Dynamics of Planetary Systems

Revised and extended, 2007

Hans Rickman

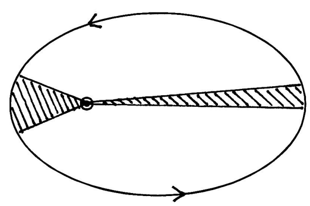
Dept. of Astronomy & Space Physics, Uppsala Univ.
PAN Space Research Center, Warsaw

These notes aim to present only the basics of our understanding of how bodies move in planetary systems, and naturally, the discussion will focus on our Solar System. For a stringent and comprehensive analysis including derivations of some formulae that we shall cite below, we refer to courses on celestial mechanics.

Kepler's laws and Newton's law of gravity

Found empirically by Johannes Kepler in the early 17th century from the observations of planetary positions aqcuired most importantly by his mentor, Tycho Brahe, these laws state that:

- the orbits of planets are ellipses, where the Sun occupies one of the foci;
- in equal times, the radius vector of a planet sweeps out equal areas, independent of its orbital position;
- the square of the period of revolution is proportional to the cube of the semimajor axis of the orbit.



These laws found their explanation with Isaac Newton's law of gravity, according to which a planet moves under the influence of a gravitational force, subject to an acceleration

$$\ddot{\mathbf{r}} = -\frac{G(M_\odot + M_p)}{r^2}\,\widehat{\mathbf{r}}. \tag{1}$$

Here $\mathbf{r}$ is the radius vector from the Sun to the planet and $\widehat{\mathbf{r}}$ is the corresponding unit vector, r is its modulus (the distance Sun-planet), G is the gravitational constant, and $M_\odot$ and M_p are the masses of the Sun and the planet. Note that Eq. (1) describes the motion of the planet with respect to the Sun, *i.e.*, in a non-inertial frame of reference. In an inertial frame like that of the center of mass, the law of gravity would look the same, replacing $(M_\odot + M_p)$ by $M_\odot$.

Let us assume for simplicity that all other forces affecting the planet (most importantly, the gravitational attractions of the other planets) can be neglected. The force field is then both *conservative* and *central*, which means that both the total energy (kinetic + potential) and the angular momentum are constants ("integrals") of motion. Given that the force varies as the inverse square of the distance (r^{-2}), an analytic solution of this "gravitational two-body problem" is obtained straightforwardly in the form of a conic section (*i.e.*, ellipse, parabola or hyperbola) satisfying all three of Kepler's laws.

If the law of gravity had not been inverse-square in r, but followed some other arbitrary function $f(r)$, the orbit would not have been a conic section with the Sun in a focus. In general, it had not been a closed curve at all, but a rosette curve oscillating between a minimum and a maximum distance ($r_{\rm min}$ and $r_{\rm max}$), gradually filling up the annulus described by the circles $r = r_{\rm min}$ and $r = r_{\rm max}$. Even in such a case, however, something analogous to Kepler's second law would have followed from the conservation of angular momentum.

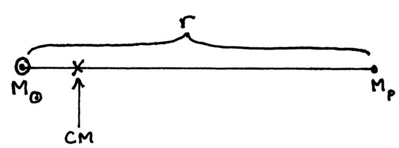
Kepler's third law and mass determination

The only one of Kepler's laws that has a potential for solving physical problems is the third one. However, for this purpose it has to be formulated in a comprehensive way. The exact form is:

$$\frac{4\pi^2}{P^2}\,a^3 = G\,(M_\odot + M_p) = GM_\odot\left(1 + \frac{M_p}{M_\odot}\right), \tag{2}$$

where P is the orbital period and a is the semimajor axis. Table 1 summarizes the values of a, P, $m_p = M_p/M_\odot$, and the mean orbital velocity $v_p = 2\pi a/P$, for the four terrestrial and the four giant planets of the Solar System. The unit of length is the astronomical unit ($1AU = 1.496 \cdot 10^8$ km), the unit of time is the year (1 year $= 3.156 \cdot 10^7$ s), and the unit of velocity is km/s. We note that m_p is always $<< 1$ and may often be neglected in Eq. (2).

It is important to note that in an inertial frame of reference, only the center of mass may be at rest, and thus both the Sun and the planet are moving (cf. the note to Eq. (1) above). Let us now consider the orbital motions of the Sun with respect to the different planets.



The distances of the Sun and the planet from the center of mass are $rM_p/(M_\odot + M_p) = rm_p/(1+m_p)$ and $rM_\odot/(M_\odot + M_p) = r/(1+m_p)$, respectively. The orbital semimajor axis

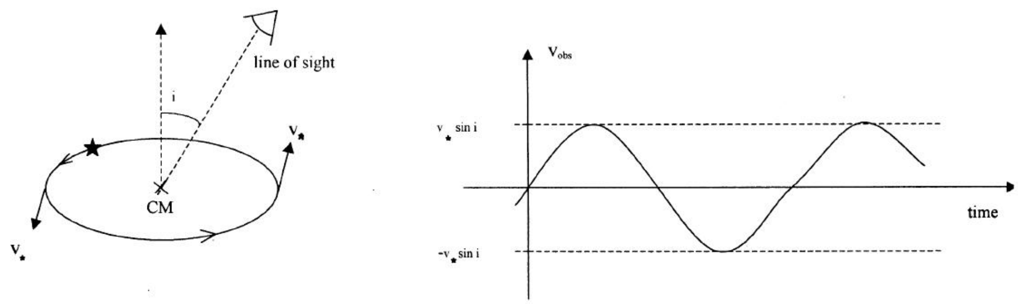
Table 1:

Planet	a (AU)	P (yr)	m_p	v_p (km/s)
Mercury	0.39	0.24	$1.7 \cdot 10^{-7}$	48
Venus	0.72	0.62	$2.4 \cdot 10^{-6}$	35
Earth	1.00	1.00	$3.0 \cdot 10^{-6}$	30
Mars	1.52	1.88	$3.2 \cdot 10^{-7}$	24
Jupiter	5.20	11.9	$9.5 \cdot 10^{-4}$	13
Saturn	9.55	29.4	$2.9 \cdot 10^{-4}$	10
Uranus	19.2	83.7	$4.4 \cdot 10^{-5}$	6.8
Neptune	30.1	164	$5.2 \cdot 10^{-5}$	5.5

and mean velocity of the Sun are $am_p/(1+m_p)$ and $2\pi am_p/P(1+m_p)$, or approximately, am_p and $2\pi am_p/P = v_p m_p$.

Exercise. Arrange the eight planets of Table 1 in order of decreasing semimajor axis and mean velocity of the Sun, respectively!

For the case of Jupiter, we get 0.005 AU and 0.013 km/s. The values for the Earth are much smaller in both cases. For comparison, the radius of the solar photosphere is $7 \cdot 10^5$ km $\simeq$ 0.005 AU. The measurement of such displacements for stars is currently not possible with astrometric observations. Hence the detection of exoplanets so far mainly rests on spectroscopic observations of stellar Doppler shifts caused by the orbital velocities. Note that in this case too, a very high accuracy is required.



Consider for simplicity a planet with mass M_p, which moves in a circular orbit around a star of mass M_*. The line of sight makes an angle i with the normal to the orbital plane. Superimposed on the other contributions to the radial velocity of the star, there will hence be a sinusoidal variation with period P and half-amplitude $v_* \sin i$, where v_* is the orbital velocity of the star in its circular motion around the center of mass. P and $v_* \sin i$ can be measured, but i is in general unknown. Hence v_* is also unknown, but the measured value of $v_* \sin i$ is a lower limit.

Applying Eq. (2) to the star-planet system, and assuming that $M_p << M_*$, we can solve for the orbital velocity of the planet:

$$v_p \simeq \left(\frac{2\pi G M_*}{P}\right)^{1/3} \tag{3}$$

M_* has to be estimated from the spectral type of the star, but due to the 1/3-power dependence, v_p is relatively insensitive to errors in this estimate. The mass of the planet can then be obtained from:

$$M_p = M_* \cdot \frac{v_*}{v_p} \tag{4}$$

and substituting $v_* \sin i$ for v_* yields a lower limit for M_p.

Note that detection of an exoplanet is easiest, if P is short and v_* is large. The latter criterion is favoured if M_p is large, v_p is large, or M_* is small. In conclusion, detectability is optimized for tight systems with heavy planets around low-mass stars.

For mass determinations within our Solar System we make use of the simultaneous availability of values for a and P in systems where one object orbits another. For instance, the mass of the Sun is readily obtained from Eq. (2) using the values in Table 1 and $G = 6.66 \cdot 10^{-11}$ kg^{-1}m^{3}s^{-2}.

Table 2:

Star	Type	Period (days)	$v_* \sin i$ (m/s)	a_p (AU)	min. M_p ($M_{Jup.}$)
51 Peg	G2.5 V	4.23	55	0.051	0.45
υ And	F7 V	4.61	74	0.056	0.65
Gliese 876	M9 V	61.1	220	0.20	1.9
16 Cyg B	G2.5 V	802	44	1.7	1.7
14 Her	K0 V	1 619	80	2.5	3.3

Exercise. Verify the value of $M_\odot = 2.0 \cdot 10^{30}$ kg using several planets!

Another example is the determination of masses of planets with natural satellites, which has been possible since a long time. In this case, combining two versions of Eq. (2) for the Sun-planet and planet-satellite systems, we get:

$$m_p = \frac{M_p}{M_\odot} \simeq \left(\frac{a_{\rm sat}}{a_{\rm pl}}\right)^3 \left(\frac{P_{\rm sat}}{P_{\rm pl}}\right)^{-2} \tag{5}$$

assuming that $M_p << M_\odot$ and $M_{\rm sat} << M_p$. Table 3 gives a few examples of satellites of giant planets, from which the corresponding values of m_p can be computed. In addition, Pluto has been included so that its very small mass can be verified.

Table 3:

Satellite	Planet	$a_{\rm sat}$ (km)	$P_{\rm sat}$ (days)
Io	*JUPITER*	$4.2 \cdot 10^5$	1.77
Dione	*SATURN*	$3.8 \cdot 10^5$	2.74
Titania	*URANUS*	$4.4 \cdot 10^5$	8.71
Triton	*NEPTUNE*	$3.5 \cdot 10^5$	5.88
Charon	*PLUTO*	$2.0 \cdot 10^4$	6.39

Exercise. Verify the values of m_p listed in Table 1, using the values of $a_{\rm sat}$, $P_{\rm sat}$, a_p and P_p!

Modern values of planetary masses rely more heavily on the tracking of artificial satellites or space probes. This is particularly important for planets lacking natural satellites (Mercury and Venus). One does not use Eq. (5) directly, since the observables are not $a_{\rm sat}$ and $P_{\rm sat}$ but rather the accurate distance and radial velocity of the satellite using the radio communication link. This holds not only when the probe orbits the planet but also during a flyby, when the probe trajectory is deflected by the gravity of the planet. But just like the case of Kepler's third law, one uses the strength of the gravity field to measure the mass of the planet.

Recent high-resolution imaging has revealed a number of satellites around near-Earth asteroids, main-belt asteroids and members of the transneptunian population. Periods are easy to measure, but distances between the object and the satellite are less accurate, and in general the masses derived are not yet very reliable.

Other methods to determine masses of objects in the Solar System include mutual perturbations of heliocentric orbits (to be described in the next Section), and modelling of nongravitational perturbations (comets).

Orbital elements and planetary perturbations

Consider first the two-body problem Sun-planet as described by Eq. (1). Using Cartesian coordinates, we can rewrite Eq. (1) as six first-order differential equations for position (x, y, z) and velocity (u, v, w):

$$\dot{x} = u; \qquad \dot{y} = v; \qquad \dot{z} = w$$

$$\dot{u} = -\frac{G(M_\odot + M_p)}{r^3}x; \qquad \dot{v} = -\frac{G(M_\odot + M_p)}{r^3}y; \qquad \dot{w} = -\frac{G(M_\odot + M_p)}{r^3}z$$

Given values of all six phase-space coordinates (x, y, z, u, v, w) at a certain time t_o, this system can be integrated in order to find the corresponding values $x(t), y(t), z(t), u(t), v(t), w(t)$ at all times. As we already mentioned, these solutions take the form of conic sections. In order to specify the solution that holds for a given object, one may use the initial values $x_o, y_o, z_o, u_o, v_o, w_o$ at $t = t_o$, but one may also use any set of six independent quantities which are constant and describe the conic section uniquely. Such sets of geometric quantities are called **orbital elements**.

The most commonly used orbital elements for planets and asteroids are:

a – the semimajor axis of the ellipse;

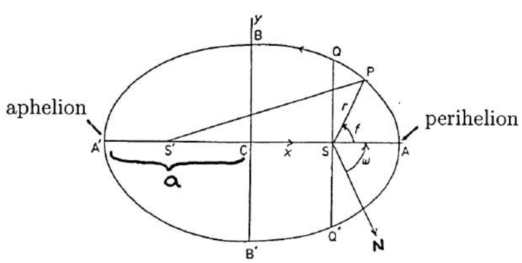
e – the eccentricity of the ellipse;

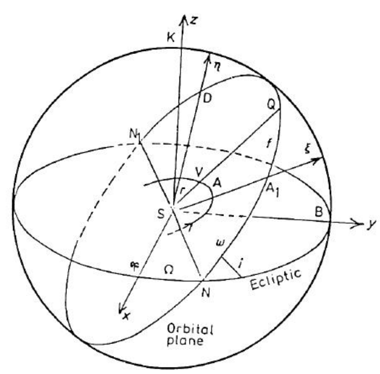
i – the inclination, *i.e.*, the angle between the plane of the ellipse and a reference plane (usually the ecliptic plane);

Ω – the longitude of the ascending node, *i.e.*, the angle along the ecliptic from a reference direction to the point where the object crosses the ecliptic northbound;

ω – the argument of perihelion, *i.e.*, the angle along the orbit from the ascending node to the direction of perihelion, where the object is closest to the Sun;

T – the time of perihelion passage.





Some useful relations follow:

$$\left.\begin{array}{ll} \text{perihelion distance} & q = a(1-e) \\ \text{aphelion distance} & Q = a(1+e) \end{array}\right\} \qquad (6)$$

$$\left.\begin{array}{ll} \text{binding energy} & E = -\frac{GM_\odot}{2a} \\ \text{speed of motion} & v^2 = GM_\odot\left(\frac{2}{r} - \frac{1}{a}\right) \\ \text{angular momentum} & H = \sqrt{GM_\odot a(1-e^2)} \end{array}\right\} \qquad (7)$$

These formulae are unproblematic for elliptic orbits ($a > 0$; $e < 1$), but with the exception of Q, they can also be used in other cases. For parabolas ($a \to \infty$; $e = 1$) q is well-defined, and we have $E = 0$; $H = \sqrt{2GM_{\odot}q}$. For hyperbolae we simply count a as negative and solve for q from Eq. (6), and E, v and H from Eq. (7).

Needless to say, for real planetary systems including the Solar System the two-body problem is a highly idealized model. Independent of whether Eq. (1) is applied to a planet or a small body, we may consider that there is always at least one other planet around, whose gravitational influence has to be taken into account. In principle, one may then resort to numerical integration of the complete equations of motion including all the gravitational forces of the system, but this method has practical difficulties due to uncertainties over the exact initial values of the phase space coordinates and errors inherent to the technique of numerical integration (*e.g.*, round-off errors). Analytical and semi-analytical results are therefore very important as well to understand this complex dynamical problem.

Consider for simplicity just one perturbing planet with mass M_1 and heliocentric radius vector $\mathbf{r}_1$, and assume that the object under study ("test body") has negligible mass ($m_p \simeq 0$). This is called the *restricted three-body problem*, and Eq. (1) transforms into:

$$\ddot{\mathbf{r}} = -\frac{GM_{\odot}}{r^3}\mathbf{r} - \frac{GM_1}{|\mathbf{r} - \mathbf{r}_1|^3}(\mathbf{r} - \mathbf{r}_1) - \frac{GM_1}{r_1^3}\mathbf{r}_1 \tag{8}$$

The three terms of the right-hand member represent (1) the acceleration exerted by the Sun on the test body, (2) the acceleration exerted by the perturbing planet on the test body, and (3) minus the acceleration by the planet on the Sun. Small departures from the Keplerian motion with constant orbital elements as given by the first term are called *perturbations*, and the second term is usually referred to as the source of **direct perturbations**, while the third term leads to **indirect perturbations**.

If close encounters occur, so that $|\mathbf{r} - \mathbf{r}_1|$ becomes very small, the direct perturbation may dominate over the solar gravity even if $M_1 << M_{\odot}$. We shall look at such cases later on (in the Solar System they involve comets and some asteroids), but for the present we limit ourselves to cases where close encounters do not occur because the orbits of the test body and perturbing planet do not intersect or come close to each other.

Eq. (8) may be rewritten as:

$$\ddot{\mathbf{r}} = -\nabla\left(-\frac{GM_{\odot}}{r}\right) - \nabla R_1 \tag{9}$$

with

$$R_1 = -GM_1\left\{\frac{1}{|\mathbf{r} - \mathbf{r}_1|} - \frac{\mathbf{r}_1 \cdot \mathbf{r}}{r_1^3}\right\} \tag{10}$$

Here ∇ denotes the gradient vector and $\mathbf{r}_1 \cdot \mathbf{r}$ is the scalar product of the vectors $\mathbf{r}_1$ and $\mathbf{r}$. The first term on the right-hand side of Eq. (9) is the acceleration in the central force field of the Sun with the potential $-GM_{\odot}/r$, and the second term is a much smaller perturbing acceleration caused by a potential R_1, which is usually called the **perturbing function**,

Now, if M_1 and R_1 were zero, the orbital elements $(a, e, i, \Omega, \omega, T)$ would be strictly constant. The fact that R_1 is not zero but nonetheless very small means that the orbital elements will change slowly with time. One of the basic tasks of classical celestial mechanics is to set up differential equations for $(\dot{a}, \dot{e}, di/dt, \dot{\Omega}, \dot{\omega}, \dot{T})$ as functions of the derivatives of R_1 with respect to $(a, e, i, \Omega, \omega, T)$, *i.e.*, as functions of the orbital elements and time, and to seek solutions that are valid over as long time as possible.

Stability and secular perturbations

The classical method is to use the fact that R_1 is proportional to the small parameter $m_1 = M_1/M_\odot$ and consider a solution for the orbital elements in terms of power series in m_1. A simple picture is the following. For any element $\mathcal{E}$, the first term $\mathcal{E}_o$ is the current value. The second term $(\delta_1\mathcal{E})$ is found as a first-order solution of the differential equations using constant values of the elements – this is a linear function of time, proportional to m_1. Inserting instead $\mathcal{E}_o + \delta_1\mathcal{E}$ and the other linear expressions, one obtains a second-order solution involving terms proportional to m_1^2, and so on. One may consider cutting at a low order, arguing that the higher-order terms are too small to have a significant influence. This may work well over a limited time, but there is no guarantee of convergence over infinite time by carrying the series to infinite order. In fact it was shown already more than a century ago that such a procedure will in general diverge.

A particular problem is posed by orbital **resonances**. If the so-called **mean motion**, or mean angular velocity $n = 2\pi/P$, of the test object is commensurable with that of the perturbing planet (*i.e.*, n/n_1 is very close to a ratio of small integers, like *e.g.* 2, 3/2, 2/3 or 1/2), some terms in the series can grow arbitrarily large. One may even worry that over long enough time, this may happen for any rational number that n/n_1 happens to be close to, and given that there is always a rational number arbitrarily close to any real number, no orbit would be stable over infinite time.

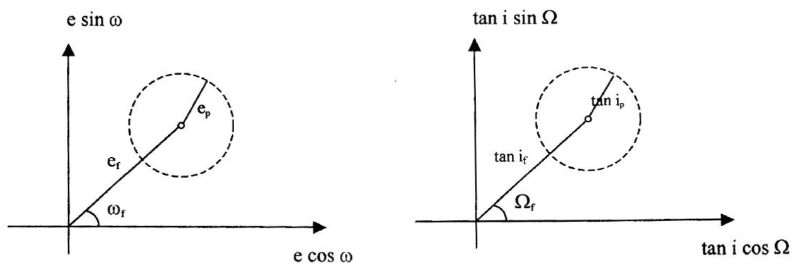
Certainly, this may be a problem of academic interest only, as far as the real Solar System is concerned, and it may be that our planetary orbits remain stable over a time much longer than the age of the Solar System ($4.6 \cdot 10^9$ years). However, the problem of the stability of the Solar System has remained one of the great challenges of dynamics ever since it arose during the 19th century. The study of this problem has developed in close relation to the theory of **chaos** in dynamical systems. This is because the fundamental question about stability is whether the orbital elements of the planets evolve only by periodic oscillations, or if they vary chaotically. Only the former case would ensure stability. Sure enough, each resonance is associated with some chaotic regions of phase space, but in most cases this chaos may take extremely long times to develop. Thus the chaos leading to eventual instability may be too slow to be important in the real Solar System.

We should also note some important theoretical breakthroughs regarding chaos in slightly perturbed, integrable dynamical systems. The *Kolmogorov-Arnold-Moser ("KAM") theorem* from the 1960's states that quasi-periodic motions will dominate the phase space, if the perturbation is small enough. The more recent and practically oriented *Nekhoroshev theorem* assures that even if chaos exists, the deviations from regular motion are bounded during a finite time.

Let us now focus on problems of immediate interest. Limiting the solutions to time spans of $\sim 10^5$ years, the main results are as follows:

- the semimajor axis a remains on the average constant;
- the eccentricity e and inclination i show periodic oscillations;
- these oscillations are coupled to linear variations in ω and Ω, respectively, over 2π.

These (e, ω) and (i, Ω) variations have periods typically $\sim 10^4$ years, and they represent the principal secular variations of the orbits. When applied to asteroids in the main belt, one finds that the solutions may be represented graphically as:

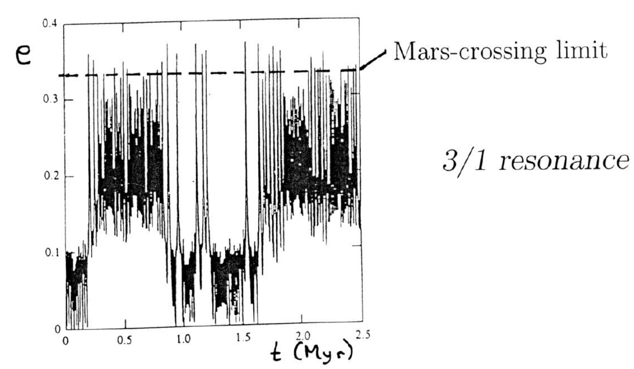


The quantities e_f and i_f are called **forced** eccentricity and inclination, respectively, while e_p and i_p are the **proper** eccentricity and inclination. It is an important aim of asteroid dynamics to reduce the observed, current values of (e, ω) and (i, Ω) to get the **proper elements** (e_p, i_p), which are intrinsic to the objects and thus allow to see genetic relationships like collisional families (where the fragments have closely similar values of e_p and i_p).

The frequencies with which (e, ω) circulates around (e_f, ω_f) and (i, Ω) around (i_f, Ω_f) are of great interest, since they are the frequencies of libration or circulation of (1) the apsidal line (perihelion-aphelion) in the orbital plane, and (2) the line of nodes in the ecliptic plane. There are similar variations of the planetary orbits due to the mutual perturbations between the planets. Thus, for some combinations of a, e_p and i_p it happens that the frequency of secular variation of the asteroid orbit coincides with that of a major perturbing planet (in the first place, Jupiter or Saturn). This is called a **secular resonance**.

Just like the mean motion resonances, the secular ones cause singularities in the series development of the perturbing function, and the motion has to be treated with special care. Very important secular resonances are present in the range of a, e, i where the main asteroid belt is situated, and they lead to large variations in eccentricity in specific regions of the belt, such that some asteroids get into planet-crossing orbits. Such asteroids may get ejected from the Solar System, or they may collide with a planet. The perihelia may also be driven by the secular resonance all the way into the Sun.

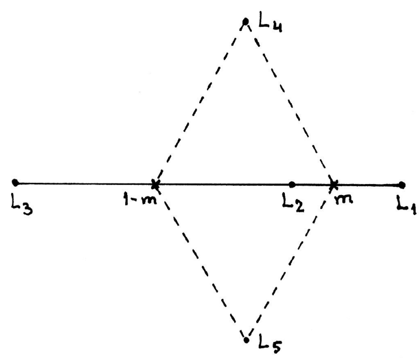
Also the main mean motion resonances like 3/1, 5/2 or 2/1 are affected by the closeness of secular resonances. An overlap of resonances leads to a chaotic zone with large excursions in eccentricity. This offers the best explanation of the Kirkwood gaps in the asteroid belt.



The circular restricted three-body problem

Consider the following problem. A nearly massless body (its mass and gravity are completely negligible) moves in the combined gravitational field of the Sun and one planet, and the orbit of the planet is circular. This is called the *circular restricted three-body problem*. It may seem highly idealized but is actually a good approximation for small bodies like asteroids or comets over short periods of time. There are some useful general results for this problem, which we shall now describe.

First, let us note that we can define a uniformly rotating frame of reference, where the Sun and the planet are both at rest. This is a frame rotating with the constant angular velocity of the planet's orbital motion. It was shown already by Lagrange that there are five points where we can place the third body at rest, and it will remain there. These equilibrium points are called **Lagrange points**, and they are denoted L_1, L_2, L_3, L_4 and L_5. The first three are situated on the line drawn through the Sun and the planet, while L_4 and L_5 occupy the third corners of the two equilateral triangles that have the Sun and the planet in the two remaining corners.

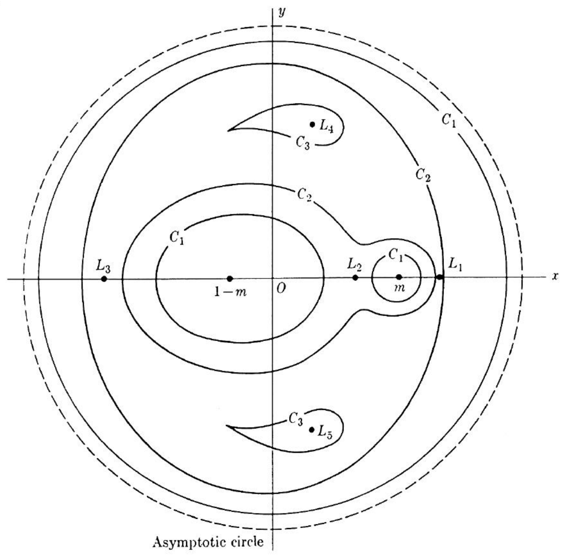


L_1, L_2 and L_3 represent unstable equilibria, but L_4 and L_5 are stable equilibrium points. In the real Solar System there are many objects that move in the vicinity of such points: first and foremost the **Trojan asteroids** with respect to Jupiter and a few other asteroids with respect to Neptune and Mars, but also Saturnian satellites sharing the orbits of Tethys and Dione.

Let us then note that the force field in the rotating system is conservative, since the two masses are at rest, and the centrifugal force also has a fixed potential. This means that the total energy of the massless body (kinetic + potential) remains constant. If we define the unit of mass as $M_\odot + M_p$, *i.e.*, $M_p = m$ and $M_\odot = 1 - m$, the unit of length as the Sun-planet distance, and the unit of time such that the angular velocity of the rotating frame is unity, and if we take the (x, y) plane to be the orbital plane of the planet with the Sun and planet situated on the x axis, we can write the constant quantity as:

$$C = -2E = \left(x^2 + y^2\right) + 2\left\{\frac{1-m}{\rho_1} + \frac{m}{\rho_2}\right\} - v^2 \qquad (11)$$

The first term in the right-hand member represents the centrifugal potential, the second term the gravitational potential, and the third term the kinetic energy per unit mass. Here ρ_1 is the distance to the Sun and ρ_2 is the distance to the planet. If we put $v = 0$, Eq. (11) defines a *surface of zero velocity*, which limits the region of space, where the massless body can move. It is instructive to plot the contours of zero velocity, where these surfaces cut the (x, y) plane for different values of C.



For very large values of C, motion can occur either in a nearly spherical volume around m (assuming that $m << 1$), in a larger, oval region around $1 - m$, or outside a large cylinder enclosing both the Sun and the planet.

If we let C decrease, the two first regions grow, while the third volume shrinks. It turns out that L_1 and L_2 are the points, where the region around the planet opens up, first at L_2 toward the Sun and then at L_1 toward the exterior. Since $m << 1$, L_1 and L_2 are at approximately the same distance from the planet, and the largest region of stable satellite motion around the planet is a sphere extending to that distance. We call this the **Hill sphere**.

The radius of the Hill sphere is:

$$\rho_H \simeq \left(\frac{m}{3}\right)^{1/3} \tag{12}$$

Exercise. Compute approximate radii of the Hill spheres of all planets of the Solar System. Verify that the known natural satellites move within these spheres. Which satellite is closest to the limit?

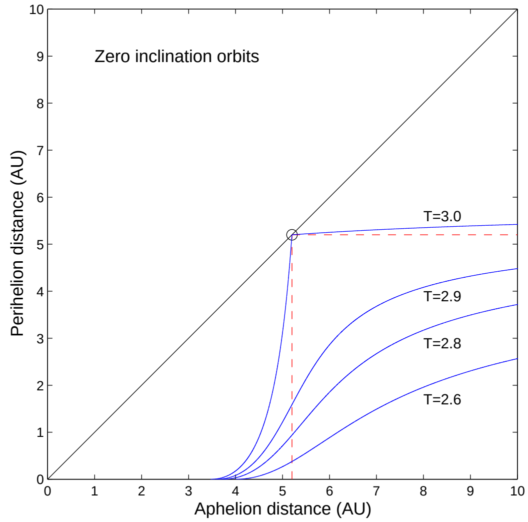
Finally, let us note that Eq. (11), which is called the **Jacobi integral**, offers a way to estimate how the orbit of a comet or asteroid may change in a general situation – even if a close encounter with the planet is involved. For comets, such encounters with Jupiter are relatively frequent, and as a result their orbital elements may change dramatically. However, one may show that, as long as the comet is not very close to either the Sun or Jupiter,

$$C \simeq \frac{1}{a} + 2\sqrt{a(1-e^2)}\cos i \tag{13}$$

Hence, even if a, e and i change considerably, they have to do so in accordance with Eq. (13) for a constant value of C. For practical purposes we use the **Tisserand parameter** T, which is just the same as C:

$$T = \frac{a_J}{a} + 2\sqrt{\frac{a}{a_J}(1-e^2)}\cos i \tag{14}$$

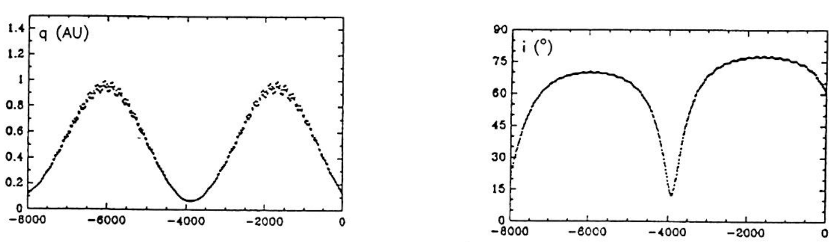
In practice, T is only a quasi-constant quantity, because Jupiter's orbit is not exactly circular and other planets add to the perturbations. But it is generally a good approximation.



If a cometary orbit changes due to close encounters, one can use Eq. (14) to draw evolutionary curves in the (a, e) plane for comets with $\cos i = 1$ – knowing that a coplanar orbit will always remain coplanar ($i \equiv 0$). Using Eqs. (6), one can also plot such curves in the (Q, q) plane as illustrated above. This shows that comets with low-inclination orbits inside that of Jupiter ($Q \lesssim a_J$) may have evolved from other orbits outside the planet ($q \gtrsim a_J$) due to close encounters.

It is interesting to note that the first term of T expresses the orbital energy of the comet with reversed sign, and the second term is the component H_z of angular momentum perpendicular to Jupiter's orbital plane (*i.e.*, approximately the ecliptic plane). In situations where no close encounters occur, a remains nearly constant, and so does H_z according to Eq. (14), but secular perturbations may change the total angular momentum H very much. The eccentricity and inclination may thus oscillate with large amplitudes according to a so-called **Kozai cycle**, so that i is closest to 90° ($|\cos i|$ is a minimum) when e is at its minimum, and conversely the maximum of e occurs when i is closest to 0 or 180°.

If an object (comet or asteroid) has $H_z \simeq 0$, e can get very close to 1, so that the perihelion distance $q = a(1 - e)$ nearly vanishes. This situation is currently observed for **Sun-grazing comets**, some of which are actually seen to hit the Sun. And it has been realized that it is also a common end state for near-Earth asteroids. When these collide with the Sun, the reason is sometimes the above-mentioned secular or mean-motion resonances and sometimes the Kozai cycle.



Kozai cycle of comet 96P/Machholz 1

Close encounters

Let us consider the acceleration of a nearly massless body (we shall call it a comet) that is close to a planet. The equations of motion for the three-body system Sun-planet-comet are:

$$\begin{aligned}
\ddot{\mathbf{r}}_\odot &= \frac{GM_p}{|\mathbf{r}_p - \mathbf{r}_\odot|^3}(\mathbf{r}_p - \mathbf{r}_\odot) && (15)\\
\ddot{\mathbf{r}}_p &= -\frac{GM_\odot}{|\mathbf{r}_p - \mathbf{r}_\odot|^3}(\mathbf{r}_p - \mathbf{r}_\odot) && (16)\\
\ddot{\mathbf{r}}_c &= -\frac{GM_\odot}{|\mathbf{r}_c - \mathbf{r}_\odot|^3}(\mathbf{r}_c - \mathbf{r}_\odot) - \frac{GM_p}{|\mathbf{r}_c - \mathbf{r}_p|^3}(\mathbf{r}_c - \mathbf{r}_p) && (17)
\end{aligned}$$

The acceleration of the comet in the *heliocentric frame* is thus:

$$\ddot{\mathbf{r}}_c - \ddot{\mathbf{r}}_\odot = \underbrace{-\frac{GM_\odot}{|\mathbf{r}_c - \mathbf{r}_\odot|^3}(\mathbf{r}_c - \mathbf{r}_\odot)}_{CENTRAL\Longrightarrow} \underbrace{- \frac{GM_p}{|\mathbf{r}_c - \mathbf{r}_p|^3}(\mathbf{r}_c - \mathbf{r}_p) - \frac{GM_p}{|\mathbf{r}_p - \mathbf{r}_\odot|^3}(\mathbf{r}_p - \mathbf{r}_\odot)}_{PERTURBING\ \Longrightarrow} \qquad (18)$$

Note that this is the same as Eq. (8). Since $|\mathbf{r}_c - \mathbf{r}_p| << |\mathbf{r}_p - \mathbf{r}_\odot|$, the last term can be neglected, and we have approximate values of the magnitudes of the other terms:

$$\textit{central}\ \ \frac{GM_\odot}{a_p^2}; \qquad \textit{perturbing}\ \ \frac{GM_p}{\Delta^2}$$

where Δ is the distance planet-comet, and a_p is the distance Sun-planet. The value of Δ, for which the two accelerations are equal, is:

$$\Delta_h = a_p \cdot \left(\frac{M_p}{M_\odot}\right)^{1/2} \qquad (19)$$

The acceleration of the comet in the *planetocentric frame* is:

$$\ddot{\mathbf{r}}_c - \ddot{\mathbf{r}}_p = \underbrace{-\frac{GM_p}{|\mathbf{r}_c - \mathbf{r}_p|^3}(\mathbf{r}_c - \mathbf{r}_p)}_{CENTRAL\Longrightarrow} \underbrace{- \frac{GM_\odot}{|\mathbf{r}_c - \mathbf{r}_\odot|^3}(\mathbf{r}_c - \mathbf{r}_\odot) + \frac{GM_\odot}{|\mathbf{r}_p - \mathbf{r}_\odot|^3}(\mathbf{r}_p - \mathbf{r}_\odot)}_{PERTURBING\ \Longrightarrow} \qquad (20)$$

Both $|\mathbf{r}_c - \mathbf{r}_\odot|$ and $|\mathbf{r}_p - \mathbf{r}_\odot|$ are $\simeq a_p$, so we get the approximate magnitudes of the accelerations:

$$\textit{central}\ \ \frac{GM_p}{\Delta^2}; \qquad \textit{perturbing}\ \ \frac{GM_\odot \Delta}{a_p^3}$$

The value of Δ, for which the two accelerations are equal, is:

$$\Delta_p = a_p \cdot \left(\frac{M_p}{M_\odot}\right)^{1/3} \qquad (21)$$

Finally, the ratio of central to perturbing acceleration is for the heliocentric frame:

$$\frac{M_\odot}{M_p} \cdot \left(\frac{\Delta}{a_p}\right)^2$$

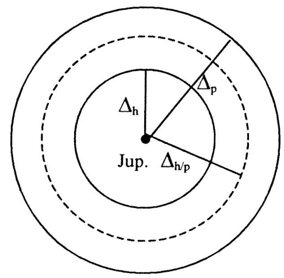
and for the planetocentric frame:

$$\frac{M_p}{M_\odot} \cdot \left(\frac{\Delta}{a_p}\right)^{-3}$$

Equality of the two ratios occurs for:

$$\Delta_{h/p} = a_p \cdot \left(\frac{M_p}{M_\odot}\right)^{2/5} \qquad (22)$$

For instance, for Jupiter the three values are: $\Delta_h = 0.16$ AU; $\Delta_p = 0.52$ AU; and $\Delta_{h/p} = 0.33$ AU.



We see that there is a small region around Jupiter ($\Delta < \Delta_h$), where the heliocentric orbit gets unstable. There is a larger region ($\Delta < \Delta_p$), where the jovicentric orbit is reasonably stable. For $\Delta_h < \Delta < \Delta_p$, both orbits may be called stable, though only marginally so.

Δ_p represents a stability criterion for planetocentric motion that differs somewhat from the earlier result for the Hill sphere. The limiting distance is a factor $3^{1/3} \simeq 1.44$ smaller for the Hill sphere, which is the more stringent of the two criteria.

An approximate way to treat close encounters is to follow the unperturbed heliocentric orbit until $\Delta = \Delta_{h/p}$, and then shift to a hyperbolic planetocentric orbit that is followed until $\Delta = \Delta_{h/p}$ again, and a new heliocentric orbit is computed. This means that the close encounters can be treated with "hyperbolic deflections" like a scattering problem, analogous to nuclear particle scattering in accelerators.

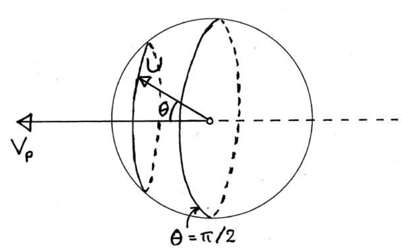
One useful result is that the planetocentric velocity has the same magnitude before and after the scattering – only the direction is changed. The event can also be considered as instantaneous, so the position of the planet is unchanged.

In fact, the approach velocity to the planet (U) can be computed very easily, since it can be shown to be related to the Tisserand parameter by:

$$U^2 = 3 - T \tag{23}$$

and naturally, since T is conserved, U is also conserved.

Exercise for the advanced student. Prove Eq. (23)



There is a simple relation between the direction of planetocentric motion and the orbital elements. Let θ be the angle between this direction and the planet's direction of heliocentric motion. Then the transverse component of the comet's heliocentric velocity in the planet's orbital plane can be written as $V_p(1+U\cos\theta)$. Hence, when θ changes during a close encounter, so does the perpendicular component of the orbital angular momentum in a corresponding way. And, due to the conservation of T, the orbital energy also changes correspondingly. In fact, for a given U, the value of θ is uniquely related to E and H_z.

We may investigate what conditions a cometary orbit must fulfill, if Jupiter shall be able to eject the comet from the Solar System by close encounters. In the units previously adopted, Jupiter's circular orbital velocity is $V_J = 1$. The velocity of escape from the Solar System is thus $V_E = \sqrt{2}$ (prove this!). The relative velocity between the comet and Jupiter must hence obey: $U > \sqrt{2} - 1$, and we get:

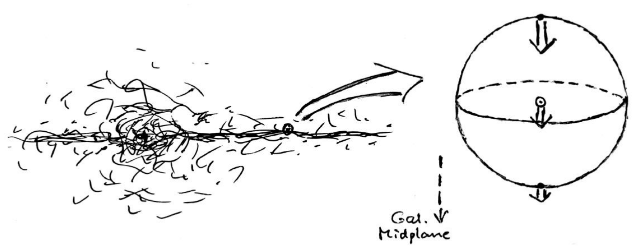
$$3 - T > (\sqrt{2} - 1)^2 \Longrightarrow T < 2\sqrt{2} \tag{24}$$

Dynamics of the Oort Cloud

The outskirts of the Solar System are occupied by a huge number of comets moving in bound orbits around the Sun. If external perturbations did not exist, the gravitational reach of the Sun would be infinite, and comets might move around the Sun at whatever distances. But, clearly, this is not the case, and we have to estimate the size of the Sun's sphere of influence in the Galaxy.

For this purpose let us consider the Hill sphere as in Eq. (12) under the assumption that the Sun moves in a circular orbit around the center of the Galaxy, where we place a point mass equal to the total mass of the Galaxy. Hence we use the approximation of the circular restricted three-body problem. The mass of the Galaxy can be estimated at about $2 \cdot 10^{11}$ solar masses, and the distance to the Galactic center is about 8.6 kpc. Eq. (12) thus yields: $\rho_H \simeq 1$ pc $\simeq 2 \cdot 10^5$ AU. We may take this to be a rough limiting distance for cometary orbits to be stable.

Determination of long-period comet orbits and integration of the planetary perturbations backward in time indicate that, more or less, the whole volume at $r < \rho_H$ is populated by comets. We call this the **Oort Cloud**. Even the observed comets, which have perihelia in the innermost part of the Solar System, spend most of the time at $r \gtrsim 10^4$ AU. Which perturbations may act in comets moving at such distances? The answer is twofold.



Compressional tidal field

1. The Solar System is situated in the smooth, general gravitational field of the Galaxy as a whole. We already saw that this limits the distances of stable motion around the Sun, but the orbits may be strongly perturbed even inside this limit. Since our place in the Galaxy is inside the disk relatively far from the center, it turns out that the gravity of the disk causes the dominant acceleration. The corresponding perturbing force on a comet in the Oort Cloud is the difference in the accelerations of the comet and the Sun due to their slight difference in distance from the central plane of the disk.

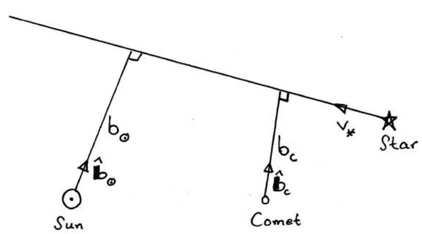
   The effect of this perturbation is similar to the above-described Kozai cycle, because there is no change in orbital energy – only in angular momentum. But there is no torque component perpendicular to the disk since the perturbing force is "vertical", so H_z is conserved (with the z direction toward the Galactic pole). Some orbits may thus undergo very large changes in perihelion distance, especially if H_z is very small. The periods are very long – typically $\sim 10^8$ yr or more – and during most of the time the perihelion is safely outside the planetary system. But there comes a time, when the comet penetrates among the planets and may come close enough to the Sun to be observed. When this happens, the orbit becomes violently unstable due to planetary perturbations, and the comet may be ejected from the Solar System or captured into orbits of shorter periods.

2. Individual stars or other local mass concentrations in the Galaxy add their own, local gravitational fields in their immediate surroundings. Therefore, whenever the Solar System passes at high speed near such local perturbers, both the Sun and each comet in the Oort Cloud receive impulses that are somewhat different due to the difference in encounter geometry. The differential impulse perturbs the heliocentric orbit of the comet.

   Such impulses are best described by means of hyperbolic deflections like those undergone by comets passing close to Jupiter. But often one may use a simplified approach, assuming that the star passes along a straight line with constant speed, while the comet is at rest relative to the Sun. This is motivated by the fact that typical encounter speeds of other stars are ~ 30 km/s, while the comet moves around the Sun with a speed $\lesssim 0.1$ km/s and the speed of escape at the star's closest approach – typically at $\sim 10^5$ AU – is only ~ 0.1 km/s as well. One generally refers to the simplification as the *classical impulse approximation*, and the formula for the relative impulse is:

$$\Delta \mathbf{v} = \frac{2GM_*}{v_*} \left\{ \frac{\hat{\mathbf{b}}_c}{b_c} - \frac{\hat{\mathbf{b}}_\odot}{b_\odot} \right\} \tag{25}$$

   where M_* and v_* are the mass and velocity of the star, $\hat{\mathbf{b}}$ is the unit vector toward the point of closest encounter, and b_c and $b_\odot$ are the respective encounter distances.

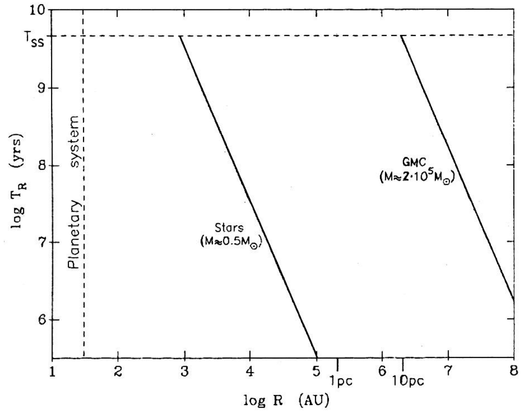


Exercise. Estimate the order of magnitude of the impulse received by a comet, if the star has the same mass as the Sun and passes at $\sim 10^4$ AU from the comet but much further away from the Sun.

Encounter frequencies in the Galaxy are estimated by the simple formula:

$$T_*(D) = \left\{ \pi D^2 n_* v_* \right\}^{-1} \tag{26}$$

where T_* is the average time between successive encounters within a distance D, n_* is the number density of stars, and v_* is the typical encounter speed with stars. We may verify that stars actually pass through the Oort Cloud on a time scale equal to the orbital periods of Oort Cloud comets – the latter being a few million years in typical cases.



Exercise. Estimate the time it typically takes before a comet in the Oort Cloud is encountered by a star to within $\sim 10^4$ AU, using the above formula and an estimated density of $n_* \sim 0.1$ pc^{-3}.

We conclude that there is a practically continuous reshuffling of the orbits of Oort Cloud comets due to slight impulses received from stellar encounters. This is enough to thermalize the cloud and make it isotropic during the age of the Solar System, and to guarantee that there will always be comets with very small values of the angular momentum, which fall into perihelia in the inner Solar System.

Actually, the above-mentioned effect of the Galactic tide is in general larger, when it comes to providing new, observable comets from the cloud. But there is an important aspect in favour of the stars. In case there had only been the Galactic tide, all the comets that started out with very small values of H_z would have been lost long ago by the planetary perturbations, as explained above, and we would not see any new comets from the Oort Cloud at the present time. This awkward situation is saved by the stars, which reshuffle angular momenta all the time, thereby providing new comets with very small H_z that can be grabbed by the Galactic tide and transferred into the inner Solar System.

In addition to the above-mentioned effects, one expects that the Solar System passes near very large mass concentrations in the Galaxy from time to time, and these may cause large disturbances of the whole Oort Cloud. One example, plotted in the figure above, is Giant Molecular Clouds (GMCs) with masses $\sim 10^6$ solar masses.

Nongravitational forces

A common feature of the dynamics of all the small bodies in the Solar System is that they are influenced by forces that do not arise from gravitational interactions. These are called nongravitational forces. Table 4 summarizes the main effects acting on each class of small bodies.

Radiation pressure

Consider a dust grain moving in a heliocentric orbit. This grain is affected by two forces with solar origin, *i.e.*, solar gravity and solar radiation pressure. If the grain is at a distance r from the Sun, the force of gravity is:

$$\mathbf{F}_g = -\frac{GM_\odot m}{r^2}\,\widehat{\mathbf{r}}, \tag{27}$$

Table 4:

Object class	Nongravitational effect
Dust particles	Radiation pressure
Larger meteoroids	Poynting-Robertson drag
Comets	Outgassing jet effect
Asteroids	Yarkovsky effect

obviously directed toward the Sun. The mass of the grain is m, and this can be expressed in terms of the density ρ and radius a. The radiation pressure force can be expressed in terms of the momentum flux of solar radiation at distance r:

$$P_\odot = \mathcal{F}_\odot / c r_{\mathrm{AU}}^2, \tag{28}$$

where $\mathcal{F}_\odot$ is the "solar constant", *i.e.*, the bolometric energy flux of sunlight at Earth's orbit, and c is the speed of light. The effective cross-section of the grain for interacting with solar photons can be expressed as:

$$A = \pi a^2 Q_{pr}, \tag{29}$$

where Q_{pr} is an efficiency factor that is ~ 2 for wavelengths $\lambda > a$ and drops toward zero for $\lambda << a$. The radiation pressure force is simply

$$\mathbf{F}_r = P_\odot \cdot A\,\hat{\mathbf{r}}. \tag{30}$$

Exercise. Verify that the grain radius for which the forces of gravity and radiation pressure are equal (at any distance from the Sun!) is of order 1 μm, using above numbers for G, $M_\odot$ and the length of the AU, and the value $\mathcal{F}_\odot = 1.36 \cdot 10^3\ \mathrm{W\,m^{-2}}$.

Since the radiation pressure force is directed outward from the Sun, the resulting force acting on the grain is $\mathbf{F}_e = -(F_g - F_r)\,\hat{\mathbf{r}}$. Since both terms F_g and F_r scale as r^{-2}, the effective force is like a scaled-down gravity. One usually expresses this by a parameter $\beta = F_r/F_g$, such that $F_e = F_g \cdot (1-\beta)$. Grains with different radii and densities will have different values of β.

Poynting-Robertson drag

This effect can be illustrated as follows. A grain orbiting the Sun will absorb photons from the solar radiation field, and it will emit its own photons due to thermal radiation. If the heat balance of the grain involves only those two processes, then the amounts of energy absorbed and emitted in unit time must be equal. However, the grain will lose angular momentum from its orbital motion!

Imagine that the transverse velocity of the grain (component perpendicular to the solar direction) is v_t. This is very small compared to the speed of light, but it does cause a slight aberration effect. As viewed in the grain's frame of reference, the sunlight appears to come from a direction slightly ahead of the Sun, the aberration angle being

$$\Psi = \arcsin(v_t/c) \tag{31}$$

The momentum gained by absorption of those photons thus deviates by the angle Ψ from the radial direction, tending to decrease v_t. On the other hand, the photons emitted by the grain have no preference for the forward or backward direction and thus do not affect v_t. The result is like that of a drag force, and two phenomena will occur.

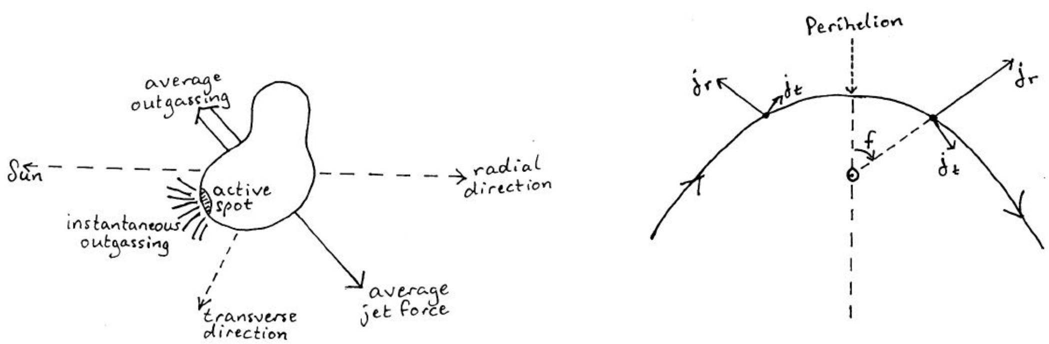
First, the orbit of the grain gets circularized, since the drag is strongest near perihelion. Second, the circular orbit of the grain will slowly "spiral" inward, toward the Sun. Typical time scales of this spiraling motion, and thus lifetimes of the grains, are $\sim 10^4$ years for the grain sizes that dominate in the zodiacal light. They increase with increasing radius, and the Poynting-Robertson effect is not important for macroscopic objects orbiting the Sun.

Cometary outgassing jet effect

The characteristic property of a cometary nucleus, which distinguishes it from an asteroid, is that it produces gas upon arrival close enough to the Sun. This gas leaves the nucleus with a finite velocity, which is related to the temperature of the surface layer where the molecules originate. If the outgassing is not uniform (equal in all directions), the nucleus is subjected to a jet force corresponding to the average speed of the bulk outflow. If the mass of the nucleus is M, the mass of a molecule is m, the average outflow velocity is $\mathbf{v}$, and the number of molecules produced per second is Q, the jet acceleration will be

$$\mathbf{j} = -\frac{Qm}{M}\mathbf{v} \tag{32}$$

Usually the dominant molecular species is water, and one may use observational estimates of Q at different positions in the orbit, and a physical model of the outgassing to estimate $\mathbf{v}$, in order to obtain $M\mathbf{j}$ as a function of orbital position. The practical use of this lies in estimating the mass of the nucleus, using the observed effects of $\mathbf{j}$ on the orbital elements of the comet.



There are several such *nongravitational effects*, and the one that is easiest to measure is a perturbation of the orbital period, ΔP. For a periodic comet an accurate determination of P is possible only by observing two consecutive apparitions of the comet. By observing a third apparition, one may then check if P is only perturbed by the planetary gravities, or if there is a measurable nongravitational ΔP effect. Such effects are reliably measured for a majority of the periodic comets, and for those with relatively long periods, the effect in question can be quite large. For instance, comet 1P/Halley arrives at each perihelion about four days too late with respect to the best gravitational prediction.

A second effect that also stands out, although the related displacement of the comet on the sky is always much smaller, is a perturbation of the longitude of perihelion, $\Delta\varpi$. This quantity can also be deduced from linkages of several apparitions of the same comet, and it typically amounts to only an arcsecond or less per orbit.

When linking cometary apparitions, one may use different models for $\mathbf{j}$. The one most commonly used since more than 30 years is based on assuming a functional dependence

$$\mathbf{j} = (A_1, A_2, A_3) \cdot g(r), \tag{33}$$

expressing $\mathbf{j}$ in a reference frame that follows the comet with the x-axis pointing radially outward from the Sun, the y-axis in the transverse direction in the orbital plane (positive along the direction of motion), and the z-axis perpendicular to this plane. Very often the third component (A_3 parameter) is neglected. The other two parameters, A_1 and A_2, can be shown to relate to $\Delta\varpi$ and ΔP, respectively.

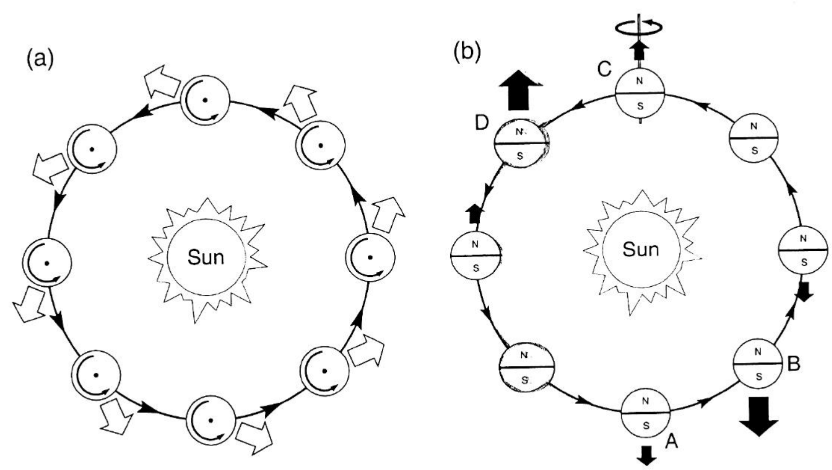
The function $g(r)$ has been chosen to approximate the variation of the sublimation flux of water ice, as the comet moves along its orbit. This falls off with r nearly as r^{-2} while close to the Sun, and much more rapidly at distances beyond $r \sim 2.8$ AU:

$$g(r) = 0.111262 \left(\frac{r}{r_o}\right)^{-2.15} \left[1 + \left(\frac{r}{r_o}\right)^{5.093}\right]^{-4.6142} \tag{34}$$

with $r_o = 2.808$ AU.

Yarkovsky effect

When we discussed the Poynting-Robertson drag on grains, we dismissed the possibility of a forward/backward asymmetry of the thermal radiation of the particle, because for such small objects there cannot be any major temperature contrasts between the day and night, or forward and backward sides. However, for large objects – several meters across or even larger – important temperature contrasts may develop, and thus the thermal radiation of those objects may in fact have an asymmetry along the transverse direction. Hence, the orbital motion may gain or lose angular momentum, and the object may slowly spiral outward or inward.



There are two different effects of this kind: the *diurnal* (left panel) and the *seasonal Yarkovsky effect* (right panel). The diurnal effect is largest, when the spin axis is perpendicular to the orbital plane. The thermal inertia induces a lag of maximum surface temperature such that most thermal photons are emitted toward the afternoon direction, causing a jet force in the opposite direction. Depending on whether the spin is prograde or retrograde, this jet force will accelerate or decelerate the orbital motion, and the asteroid will drift outward or inward.

The seasonal effect is largest, when the spin axis is in the orbital plane. Once again due to thermal inertia, the maximum temperature on the poles occurs somewhat after the orbital position, when the pole faces the Sun. The jet force that is directed toward the dark pole thus has a transverse component that decelerates the orbital motion. The asteroid will hence drift toward the Sun, as if it were affected by a drag force.

The Yarkovsky effects on sizeable asteroids are extremely small, but when they act over Gyr time scales, the orbits may change appreciably. This is identified as a main contributor to the delivery of new asteroids into the resonant zones, where they may escape into planet-crossing orbits (see above).

In addition, very accurate data on distance and velocity have been obtained by radar observations during several occasions for near-Earth asteroid (6489) *Golevka*. These have shown that a nongravitational effect (interpreted as the Yarkovsky effect) caused a relative displacement of this ~ 0.5 km diameter asteroid by 15 km from the first observations in 1991 until the most recent in 2003. In cases where such small asteroids are predicted to pass very close to the Earth and get deflected by Earth's gravity into orbits that may lead to impact on our planet (this does happen in reality!), a prediction of the Yarkovsky effect would be essential in order to know if the deflection will be dangerous or not.