V Mannings, AP Boss, SS Russell, pp. 639–72. University of Arizona Press

Lagrange AM, Beust H, Mouillet D, Deleuil M, Feldman PD, et al. 1998. The β Pictoris circumstellar disk. XXIV. Clues to the origin of the stable gas. *A&A* 330:1091–1108

Laplace PS. 1796. *Exposition du système du Monde.* Paris

Latham DW, Mazeh T, Stefanik RP, Mayor M, Burki G. 1989. The unseen companion of HD 114762: a probable brown dwarf. *Nature* 339:38–40

Lattanzi MG, Spagna A, Sozzetti A, Casertano S. 2000. Space-borne global astrometric surveys: the hunt for extrasolar planets. *MNRAS* 317:211–24

Laughlin G, Adams FC. 1999. Stability and chaos in the υ Andromedae planetary system. *ApJ* 526:881–89

Laughlin G, Chambers JE. 2001. Short-term dynamical interactions among extrasolar planets. *ApJ* 551:L109–13

Laughlin G, Chambers JE. 2002. Extrasolar Trojans: the viability and detectability of planets in the 1:1 resonance. *ApJ* 124:592–600

Laughlin G, Chambers JE, Fischer D. 2002. A dynamical analysis of the 47 Ursae Majoris planetary system. *ApJ* 579:455–67

Lawson PR, ed. 2000. *Principles of long baseline stellar interferometry. Course notes from the 1999 Michelson summer school.* JPL Publication 00–009, Pasadena, CA

le Poole RS, Quirrenbach A. 2002. Optimized beam-combination schemes for each channel for PRIMA. In *Interferometry for optical astronomy II,* ed. WA Traub, pp. 496–502. SPIE Vol. 4838. Bellingham, WA

Lecavelier des Etangs A. 1998. Planetary migration and sources of dust in the β Pictoris disk. *A&A* 337:501–11

Lecavelier des Etangs A, Vidal-Madjar A, Ferlet R. 1999. Photometric stellar variation due to extrasolar comets. *A&A* 343:916–22

Lecavelier des Etangs A, Vidal-Madjar A, Roberge A, Feldman PD, Deleuil, M, et al. 2001. Deficiency of molecular hydrogen in the disk of β Pictoris. *Nature* 412:706–08

Lee MH, Peale SJ. 2003. Secular evolution of hierarchical planetary systems. *ApJ* 592:1201–1216

Lestrade JF. 2000a. Stellar VLBI. In *Proceedings of the 5th European VLBI Network Symposium,* ed. JE Conway, AG Polatidis, RS Booth, Y. Pihlström, pp. 155–61. Onsala, Sweden

Lestrade JF. 2000b. Potential of ALMA for extra-solar planet search by astrometry. Unpublished.

Lewis GF, Ibata RA. 2000. Probing the atmospheres of planets orbiting microlensed stars via polarization variability. *ApJ* 539:L63–66

Lin DNC, Bodenheimer P, Richardson DC. 1996. Orbital migration of the planetary companion of 51 Pegasi to its present location. *Nature* 380:606–07

Lin DNC, Ida S. 1997. On the origin of massive eccentric planets. *ApJ* 477:781–91

Lindegren L, Perryman MAC. 1996. GAIA: global astrometric interferometer for astrophysics. *A&AS* 116:579–95

Lineweaver C, Grether D. 2002. The observational case for Jupiter being a typical massive planet. *Astrobiology* vol.2 issue 3, 325–334

Linfield RP, Colavita MM, Lane, BF. 2001. Atmospheric turbulence measurements with the Palomar Testbed Interferometer. *ApJ* 554:505–13

Liseau R, Artymowicz P. 1998. High sensitivity search for molecular gas in the β Pic disk – On the low gas-to-dust mass ratio of the circumstellar disk around β Pictoris. *A&A* 334:935–42

Liseau R, Brandeker A, Fridlund M, Olofsson G, Takeuchi T, Artymowicz P. 2003. The 1.2 mm image of the β Pictoris disk. *A&A* 402:183–187

Lissauer JJ, Rivera EJ. 2001. Stability analysis of the planetary system orbiting υ Andromedae. II. Simulations using new Lick observatory fits. *ApJ* 554:1141–50

Lloyd JP, Oppenheimer BR, Graham JR. 2002. The potential of differential astrometric interferometry from the high Antarctic plateau. *PASA* 19:318–22

Lucas PW, Roche PF, Allard F, Hauschildt PH. 2001. Infrared spectroscopy of substellar objects in Orion. *MNRAS* 326:695–721

Luhman KL, Jayawardhana R. 2002. An adaptive optics search for companions to stars with planets. *ApJ* 566:1132–46

Luu JX, Jewitt DC. 2002. Kuiper Belt objects: relics from the accretion disk of the Sun. *ARAA* 40:63–101

Lyne AG, Bailes M. 1992. No planet orbiting PSR1829−10. *Nature* 355:213

Malfait K, Waelkens C, Waters LBFM, Vandenbussche B, Huygen E, de Graauw MS. 1998. The spectrum of the young star HD 100546 observed with the Infrared Space Observatory. *A&A* 332:L25–28

Mallen-Ornélas G, Seager S, Yee HKC, Minniti D, Gladders MD et al. 2003. The EXPLORE project I: a deep search for transiting extrasolar planets. *ApJ* 582:1123–1140

Mann I. 2001. Dust in the Solar System and in other planetary systems. In *Solar and extra-solar planetary systems, ed. IP Williams, N Thomas,* pp. 218–42. Lecture Notes in Physics, Springer, Berlin

Mao S, Paczyński B. 1991. Gravitational microlensing by double stars and planetary systems. *ApJ* 374:L37–40

Marcy GW, Butler RP. 1992. Precision radial velocities with an iodine absorption cell. *PASP* 104:270–77

Marcy GW, Butler RP. 1995. The planet around 51 Pegasi. *AAS abstract* 187:70.04

Marcy GW, Butler RP. 1996. A planetary companion to 70 Virginis. *ApJ* 464:L147–51

Marcy GW, Butler RP. 1998. Doppler detection of extra-solar planets. In *Cool stars, stellar systems and the Sun, tenth Cambridge workshop,* ed.

Donahue RA, Bookbinder JA, pp. 9–24. ASP Conf. Ser. Vol. 158. San Francisco, CA

Marcy GW, Butler RP. 2000. Planets orbiting other Suns. *PASP* 112:137–40

Marcy GW, Butler RP, Fischer D, Laughlin G, Vogt SS, et al. 2002. A planet at 5 AU around 55 Cancri. *ApJ* 581:1375–88

Marcy GW, Butler RP, Fischer DA, Vogt SS. 2003a. Properties of extrasolar planets. In *Scientific frontiers in research on extrasolar planets,* ed. D Deming, S Seager. ASP Conf. Ser., 294:1–16

Marcy GW, Butler RP, Fischer D, Vogt SS, Lissauer JJ, Rivera EJ. 2001a. A pair of resonant planets orbiting GJ 876. *ApJ* 556:296–301

Marcy GW, Butler RP, Vogt SS. 2000a. Sub-Saturn planetary candidates of HD 16141 and HD 46375. *ApJ* 536:L43–46

Marcy GW, Butler RP, Vogt SS, Fischer D, Lissauer JJ. 1998. A planetary companion to a nearby M4 dwarf, Gliese 876. *ApJ* 505:L147–49

Marcy GW, Butler RP, Vogt SS, Liu MC, Laughlin G, et al. 2001b. Two substellar companions orbiting HD 168443. *ApJ* 555:418–25

Marcy GW, Butler RP, Williams E, Bildsten L, Graham JR, et al. 1997. The planet around 51 Pegasi. *ApJ* 481:926–35

Marcy GW, Fischer DA, Butler RP, Vogt SS. 2003b. Systems of multiple planets. *SSR*, in press

Mardling RA, Lin DNC. 2002. Calculating the tidal, spin, and dynamical evolution of extrasolar planetary systems. *ApJ* 573:829–44

Marley MS, Gelino C, Stephens D, Lunine JI, Freedman, R. 1999. Reflected spectra and albedos of extrasolar giant planets. I. Clear and cloudy atmospheres. *ApJ* 513:879–93

Matthews JM, Kuschnig R, Walker GAH, Pazder J, Johnson R, et al. 2000. Ultraprecise photometry from space: The MOST microsat mission. In *The Impact of Large-Scale Surveys on Pulsating Star Research*, ed. L Szabados, D Kurtz, pp. 74–75. IAU Coll. 176, ASP Conf. Ser. Vol. 203. San Francisco, CA

Mayor M, Queloz D. 1995. A Jupiter-mass companion to a Solar-type star. *Nature* 378:355–59

Mayor M., Udry S., Naef D., Pepe F., Queloz D., et al 2004, The CORALIE survey for southern extra-solar planets. XII Orbital solutions for 16 extra-solar planets detected with CORALIE*A&A* 415:391:402

Mazeh T, Naef D, Torres G, Latham DW, Mayor M, et al. 2000. The spectroscopic orbit of the planetary companion transiting HD 209458. *ApJ* 532:L55–58

Mazeh T, Zucker S, Dalla Torre A, van Leeuwen F. 1999. Analysis of the HIPPARCOS measurements of υ Andromedae: a mass estimate of its outermost known planetary companion. *ApJ* 522:L149–51

McCarthy DW, Probst RG, Low FJ. 1985. Infrared detection of a close cool companion to Van Biesbroeck 8. *ApJ* 290:L9–13

McCaughrean MJ, O'Dell CR. 1996. Direct imaging of circumstellar disks in the Orion Nebula. *AJ* 111:1977–86

McCaughrean MJ, Stapelfeldt KR, Close LM. 2000. High-resolution optical and near-infrared imaging of young circumstellar disks. In *Protostars and planets IV,* ed. V Mannings, AP Boss, SS Russell, pp. 485–507. University of Arizona Press

McLean IS. 1997. *Electronic imaging in astronomy.* Chichester: Praxis Publishing. 472 pp.

Meisner J, le Poole RS. 2002. Dispersion affecting the VLTI and 10 micron interferometry using MIDI. In *Interferometry for optical astronomy II,* ed. WA Traub, SPIE Vol. 4838, in press

Melnick GJ, Neufeld DA, Ford KES, Hollenbach DJ, Ashby MLN. 2001. Discovery of water vapour around IRC+10216 as evidence for comets orbiting another star. *Nature* 412:160–63

Men'shchikov AB, Henning T, Fischer O. 1999. Self-consistent model of the dusty torus around HL Tauri. *ApJ* 519:257–78

Mennesson B, Mariotti JM. 1997. Array configurations for a space infrared nulling interferometer dedicated to the search for earthlike extrasolar planets. *Icarus* 128:202–12

Mennesson B, Ollivier M, Ruilier C. 2002. Use of single-mode waveguides to correct the optical defects of a nulling interferometer. *JOSA A* 19:596–602

Meyer MR, Backman D, Beckwith SVW, Brooke TY, Carpenter JM, et al. 2001. The formation and evolution of planetary systems: SIRTF legacy science in the VLT era. In *The Origin of Stars and Planets: The VLT View*, ESO

Mieremet AL, Braat JJM. 2002a. Nulling interferometry without achromatic phase shifters. *Appl Opt* 41:4697–703

Mieremet AL, Braat JJM. 2002b. Towards the deepest possible null. In *Hunting for planets – GENIE VLTI instrument: a DARWIN technology demonstrator,* in press

Mieremet AL, Braat J, Bokhove H, Ravel K. 2000. Achromatic phase shifting using adjustable dispersive elements. In *Interferometry in optical astronomy,* ed. PJ Léna, A Quirrenbach, pp. 1035–41. SPIE Vol. 4006. Bellingham, WA

Miller SL. 1953. A production of amino acids under possible primitive Earth conditions. *Science* 117:528–29.

Milman MH, Turyshev SG. 2000. Observational model for the Space Interferometry Mission. In *Interferometry in optical astronomy,* ed. PJ Léna, A Quirrenbach, pp. 828–37. SPIE Vol. 4006. Bellingham, WA

Misner CW, Thorne KS, Wheeler JA. 1973. *Gravitation.* New York: W.H. Freeman and Company. 1279 pp.

Mochejska BJ, Stanek KZ, Sasselov DD, Szentgyorgyi AH. 2002. Planets In Stellar Clusters Extensive Search. I. Discovery of 47 low-amplitude variables in the metal-rich cluster NGC 6791 with millimagnitude image subtraction photometry. *AJ* 123:3460–72

Monet D, Bird A, Canzian B, Harris H, Reid N, et al. 1996. *USNO-A1.0.* U.S. Naval Observatory, Washington DC

Mora A, Eiroa C, Natta A, Grady CA, de Winter D, et al. 2003. Dynamics of the circumstellar gas in BF Orionis, SV Cephei, WW Vulpeculae and XY Persei. 2003. *A&A*, submitted

Mora A, Natta A, Eiroa C, Grady CA, de Winter D, et al. 2002. A dynamical study of the circumstellar gas in UX Orionis. *A&A* 393:259–71

Morgan RM, Burge J, Woolf N. 2000. Nulling interferometric beam combiner utilizing dielectric plates: experimental results in the visible broadband. In *Interferometry in optical astronomy,* ed. PJ Léna, A Quirrenbach, pp. 340–48. SPIE Vol. 4006. Bellingham, WA

Mouillet D, Larwood JD, Papaloizou JCB, Lagrange AM. 1997. A planet on an inclined orbit as an explanation of the warp in the β Pictoris disc. *MNRAS* 292:896–904

Moutou C, Coustenis A, Schneider J, St Gilles R, Mayor M, et al. 2001. Search for spectroscopical signatures of transiting HD 209458 B's exosphere. *A&A* 371:260–66

Mozurkewich D, Johnston KJ, Simon RS, Bowers PF, Gaume R, et al. 1991. Angular diameter measurements of stars. *AJ* 101:2207–19

Murray CD, Dermott SF. 1999. *Solar system dynamics.* Cambridge: Cambridge University Press. 592 pp.

Murray N, Chaboyer B. 2002. Are stars with planets polluted? *ApJ* 566:442–51

Murray N, Chaboyer B, Arras P, Hansen B, Noyes RW. 2001. Stellar pollution in the Solar neighborhood. *ApJ* 555:801–15

Naef D, Latham DW, Mayor M, Mazeh T, Beuzit JL. 2001a. HD 80606 b, a planet on an extremely elongated orbit. *A&A* 375:L27–30

Naef D, Mayor M, Pepe F, Queloz D, Santos NC, et al. 2001b. The CORALIE survey for southern extrasolar planets. V. 3 new extrasolar planets. *A&A* 375:205–18

Naef D, Mayor M, Beuzit J-L, Perrier C., Queloz et al. 2004, The ELODIE survey for northern extrasolar planets III: Three planetary candidates detected with ELODIE. it A&A 414:351–359

NASA. 1999a. *SIM Space Interferometry Mission: taking the measure of the Universe,* eds. R Danner & S Unwin. JPL 400–811. Pasadena, CA. 139 pp.

NASA. 1999b. *TPF: Terrestrial Planet Finder,* eds. CA Beichman, NJ Woolf, CA Lindensmith. JPL 99–3. Pasadena, CA. 158 pp.

Natta A, Grinin VP, Tambovtseva LV. 2000. An interesting episode of accretion activity in UX Orionis. *ApJ* 542:421–27

Nauenberg M. 2002a. Determination of masses and other properties of extrasolar planetary systems with more than one planet. *ApJ* 568:369–76

Nauenberg M. 2002b. Stability and eccentricity for two planets in 1:1 resonance, and their possible occurence in extrasolar planetary systems. *ApJ* 124:2332–38

Nelson AF, Angel JRP. 1998. The range of masses and periods explored by radial velocity searches for planetary companions. *ApJ* 500:940–57

Nelson RP, Papaloizou JCB. 2002. Possible commensurabilities among pairs of extrasolar planets. *MNRAS* 333:L26–30

Nisenson P, Contos A, Korzennik S, Noyes R, Brown T. 1999. The Advanced Fiber-optic Echelle (AFOE) and extrasolar planet searches. In *Precise stellar radial velocities,* ed. JB Hearnshaw, CD Scarfe, pp. 143–53. IAU Coll. 170, ASP Conf. Ser. Vol. 185. San Francisco, CA

Noyes RW, Hartmann LW, Baliunas SL, Duncan DK, Vaughan AH. 1984. Rotation, convection, and magnetic activity in lower main-sequence stars. *ApJ* 279:763–77

Noyes RW, Jha S, Korzennik SG, Krockenberger M, Nisenson P, et al. 1997. A planet orbiting the star ρ Coronae Borealis. *ApJ* 483:L111–14

O'Dell CR, Wen Z, Hu X. 1993. Discovery of new objects in the Orion nebula on HST images: shocks, compact sources, and protoplanetary disks. *ApJ* 410:696–700

O'Leary BT. 1966. On the occurrence and nature of planets outside the Solar System. *Icarus* 5:419–36

Olofsson G, Liseau R, Brandeker A. 2001. Widespread atomic gas emission reveals the rotation of the β Pictoris disk. *ApJ* 563:L77–80

Ozernoy LM, Gorkavyi NN, Mather JC, Taidakova TA. 2000. Signatures of exosolar planets in dust debris disks. *ApJ* 537:L147–51

Paczyński B. 1986. Gravitational microlensing by the Galactic halo. *ApJ* 304:1–5

Paczyński B. 1996. Gravitational microlensing in the Local Group. *ARAA* 34:419–59

Pancharatnam S. 1956. Generalized theory of interference, and its applications. Part I. Coherent pencils. *Proc Ind Acad Sci A* 44:247–62

Pantin E, Lagage PO, Artymowicz P. 1997. Mid-infrared images and models of the β Pictoris dust disk. *A&A* 327:1123–36

Papaloizou JCB, Terquem C. 2001. Dynamical relaxation and massive extrasolar planets. *MNRAS* 325:221–30

Patience J, White RJ, Ghez AM, McCabe C, McLean IS, et al. 2002. Stellar companions to stars with planets. *ApJ* 581:654–65

Patterson RJ, Majewski SR, Kundu A, Kunkel WE, Johnston KV, et al. 1999. The Grid Giant Star Survey for the SIM astrometric grid. *AAS abstract* 195:46.03

Pätzold M, Rauer H. 2002. Where are the massive close-in extrasolar planets? *ApJ* 568:L117–20

Paulson DB, Saar SH, Cochran WD, Hatzes AP. 2002. Searching for planets in the Hyades. II. Some implications of stellar magnetic activity. *AJ* 124:572–82

Peale SJ. 1997. Expectations from a microlensing search for planets. *Icarus* 127:269–89

Pepe F, Mayor M, Delabre B, Kohler D, Lacroix, D, et al. 2000. HARPS: a new high-resolution spectrograph for the search of extrasolar planets. In *Optical and IR telescope instrumentation and detectors,* ed. M Iye, AF Moorwood, pp. 582–92. SPIE Vol. 4008. Bellingham, WA

Perrier C, Mariotti JM. 1987. On the binary nature of Van Biesbroeck 8. *ApJ* 312:L27–30

Perryman MAC, de Boer KS, Gilmore G, Høg E, Lattanzi MG, et al. 2001. GAIA: composition, formation and evolution of the Galaxy. *A&A* 369:339–63

Perryman MAC, Peacock, A. 2000. The astronomical potential of optical STJs. In *Imaging the Universe in three dimensions,* ed. W van Breugel, J Bland-Hawthorn, pp. 487–94. ASP Conf. Ser. Vol. 195

Phinney ES, Hansen BMS. 1993. The pulsar planet production process. In *Planets around pulsars,* ed. JA Phillips, SE Thorsett, SR Kulkarni, pp. 371–90. ASP Conf. Ser. Vol. 36

Podsiadlowski P. 1993. Planet formation scenarios. In *Planets around pulsars,* ed. JA Phillips, SE Thorsett, SR Kulkarni, pp. 149–65. ASP Conf. Ser. Vol. 36

Potter DE, Close LM, Roddier F, Roddier C, Graves JE, Northcott M. 2000. A high-resolution polarimetry map of the circumbinary disk around UY Aurigae. *ApJ* 540:422–28

Pravdo SH, Shaklan SB. 1996. Astrometric detection of extrasolar planets: results of a feasibility study with the Palomar 5 Meter Telescope. *ApJ* 465:264–77

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. *Numerical Recipes in C, the art of scientific computing, 2nd edition.* Cambridge: Cambridge University Press. 994 pp.

Queloz D. 1999. Indirect searches: Doppler spectroscopy and pulsar timing. In *Planets outside the Solar System: theory and observations,* ed. JM Mariotti, D Alloin, pp. 229–47. NATO ASI Vol. 532, Dordrecht: Kluwer

Queloz D, Casse M, Mayor M. 1999. The fiber-fed spectrograph, a tool to detect planets. In *Precise stellar radial velocities,* ed. JB Hearnshaw, CD Scarfe, pp. 13–21. IAU Coll. 170, ASP Conf. Ser. Vol. 185. San Francisco, CA

Queloz D, Eggenberger A, Mayor M, Perrier C, Beuzit JL, et al. 2000a. Detection of a spectroscopic transit by the planet orbiting the star HD 209458. *A&A* 359:L13–17

Queloz D, Henry GW, Sivan JP, Baliunas SL, Beuzit JL, et al. 2001. No planet for HD 166435. *A&A* 379:279–87

Quillen AC, Holman M. 2000. Production of star-grazing and star-impacting planetesimals via orbital migration of extrasolar planets *AJ* 119:397–402

Quillen AC, Thorndike S. 2002. Structure in the ε Eridani dusty disk caused by mean motion resonances with a 0.3 eccentricity planet at periastron. *ApJ* 578:L149–52

Quintana EV, Lissauer JJ, Chambers JE, Duncan MJ. 2002. Terrestrial planet formation in the α Centauri system. *ApJ* 576:982–96

Quirrenbach A. 2000a. Astrometry with the VLT Interferometer. In *From extrasolar planets to cosmology: the VLT opening symposium,* ed. J Bergeron, A Renzini, pp. 462–67. Berlin/Heidelberg: Springer-Verlag

Quirrenbach A. 2000b. Observing through the turbulent atmosphere. In *Principles of long baseline stellar interferometry. Course notes from the 1999 Michelson summer school,* ed. PR Lawson, pp. 71–85. JPL Publication 00–009, Pasadena, CA

Quirrenbach A. 2001. Optical interferometry. *ARAA* 39:353–401

Quirrenbach A. 2002a. The Space Interferometry Mission (SIM) and Terrestrial Planet Finder (TPF). In *From optical to millimetric interferometry: scientific and technological challenges,* ed. J Surdej, JP Swings, D Caro, A Detal, pp. 51–67. Université de Liège

Quirrenbach A. 2002b. Site testing and site monitoring for extremely large telescopes. In *Astronomical site evaluation in the visible and radio range,* ed. J Vernin, Z Benkhaldoun C Muñoz-Tuñón, pp. 516–22. ASP Conf. Ser. Vol. 266. San Francisco, CA

Quirrenbach A, Cooke J, Mitchell D, Safizadeh N, Deeg H, Export team. 2000. EXPORT: Search for transits in open clusters with the Jakobus Kapteyn and Lick 1 m telescopes. In *Disks, planetesimals and planets,* ed. F Garzón, C Eiroa, D de Winter, TJ Mahoney, pp. 566–71. ASP Conf. Ser. Vol. 219. San Francisco, CA

Quirrenbach A, Coudé du Foresto V, Daigne G, Hofmann KH, Hofmann R, et al. 1998. PRIMA – study for a dual-beam instrument for the VLT Interferometer. In *Astronomical interferometry,* ed. RD Reasenberg, pp. 807–17. SPIE Vol. 3350. Bellingham, WA

Quirrenbach A, Mozurkewich D, Buscher DF, Hummel CA, Armstrong JT. 1994. Phase-referenced visibility averaging in optical long-baseline interferometry. *A&A* 286:1019–27

Quirrenbach A, Mozurkewich D, Buscher DF, Hummel CA, Armstrong JT. 1996. Angular diameter and limb darkening of Arcturus. *A&A* 312:160–66

Rauer H, Bockelée-Morvan D, Coustenis A, Guillot T, Schneider J. 2000. Search for an exosphere around 51 Pegasi B with ISO. *A&A* 355:573–80

Reid IN. 2002. On the nature of stars with planets. *PASP* 114:306–29

Reuyl D, Holmberg E. 1943. On the existence of a third component in the system 70 Ophiuchi. *ApJ* 97:41–45

Rhee J, Slesnick CL, Crane JD, Polak AA, Patterson RJ, et al. 2001. Preliminary results of the Grid Giant Star Survey (GGSS) for the Space Interferometry Mission (SIM) astrometric grid. *AAS abstract* 198:62.03

Rhie SH, Becker AC, Bennett DP, Fragile PC, Johnson BR, et al. 1999. Observations of the binary microlens event MACHO 98-SMC-1 by the Microlensing Planet Search collaboration. *ApJ* 522:1037–45

Rhie SH, Bennett DP, Becker AC, Peterson BA, Fragile PC, et al. 2000. On planetary companions to the MACHO 98-BLG-35 microlens star. *ApJ* 533:378–91

Ridgway ST, Roddier F. 2000. *An infrared Very Large Array for the 21st century.* In *Interferometry in optical astronomy,* ed. PJ Léna, A Quirrenbach, pp. 940–50. SPIE Vol. 4006. Bellingham, WA

Rivera E, Lissauer JJ. 2001. Dynamical models of the resonant pair of planets orbiting the star GJ 876. *ApJ* 558:392–402

Roberge A, Feldman PD, Lagrange AM. 2000. High-resolution Hubble Space Telescope STIS spectra of C I and CO in the Beta Pictoris circumstellar disk. *ApJ* 538:904–10

Roberge A, Feldman PD, Lecavelier des Etangs A, Vidal-Madjar A, Deleuil M, et al. 2002. Far Ultraviolet Spectroscopic Explorer observations of possible infalling planetesimals in the 51 Ophiuchi circumstellar disk. *ApJ* 568:343–51

Robichon N, Arenou F. 2000. HD 209458 planetary transits from Hipparcos photometry. *A&A* 355:295–98

Robinson LB, Wei MZ, Borucki WJ, Dunham EW, Ford CH, Granados AF. 1995. Test of CCD precision limits for differential photometry. *PASP* 107:1094–98

Roddier C, Roddier F, Northcott MJ, Graves JE, Jim K. 1996. Adaptive optics imaging of GG Tauri: optical detection of the circumbinary ring. *ApJ* 463:326–35

Roddier, F. 1981. The effects of atmospheric turbulence in optical astronomy. *Prog Opt* XIX:281–376

Roddier F. 1989. Optical propagation and image formation through the turbulent atmosphere. In *Diffraction-limited imaging with very large telescopes,* ed. DM Alloin, JM Mariotti, pp. 33–52. NATO ASI Vol. 274, Dordrecht: Kluwer

Romani RW, Miller AJ, Cabrera B, Figueroa-Feliciano E, Nam SW. 1999. First astronomical application of a cryogenic transition edge sensor spectrophotometer. *ApJ* 521:L153–56

Rosenblatt F. 1971. A two-color photometric method for detection of extra solar planetary systems. *Icarus* 14:71–93

Rouan D, Baglin A, Copet E, Schneider J, Barge P, et al. 2000. The exosolar planets program of the COROT satellite. *Earth, Moon, and Planets* 81:79–82

Rubenstein EP, Schaefer BE. 2000. Are superflares on Solar analogues caused by extrasolar planets? *ApJ* 529:1031–33

Ryan P. 2002. Scintillation correlations in the near-infrared. *PASP* 114:462–70

Saar SH, Butler RP, Marcy GW. 1998. Magnetic activity-related radial velocity variations in cool stars: First results from the Lick extrasolar planet survey. *ApJ* 498:L153–57

Saar SH, Donahue RA. 1997. Activity-related radial velocity variation in cool stars. *ApJ* 485:319–27

Saar SH, Fischer D. 2000. Correcting radial velocities for long-term magnetic activity variations. *ApJ* 534:L105–08

Sackett PD. 1999. Searching for unseen planets via occultation and microlensing. In *Planets outside the Solar System: theory and observations,* ed. JM Mariotti, D Alloin, pp. 189–227. NATO ASI Vol. 532, Dordrecht: Kluwer

Sackett PD. 2000. Results from microlensing searches for extrasolar planets. In *Planetary Systems in the UUniverse*, IAU Symp 202 in press

Safizadeh N, Dalal N, Griest K. 1999. Astrometric microlensing as a method of discovering and characterizing extrasolar planets. *ApJ* 522:512–17

Sahu KC, Anderson J, King IR. 2002. A reexamination of the "planetary" lensing events in M 22. *ApJ* 565:L21–24

Sahu KC, Casertano S, Livio M, Gilliland RL, Panagia N, et al. 2001. Gravitational microlensing by low-mass objects in the globular cluster M22. *Nature* 411:1022–24

Salaris M, Weiss A. 1998. Metal-rich globular clusters in the galactic disk: new age determinations and the relation to halo clusters *A&A* 335:943–53

Sandquist E, Taam RE, Lin DNC, Burkert A. 1998. Planet consumption and stellar metallicity enhancements. *ApJ* 506:L65–68

Santos NC, Israelian G, Mayor M. 2001a. The metal-rich nature of stars with planets. *A&A* 373:1019–31

Santos NC, Israelian G, Mayor M, Rebolo R, Udry S. 2003a. Statistical properties of exoplanets. II. Metallicity, orbital parameters, and space velocities. *A&A* 398:363–76

Santos NC, Mayor M, Naef D, Pepe F, Queloz D, et al. 2000. The CORALIE survey for southern extra-solar planets. IV. Intrinsic stellar limitations to planet searches with radial-velocity techniques. *A&A* 361:265–72

Santos NC, Mayor M, Naef D, Pepe F, Queloz D, et al. 2001b. The CORALIE survey for southern extra-solar planets. VI. New long period giant planets around HD 28185 and HD 213240. *A&A* 379:999–1004

Santos NC, Mayor M, Queloz D, Udry S. 2002. Extra-solar planets. *The Messenger*, 110:32–38

Santos NC, Udry S, Mayor M, et al. 2003b, *A&A* 406:373–381

Santos NC et al. 2004, The HARPS survey for southern extra-solar planets: II A 14 Earth-masses exoplanet around μ Arae

Sartoretti P, Schneider J. 1999. On the detection of satellites of extrasolar planets with the method of transits. *A&AS* 134:553–60

Sato B, Ando H, Kambe E, et al. 2003. A planetary Companion to the G-type Giant Star HD 104985, *ApJ* 597:L157–L160

Schneider G, Becklin EE, Smith BA, Weinberger AJ, Silverstone M, Hines DC. 2001. NICMOS coronagraphic observations of 55 Cancri. *ApJ* 121:525–37

Schneider G, Smith BA, Becklin EE, Koerner DW, Meier R, et al. 1999. NICMOS imaging of the HR 4796 A circumstellar disk. *ApJ* 513:L127–30

Schneider G, Wood K, Silverstone MD, Hines DC, Koerner DW, et al. 2003. NICMOS coronagraphic observations of the GM Aurigae circumstellar disk. *ApJ* 125:1467–79

Schneider J, Chevreton M. 1990. The photometric search for Earth-sized extrasolar planets by occultation in binary systems. *A&A* 232:251–57

Schneider P, Weiß A. 1986. The two-point-mass lens – detailed investigation of a special asymmetric gravitational lens. *A&A* 164:237–59

Schroeder DJ. 2000. *Astronomical optics, 2nd edition.* San Diego: Academic Press. 478 pp.

Schultz AB, Jordan I, Hart HM, Bruhweiler F, Fraquelli DA, et al. 2000. Imaging planets about other stars with UMBRAS II. In *Infrared spaceborne remote sensing VIII,* ed. M Strojnik, BF Andresen, pp. 132–40. SPIE Vol. 4131. Bellingham, WA

Schultz AB, Schroeder DJ, Jordan I, Bruhweiler F, DiSanti MA, et al. 1999. Imaging planets about other stars with UMBRAS. In *Infrared spaceborne remote sensing VII,* ed. M Strojnik, BF Andresen, pp. 49–58. SPIE Vol. 3759. Bellingham, WA

Seager S, Hui L. 2002. Constraining the rotation rate of transiting extrasolar planets by oblateness measurements. *ApJ* 574:1004–10

Seager S, Mallén-Ornelas G. 2002. On the unique solution of planet and star parameters from an extrasolar planet transit light curve. *ApJ* 585:1038–1055

Seager S, Sasselov DD. 1998. Extrasolar giant planets under strong stellar irradiation. *ApJ* 502:L157–61

Seager S, Sasselov DD. 2000. Theoretical transmission spectra during extrasolar giant planet transits. *ApJ* 537:916–21

Seager S, Whitney BA, Sasselov DD. 2000. Photometric light curves and polarization of close-in extrasolar giant planets. *ApJ* 540:504–20

Sekanina Z. 2002. Statistical investigation and modeling of sungrazing comets discovered with the Solar and Heliospheric Observatory. *ApJ* 566:577–98

Serabyn E. 1999. Nanometer-level path-length control scheme for nulling interferometry. *Appl Opt* 38:4213–16

Serabyn E. 2000. Nulling interferometry: symmetry requirements and experimental results. In *Interferometry in optical astronomy,* ed. PJ Léna, A Quirrenbach, pp. 328–39. SPIE Vol. 4006. Bellingham, WA

Serabyn E, Colavita MM. 2001. Fully symmetric nulling beam combiners. *Appl Opt* 40:1668–71

Serabyn E, Wallace JK, Hardy GJ, Schmidtlin EGH, Nguyen HT. 1999. Deep nulling of visible laser light. *Appl Opt* 38:7128–32

Shaklan S. 1990. Fiber optic beam combiner for multiple-telescope interferometry. *Opt Eng* 29:684–89

Shaklan SB, Roddier F. 1988. Coupling starlight into single-mode fiber optics. *Appl Opt* 27:2334–38

Shao M, Colavita MM. 1992. Potential of long-baseline infrared interferometry for narrow-angle astrometry. *A&A* 262:353–58

Shao M, Colavita MM, Hines B, Staelin D, Hutter DJ, et al. 1988. The Mark III stellar interferometer. *A&A* 193:357–71

Shu FH, Adams FC, Lizano S. 1987. Star formation in molecular clouds: observation and theory. *ARAA* 25:23–81

Shu F, Najita J, Ostriker E, Wilkin F, Ruden S, Lizano S. 1994. Magnetocentrifugally driven flows from young stars and disks. I. A generalized model. *ApJ* 429:781–96

Siess L, Livio M. 1999. The accretion of brown dwarfs and planets by giant stars – II. Solar-mass stars on the red giant branch. *MNRAS* 308:1133–49

Sigurdsson S. 1992. Planets in globular clusters? *ApJ* 399:L95–97

Sigurdsson S. 1993. Genesis of a planet in Messier 4. *ApJ* 415:L43–46

Sigurdsson S. 1995. Assessing the environmental impact on PSR B1620−26 in M4. *ApJ* 452:323–31

Skrutskie MF, Forrest WJ, Shure MA. 1987. Direct infrared imaging of VB 8. *ApJ* 312:L55–58

Smith BA, Terrile RJ. 1984. A circumstellar disk around β Pictoris. *Science* 226:1421–24

Snellgrove MD, Papaloizou JCB, Nelson RP. 2001. On disc driven inward migration of resonantly coupled planets with application to the system around GJ 876. *A&A* 374:1092–99

Sozzetti A, Casertano S, Lattanzi MG, Spagna A. 2001. Detection and measurement of planetary systems with GAIA. *A&A* 373:L21–24

Spangler C, Sargent AI, Silverstone MD, Becklin EE, Zuckerman B. 2001. Dusty debris around Solar-type stars: temporal disk evolution. *ApJ* 555:932–44

Stapelfeldt KR, Krist JE, Ménard F, Bouvier J, Padgett DL, Burrows CJ. 1998. An edge-on circumstellar disk in the young binary system HK Tauri *ApJ* 502:L65–69

Steinacker J, Henning T. 2003. Detection of gaps in circumstellar disks. *ApJ* 583:L35–38

Stern SA. 1994. The detectability of extrasolar terrestrial and giant planets during their luminous final accretion. *AJ* 108:2312–17

Stern SA, Shull JM, Brandt JC. 1990. Evolution and detectability of comet clouds during post-main-sequence stellar evolution. *Nature* 345:305–08

Stetson PB. 1987. DAOPHOT – A computer program for crowded-field stellar photometry. *PASP* 99:191–222

Strand KA. 1943. 61 Cygni as a triple system. *PASP* 55:29–32

Street RA, Horne K, Lister TA, Penny A, Tsapras Y, et al. 2002. Variable stars in the field of open cluster NGC 6819. *MNRAS* 330:737–54

Street RA, Horne K, Lister TA, Penny A, Tsapras Y, et al. 2003. Searching for planetary transits in the field of open cluster NGC 6819. *MNRAS* 340:1287–97

Street RA, Horne K, Penny A, Tsapras Y, Quirrenbach A, et al. 2000. A search for planetary transits in open clusters. In *Disks, planetesimals and planets,* ed. F Garzón, C Eiroa, D de Winter, TJ Mahoney, pp. 572–77. ASP Conf. Ser. Vol. 219. San Francisco, CA

Strom RG, Smolders B, van Ardenne A. 2001. Active adaptive arrays: the ASRTRON approach to SKA. *ApSS* 278:209–12

Struck C, Cohanim BE, Willson LA. 2002. Models of planets and brown dwarfs in Mira winds. *ApJ* 572:L83–86

Struve O. 1952. Proposal for a project of high-precision stellar radial velocity work. *The Observatory* 72:199–200

Sudarsky D, Burrows A, Pinto P. 2000. Albedo and reflection spectra of extrasolar giant planets. *ApJ* 538:885–903

Sudarsky D, Burrows A, Hubeny I. 2003. Theoretical spectra and atmospheres of extrasolar giant planets. *ApJ* 588:1121–48

Szostak JW, Bartel DP, Luisi PL. 2001. Synthesizing life. *Nature* 409:387–90

Tabachnik S, Tremaine S. 2002. Maximum likelihood method for estimating the mass and period distributions of extra-solar planets. *MNRAS* 335:151–58

Tallon M, Tallon-Bosc I. 1992. The object-image relationship in Michelson stellar interferometry. *A&A* 253:641–45

Tarter J. 2001. The search for extraterrestrial intelligence (SETI). *ARAA* 39:511–48

Tatarski, V.I. 1961. *Wave propagation in a turbulent medium.* New York: McGraw-Hill. 285 pp.

Telesco CM, Fisher RS, Piña RK, Knacke RF, Dermott SF, et al. 2000. Deep 10 and 18 micron imaging of the HR 4796 A circumstellar disk: transient dust particles and tentative evidence for a brightness asymmetry. *ApJ* 530:341

Thi WF, Blake GA, van Dishoeck EF, van Zadelhoff GJ, Horn JMM, et al. 2001. Substantial reservoir of molecular hydrogen in the debris disks around young stars. *Nature* 409:60–63

Thompson AR, Moran JM, Swenson GW. 1986. *Interferometry and synthesis in radio astronomy.* New York: Wiley. 534 pp.

Thorsett SE, Arzoumanian Z, Camilo F, Lyne AG. 1999. The triple pulsar system PSR B1620−26 in M4. *ApJ* 523:763–70

Tough A. 2000. Five strategies for detecting intelligence. In *Bioastronomy '99 – a new era in bioastronomy,* ed. GA Lemarchand, KJ Meech, pp. 445–47. ASP Conf. Ser. Vol. 213. San Francisco, CA

Trilling DE, Benz W, Guillot T, Lunine JI, Hubbard WB, Burrows A. 1998. Orbital evolution and migration of giant planets: modeling extrasolar planets. *ApJ* 500:428–39

Trilling DE, Brown RH. 1998. A circumstellar dust disk around a star with a known planetary companion. *Nature* 395:775–77

Trilling D.E., 2000, *ApJ*, 537, 61–64

Trimble V. 1999. Milky Way and galaxies. In *Allen's astrophysical quantities,* ed. AN Cox, pp. 569–83. New York: Springer

Tsapras Y, Street RA, Horne K, Penny A, Clarke F, et al. 2001. Can Jupiters be found by monitoring Galactic bulge microlensing events from northern sites? *MNRAS* 325:1205–12

Udalski A, Paczyński B, Żebruń K, Szymański M, Kubiak M, et al. 2002. The Optical Gravitational Lensing Experiment. Search for planetary and low-luminosity object transits in the Galactic disk. Results of 2001 campaign. *Acta Astron* 52:115–128

Udalski A, Szymański M, Kałużny J, Kubiak M, Krzemiński W, et al. 1993. The Optical Gravitational Lensing Experiment. Discovery of the first candi-

date microlensing event in the direction of the Galactic bulge. *Acta Astron* 43:289–94

Udalski A, Szymański M, Mao S, di Stefano R, Kałużny J, et al. 1994. The optical gravitational lensing experiment: OGLE no. 7: Binary microlens or a new unusual variable? *ApJ* 436:L103–06

Udalski A, Żebruń K, Szymański M, Kubiak M, Pietrzyński G, et al. 2000. The Optical Gravitational Lensing Experiment. Catalog of microlensing events in the Galactic bulge. *Acta Astron* 50:1–65

Udry S, Mayor M, Clausen JV, Freyhammer LM, Helt BE, et al. 2003a. The CORALIE survey for southern extra-solar planets. X. A hot Jupiter orbiting HD 73256. *A&A* 407:679–684

Udry S, Mayor M, Naef D, Pepe F, Queloz D, et al. 2000. The CORALIE survey for southern extra-solar planets. II. The short-period planetary companions to HD 75289 and HD 130322. *A&A* 356:590–98

Udry S, Mayor M, Naef D, Pepe F, Queloz D, et al. 2002. The CORALIE survey for southern extra-solar planets – VIII. The very low-mass companions of HD 141937, HD 162020, HD 168443 and HD 202206: Brown dwarfs or "superplanets"? *A&A* 390:267–79

Udry S, Mayor M, Queloz D. 2001. CORALIE-ELODIE new planets and planetary systems. Looking for fossil traces of formation and evolution. In *Planetary Systems in the Universe,* ed. A Penny, P Artymowicz, AM Lagrange, S Russel. IAU Symp. 202, ASP Conf. Ser. San Francisco, CA, in press

Udry S, Mayor M, Santos NC. 2003b. Statistical properties of exoplanets. I. The period distribution: constraints for the migration scenario. *A&A*, 407:369–376

van de Kamp P. 1963. Astrometric study of Barnard's Star from plates taken with the 24-inch Sproul refractor. *AJ* 68:515–21

van Dishoeck EJ, Blake GA. 1998. Chemical evolution of star-forming regions. *ARAA* 36:317–68

Van Hamme W. 1993. New limb-darkening coefficients for modeling binary star light curves. *AJ* 106:2096–117

Vidal-Madjar A, Lagrange-Henri AM, Feldman PD, Beust H, Lissauer JJ. 1994. HST-GHRS observations of β Pictoris: additional evidence for infalling comets. *A&A* 290:245–58

Vogt SS. 1987. The Lick Observatory Hamilton Echelle Spectrometer. *PASP* 99:1214–28

Vogt SS, Butler RP, Marcy GW, Fischer DA, Pourbaix D, et al. 2002. Ten low-mass companions from the Keck precision velocity survey. *ApJ* 568:352–62

Vogt SS, Marcy GW, Butler RP, Apps K. 2000. Six new planets from the Keck precision velocity survey. *ApJ* 536:902–14

Wahhaj Z, Koerner DW, Ressler ME, Werner MW, Backman DE, Sargent AI. 2003. The Inner Rings of β Pictoris. *ApJ* 584:L27–31

Walker GAH, Walker AR, Irwin AW, Larson AM, Yang SLS, Richardson DC. 1995. A search for Jupiter-mass companions to nearby stars. *Icarus* 116:359–75

Wallace K, Hardy G, Serabyn E. 2000. Deep and stable interferometric nulling of broadband light with implications for observing planets around nearby stars. *Nature* 406:700–02

Wallner O, Kudielka K, Leeb WR. 2001. Nulling interferometry for spectroscopic investigation of exoplanets – a statistical analysis of imperfections. In *The search for extraterrestrial intelligence (SETI) in the optical spectrum III,* ed. SA Kingsley, R Bhathal, pp. 47–55. SPIE Vol. 4273. Bellingham, WA

Wambsganss J. 1997. Discovering Galactic planets by gravitational microlensing: magnification patterns and light curves. *MNRAS* 284:172–88

Worley CE and Douglass GG, 1997, A8A5 125, 523. "The Wostinjtan Double star catolog"

Weinberger AJ, Becklin EE, Schneider G, Smith BA, Lowrance PJ, et al. 1999. The circumstellar disk of HD 141569 imaged with NICMOS. *ApJ* 525:L53–56

Weinberger AJ, Becklin EE, Zuckerman B. 2003. First spatially resolved mid-infrared spectroscopy of β Pictoris. *ApJ* 584:L33–37

Weinberger AJ, Rich RM, Becklin EE, Zuckerman B, Matthews K. 2000. Stellar companions and the age of HD 141569 and its circumstellar disk. *ApJ* 544:937–43

Wetherill GW. 1990. Formation of the Earth. *Ann Rev Earth Planet Sci* 18:205–56

Wiedemann G, Deming D, Bjoraker G. 2001. A sensitive search for methane in the infrared spectrum of τ Bootis. *ApJ* 546:1068–74

Willacy K, Cherchneff I. 1998. Silicon and sulphur chemistry in the inner wind of IRC+10216. *A&A* 330:676–84

Wolf S, Gueth F, Henning T, Kley W. 2002. Detecting planets in protoplanetary disks: a prospective study. *ApJ* 566:L97–99

Wolszczan A. 1994. Confirmation of Earth mass planets orbiting the millisecond pulsar PSR B1257+12. *Science* 264:538–42

Wolszczan A. 1999. Detecting planets around pulsars. In *Pulsar timing, general relativity and the internal structure of neutron stars,* ed. Z Arzoumanian, F van der Hooft, EPJ van den Heuvel, pp. 101–11. Koninklijke Nederlandse Akademie van Wetenschappen, Amsterdam

Wolszczan A, Doroshenko O, Konacki M, Kramer M, Jessner A, et al. 2000. Timing observations of four millisecond pulsars with the Arecibo and Effelsberg radio telescopes. *ApJ* 528:907–12

Wolszczan A, Frail DA. 1992. A planetary system around the millisecond pulsar PSR 1257+12. *Nature* 355:145–47

Woolf N, Angel JRP. 1997. Planet Finder options I: new linear nulling array configurations. In *Planets beyond the Solar System and the next generation of space missions,* ed. D Soderblom, pp. 285–93. ASP Conf. Ser. Vol. 119. San Francisco, CA

Worley CE, Douglass GG. 1997. The Washington Double Star Catalog. *A & A* 125:523

Wyatt MC, Dermott SF, Telesco CM, Fisher RS, Grogan K, et al. 1999. How observations of circumstellar disk asymmetries can reveal hidden planets: pericenter glow and its application to the HR 4796 disk. *ApJ* 527:918–44

Youdin AN, Shu FH. 2002. Planetesimal formation by gravitational instability. *ApJ* 580:494–505

Yura HT, Fried DL. 1998. Variance of the Strehl ratio of an adaptive optics system. *JOSA A* 15:2107–10

Zapatero Osorio MR, Béjar VJS, Martín EL, Rebolo R, Barrado y Navascués D, et al. 2000. Discovery of young, isolated planetary mass objects in the σ Orionis star cluster. *Science* 290:103–07

Zapatero Osorio MR, Béjar VJS, Martín EL, Rebolo R, Barrado y Navascués D, et al. 2002. A methane, isolated planetary-mass object in Orion. *ApJ* 578:536–42

Zarka P, Treumann RA, Ryabov BP, Ryabov VB. 2001. Magnetically-driven planetary radio emissions and application to extrasolar planets. *ApSS* 277:293–300

Zucker S, Mazeh T. 2001a. Analysis of the Hipparcos observations of the extrasolar planets and the brown dwarf candidates. *ApJ* 562:549–57

Zucker S, Mazeh T. 2001b. Derivation of the mass distribution of extrasolar planets with MAXLIMA, a maximum likelihood algorithm. *ApJ* 562:1038–44

Zucker S, Mazeh T. 2002. On the mass-period correlation of the extrasolar planets. *ApJ* 568:L113–16

Zucker S, Naef D, Latham D, Mayor M, Mazeh T, et al. 2002. A Planet candidate in the stellar triple system HD 178911. *ApJ* 568:363–68

Zuckerman B. 2001. Dusty circumstellar disks. *ARAA* 39:549–80

Zuckerman B, Becklin EE. 1992. Companions to white dwarfs: very low-mass stars and the brown dwarf candidate GD 165 B. *ApJ* 386:260–64

Zuckerman B, Song I, Bessell MS, Webb RA. 2001. The β Pictoris moving group. *ApJ* 562:L87–90