3

The Orbits of Stars

In this chapter we examine the orbits of individual stars in gravitational fields such as those found in stellar systems. Thus we ask the questions, "What kinds of orbits are possible in a spherically symmetric, or an axially symmetric potential? How are these orbits modified if we distort the potential into a bar-like form?" We shall obtain analytic results for the simpler potentials, and use these results to develop an intuitive understanding of how stars move in more general potentials.

In §§3.1 to 3.3 we examine orbits of growing complexity in force fields of decreasing symmetry. The less symmetrical a potential is the less likely it is that we can obtain analytic results, so in §3.4 we review techniques for integrating orbits in both a given gravitational field, and the gravitational field of a system of orbiting masses. Even numerically integrated orbits in gravitational fields of low symmetry often display a high degree of regularity in their phase-space structures. In §3.5 we study this structure using analytic models, and develop analytic tools of considerable power, including the idea of adiabatic invariance, which we apply to some astronomical problems in §3.6. In §3.7 we develop Hamiltonian perturbation theory, and use it to study the phenomenon of orbital resonance and the role it plays in generating orbital chaos. In §3.8 we draw on techniques developed throughout the chapter to understand how elliptical galaxies are affected by the existence of central stellar cusps and massive black holes at their centers.

All of the work in this chapter is based on a fundamental approximation:

although galaxies are composed of stars, we shall neglect the forces from individual stars and consider only the large-scale forces from the overall mass distribution, which is made up of thousands of millions of stars. In other words, we assume that the gravitational fields of galaxies are *smooth*, neglecting small-scale irregularities due to individual stars or larger objects like globular clusters or molecular clouds. As we saw in §1.2, the gravitational fields of galaxies *are* sufficiently smooth that these irregularities can affect the orbits of stars only after many crossing times.

Since we are dealing only with gravitational forces, the trajectory of a star in a given field does not depend on its mass. Hence, we examine the dynamics of a particle of unit mass, and quantities such as momentum, angular momentum, and energy, and functions such as the Lagrangian and Hamiltonian, are normally written per unit mass.

3.1 Orbits in static spherical potentials

We first consider orbits in a static, spherically symmetric gravitational field. Such fields are appropriate for globular clusters, which are usually nearly spherical, but, more important, the results we obtain provide an indispensable guide to the behavior of orbits in more general fields.

The motion of a star in a centrally directed gravitational field is greatly simplified by the familiar law of conservation of angular momentum (see Appendix D.1). Thus if

$$\mathbf{r} = r\hat{\mathbf{e}}_r \tag{3.1}$$

denotes the position vector of the star with respect to the center, and the radial acceleration is

$$\mathbf{g} = g(r)\hat{\mathbf{e}}_r, \tag{3.2}$$

the equation of motion of the star is

$$\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = g(r)\hat{\mathbf{e}}_r. \tag{3.3}$$

If we remember that the cross product of any vector with itself is zero, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{r}\times\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right) = \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\times\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} + \mathbf{r}\times\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = g(r)\mathbf{r}\times\hat{\mathbf{e}}_r = 0. \tag{3.4}$$

Equation (3.4) says that $\mathbf{r}\times\dot{\mathbf{r}}$ is some constant vector, say $\mathbf{L}$:

$$\mathbf{r}\times\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} = \mathbf{L}. \tag{3.5}$$

Of course, $\mathbf{L}$ is simply the angular momentum per unit mass, a vector perpendicular to the plane defined by the star's instantaneous position and velocity vectors. Since this vector is constant, we conclude that the star moves in a plane, the **orbital plane**. This finding greatly simplifies the determination of the star's orbit, for now that we have established that the star moves in a plane, we may simply use plane polar coordinates (r, ψ) in which the center of attraction is at $r = 0$ and ψ is the azimuthal angle in the orbital plane. In terms of these coordinates, the Lagrangian per unit mass (Appendix D.3) is

$$\mathcal{L} = \tfrac{1}{2}\big[\dot{r}^2 + (r\dot{\psi})^2\big] - \Phi(r), \tag{3.6}$$

where Φ is the gravitational potential and $g(r) = -\mathrm{d}\Phi/\mathrm{d}r$. The equations of motion are

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\mathcal{L}}{\partial\dot{r}} - \frac{\partial\mathcal{L}}{\partial r} = \ddot{r} - r\dot{\psi}^2 + \frac{\mathrm{d}\Phi}{\mathrm{d}r}, \tag{3.7a}$$

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\mathcal{L}}{\partial\dot{\psi}} - \frac{\partial\mathcal{L}}{\partial\psi} = \frac{\mathrm{d}}{\mathrm{d}t}\big(r^2\dot{\psi}\big). \tag{3.7b}$$

The second of these equations implies that

$$r^2\dot{\psi} = \text{constant} \equiv L. \tag{3.8}$$

It is not hard to show that L is actually the length of the vector $\mathbf{r} \times \dot{\mathbf{r}}$, and hence that (3.8) is just a restatement of the conservation of angular momentum. Geometrically, L is equal to twice the rate at which the radius vector sweeps out area.

To proceed further we use equation (3.8) to replace time t by angle ψ as the independent variable in equation (3.7a). Since (3.8) implies

$$\frac{\mathrm{d}}{\mathrm{d}t} = \frac{L}{r^2}\frac{\mathrm{d}}{\mathrm{d}\psi}, \tag{3.9}$$

equation (3.7a) becomes

$$\frac{L^2}{r^2}\frac{\mathrm{d}}{\mathrm{d}\psi}\left(\frac{1}{r^2}\frac{\mathrm{d}r}{\mathrm{d}\psi}\right) - \frac{L^2}{r^3} = -\frac{\mathrm{d}\Phi}{\mathrm{d}r}. \tag{3.10}$$

This equation can be simplified by the substitution

$$u \equiv \frac{1}{r}, \tag{3.11a}$$

which puts (3.10) into the form

$$\frac{\mathrm{d}^2u}{\mathrm{d}\psi^2} + u = \frac{1}{L^2u^2}\frac{\mathrm{d}\Phi}{\mathrm{d}r}\big(1/u\big). \tag{3.11b}$$

The solutions of this equation are of two types: along **unbound** orbits $r \rightarrow \infty$ and hence $u \rightarrow 0$, while on **bound** orbits r and u oscillate between finite limits. Thus each bound orbit is associated with a periodic solution of this equation. We give several analytic examples later in this section, but in general the solutions of equation (3.11b) must be obtained numerically.

Some additional insight is gained by deriving a "radial energy" equation from equation (3.11b) in much the same way as we derive the conservation of kinetic plus potential energy in Appendix D; we multiply (3.11b) by $\mathrm{d}u/\mathrm{d}\psi$ and integrate over ψ to obtain

$$\left(\frac{\mathrm{d}u}{\mathrm{d}\psi}\right)^2 + \frac{2\Phi}{L^2} + u^2 = \text{constant} \equiv \frac{2E}{L^2}, \tag{3.12}$$

where we have used the relation $\mathrm{d}\Phi/\mathrm{d}r = -u^2(\mathrm{d}\Phi/\mathrm{d}u)$.

This result can also be derived using Hamiltonians (Appendix D.4). From (3.6) we have that the momenta are $p_r = \partial\mathcal{L}/\partial\dot{r} = \dot{r}$ and $p_\psi = \partial\mathcal{L}/\partial\dot{\psi} = r^2\dot{\psi}$, so with equation (D.50) we find that the Hamiltonian per unit mass is

$$\begin{aligned} H(r, p_r, p_\psi) &= p_r\dot{r} + p_\psi\dot{\psi} - \mathcal{L} \\ &= \tfrac{1}{2}\Big(p_r^2 + \frac{p_\psi^2}{r^2}\Big) + \Phi(r) \\ &= \tfrac{1}{2}\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 + \tfrac{1}{2}\left(r\frac{\mathrm{d}\psi}{\mathrm{d}t}\right)^2 + \Phi(r). \end{aligned} \tag{3.13}$$

When we multiply (3.12) by $L^2/2$ and exploit (3.9), we find that the constant E in equation (3.12) is simply the numerical value of the Hamiltonian, which we refer to as the energy of that orbit.

For bound orbits the equation $\mathrm{d}u/\mathrm{d}\psi = 0$ or, from equation (3.12)

$$u^2 + \frac{2[\Phi(1/u) - E]}{L^2} = 0 \tag{3.14}$$

will normally have two roots u_1 and u_2 between which the star oscillates radially as it revolves in ψ (see Problem 3.7). Thus the orbit is confined between an inner radius $r_1 = u_1^{-1}$, known as the **pericenter** distance, and an outer radius $r_2 = u_2^{-1}$, called the **apocenter** distance. The pericenter and apocenter are equal for a circular orbit. When the apocenter is nearly equal to the pericenter, we say that the orbit has small **eccentricity**, while if the apocenter is much larger than the pericenter, the eccentricity is said to be near unity. The term "eccentricity" also has a mathematical definition, but only for Kepler orbits—see equation (3.25a).

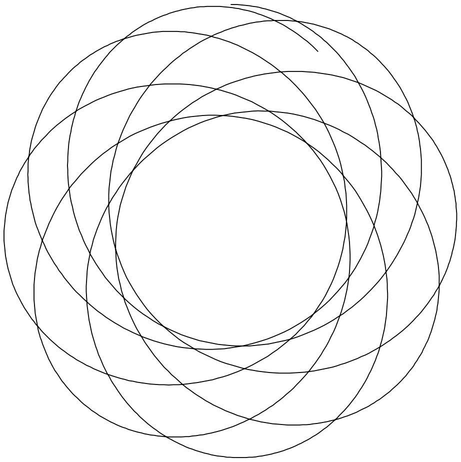


Figure 3.1 A typical orbit in a spherical potential (the isochrone, eq. 2.47) forms a rosette.

The **radial period** T_r is the time required for the star to travel from apocenter to pericenter and back. To determine T_r we use equation (3.8) to eliminate $\dot{\psi}$ from equation (3.13). We find

$$\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 = 2(E - \Phi) - \frac{L^2}{r^2}, \tag{3.15}$$

which may be rewritten

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \pm\sqrt{2[E - \Phi(r)] - \frac{L^2}{r^2}}. \tag{3.16}$$

The two possible signs arise because the star moves alternately in and out. Comparing (3.16) with (3.14) we see that $\dot{r} = 0$ at the pericenter and apocenter distances r_1 and r_2, as of course it must. From equation (3.16) it follows that the radial period is

$$T_r = 2\int_{r_1}^{r_2} \frac{\mathrm{d}r}{\sqrt{2[E - \Phi(r)] - L^2/r^2}}. \tag{3.17}$$

In traveling from pericenter to apocenter and back, the azimuthal angle ψ increases by an amount

$$\Delta\psi = 2\int_{r_1}^{r_2} \frac{\mathrm{d}\psi}{\mathrm{d}r}\mathrm{d}r = 2\int_{r_1}^{r_2} \frac{L}{r^2}\frac{\mathrm{d}t}{\mathrm{d}r}\mathrm{d}r. \tag{3.18a}$$

Substituting for $\mathrm{d}t/\mathrm{d}r$ from (3.16) this becomes

$$\Delta\psi = 2L\int_{r_1}^{r_2} \frac{\mathrm{d}r}{r^2\sqrt{2[E - \Phi(r)] - L^2/r^2}}. \tag{3.18b}$$

The **azimuthal period** is

$$T_\psi = \frac{2\pi}{|\Delta\psi|} T_r; \tag{3.19}$$

in other words, the mean angular speed of the particle is $2\pi/T_\psi$. In general $\Delta\psi/2\pi$ will not be a rational number. Hence the orbit will not be closed: a typical orbit resembles a rosette and eventually passes close to every point in the annulus between the circles of radii r_1 and r_2 (see Figure 3.1 and Problem 3.13). There are, however, two and only two potentials in which all bound orbits are closed.

(a) Spherical harmonic oscillator We call a potential of the form

$$\Phi(r) = \tfrac{1}{2}\Omega^2 r^2 + \text{constant} \tag{3.20}$$

a spherical harmonic oscillator potential. As we saw in §2.2.2b, this potential is generated by a homogeneous sphere of matter. Equation (3.11b) could be solved analytically in this case, but it is simpler to use Cartesian coordinates (x, y) defined by $x = r\cos\psi$, $y = r\sin\psi$. In these coordinates, the equations of motion are simply

$$\ddot{x} = -\Omega^2 x \quad ; \quad \ddot{y} = -\Omega^2 y, \tag{3.21a}$$

with solutions

$$x = X\cos(\Omega t + \epsilon_x) \quad ; \quad y = Y\cos(\Omega t + \epsilon_y), \tag{3.21b}$$

where X, Y, ϵ_x, and ϵ_y are arbitrary constants. Every orbit is closed since the periods of the oscillations in x and y are identical. The orbits form ellipses centered on the center of attraction. The azimuthal period is $T_\psi = 2\pi/\Omega$ because this is the time required for the star to return to its original azimuth. During this time, the particle completes two in-and-out cycles, so the radial period is

$$T_r = \tfrac{1}{2}T_\psi = \frac{\pi}{\Omega}. \tag{3.22}$$

(b) Kepler potential When the star is acted on by an inverse-square field $g(r) = -GM/r^2$ due to a point mass M, the corresponding potential is $\Phi = -GM/r = -GMu$. Motion in this potential is often called **Kepler motion**. Equation (3.11b) becomes

$$\frac{\mathrm{d}^2 u}{\mathrm{d}\psi^2} + u = \frac{GM}{L^2}, \tag{3.23}$$

the general solution of which is

$$u(\psi) = C\cos(\psi - \psi_0) + \frac{GM}{L^2}, \tag{3.24}$$

where $C > 0$ and ψ_0 are arbitrary constants. Defining the orbit's **eccentricity** by

$$e \equiv \frac{CL^2}{GM} \tag{3.25a}$$

and its **semi-major axis** by

$$a \equiv \frac{L^2}{GM(1-e^2)}, \tag{3.25b}$$

equation (3.24) may be rewritten

$$r(\psi) = \frac{a(1-e^2)}{1+e\cos(\psi-\psi_0)}. \tag{3.26}$$

An orbit for which $e \geq 1$ is unbound, since $r \to \infty$ as $(\psi - \psi_0) \to \pm\cos^{-1}(-1/e)$. We discuss unbound orbits in §3.1d below. Bound orbits have $e < 1$ and along them r is a periodic function of ψ with period 2π, so the star returns to its original radial coordinate after exactly one revolution in ψ. Thus bound Kepler orbits are closed, and one may show that they form ellipses with the attracting center at one focus. The pericenter and apocenter distances are

$$r_1 = a(1-e) \quad \text{and} \quad r_2 = a(1+e). \tag{3.27}$$

In many applications, equation (3.26) for r along a bound Kepler orbit is less convenient than the parameterization

$$r = a(1 - e\cos\eta), \tag{3.28a}$$

where the parameter η is called the **eccentric anomaly** to distinguish it from the **true anomaly**, $\psi - \psi_0$. By equating the right sides of equations (3.26) and (3.28a) and using the identity $\cos\theta = (1-\tan^2\frac{1}{2}\theta)/(1+\tan^2\frac{1}{2}\theta)$, it is straightforward to show that the true and eccentric anomalies are related by

$$\sqrt{1-e}\,\tan\tfrac{1}{2}(\psi-\psi_0) = \sqrt{1+e}\,\tan\tfrac{1}{2}\eta. \tag{3.29}$$

Equation (3.326) gives alternative relations between the two anomalies.

Taking $t = 0$ to occur at pericenter passage, from $L = r^2\dot\psi$ we have

$$t = \int_{\psi_0}^{\psi} \frac{\mathrm{d}\psi}{\dot\psi} = \int \mathrm{d}\psi\,\frac{r^2}{L} = \frac{a^2}{L}\int_0^{\eta} \mathrm{d}\eta\,\frac{\mathrm{d}\psi}{\mathrm{d}\eta}(1-e\cos\eta)^2. \tag{3.30}$$

Evaluating $\mathrm{d}\psi/\mathrm{d}\eta$ from (3.29), integrating, and using trigonometrical identities to simplify the result, we obtain finally

$$t = \frac{a^2}{L}\sqrt{1-e^2}\,(\eta - e\sin\eta) = \frac{T_r}{2\pi}(\eta - e\sin\eta), \tag{3.28b}$$

where the second equality follows because the bracket on the right increases by 2π over an orbital period. This is called **Kepler's equation**, and the quantity $2\pi t/T_r$ is sometimes called the **mean anomaly**. Hence

$$T_r = T_\psi = \frac{a^2}{L}\sqrt{1-e^2} = 2\pi\sqrt{\frac{a^3}{GM}}, \tag{3.31}$$

where the second equality uses (3.25b).

From (3.12) the energy per unit mass of a particle on a Kepler orbit is

$$E = -\frac{GM}{2a}. \tag{3.32}$$

To unbind the particle, we must add the **binding energy** $-E$.

The study of motion in nearly Kepler potentials is central to the dynamics of planetary systems (Murray & Dermott 1999).

We have shown that a star on a Kepler orbit completes a radial oscillation in the time required for ψ to increase by $\Delta\psi = 2\pi$, whereas a star that orbits in a harmonic-oscillator potential has already completed a radial oscillation by the time ψ has increased by $\Delta\psi = \pi$. Since galaxies are more extended than point masses, and less extended than homogeneous spheres, a typical star in a spherical galaxy completes a radial oscillation after its angular coordinate has increased by an amount that lies somewhere in between these two extremes; $\pi < \Delta\psi < 2\pi$ (cf. Problem 3.17). Thus, we expect a star to oscillate from its apocenter through its pericenter and back in a shorter time than is required for one complete azimuthal cycle about the galactic center.

It is sometimes useful to consider that an orbit in a non-Kepler force field forms an approximate ellipse, though one that **precesses** by $\psi_\mathrm{p} = \Delta\psi - 2\pi$ in the time needed for one radial oscillation. For the orbit shown in Figure 3.1, and most galactic orbits, this precession is in the sense opposite to the rotation of the star itself. The angular velocity Ω_p of the rotating frame in which the ellipse appears closed is

$$\Omega_\mathrm{p} = \frac{\psi_\mathrm{p}}{T_r} = \frac{\Delta\psi - 2\pi}{T_r}. \tag{3.33}$$

Hence we say that Ω_p is the **precession rate** of the ellipse. The concept of closed orbits in a rotating frame of reference is crucial to the theory of spiral structure—see §6.2.1, particularly Figure 6.12.

(c) Isochrone potential The harmonic oscillator and Kepler potentials are both generated by mass distributions that are qualitatively different from the mass distributions of galaxies. The only known potential that could be generated by a realistic stellar system for which all orbits are analytic is the isochrone potential of equation (2.47) (Hénon 1959).

Box 3.1: Timing the local group

The nearest giant spiral galaxy is the Sb galaxy M31, at a distance of about $(740 \pm 40)\,\mathrm{kpc}$ (BM §7.4.1). Our galaxy and M31 are by far the two largest members of the Local Group of galaxies. Beyond these, the next nearest prominent galaxies are in the Sculptor and M81 groups, at a distance of 3 Mpc. Thus the Local Group is an isolated system.

The line-of-sight velocity of the center of M31 relative to the center of the Galaxy is $-125\,\mathrm{km\,s^{-1}}$ (for a solar circular speed $v_0 = 220\,\mathrm{km\,s^{-1}}$, eq. 1.8); it is negative because the two galaxies are approaching one another. It seems that gravity has halted and reversed the original motion of M31 away from the Galaxy. Since M31 and the Galaxy are by far the most luminous members of the Local Group, we can treat them as an isolated system of two point masses, and estimate their total mass (Kahn & Woltjer 1959; Wilkinson & Evans 1999). Moreover, the original Hubble recession corresponded to an orbit of zero angular momentum, so we expect the angular momentum of the current orbit to be negligible. Thus we assume that the eccentricity $e = 1$.

We may now apply equations (3.28) for a Kepler orbit. Taking the log of both equations, differentiating with respect to η, and taking the ratio, we obtain

$$\frac{\mathrm{d}\ln r}{\mathrm{d}\ln t} = \frac{t}{r}\frac{\mathrm{d}r}{\mathrm{d}t} = \frac{e\sin\eta(\eta - e\sin\eta)}{(1 - e\cos\eta)^2}. \tag{1}$$

We set $e = 1$, and require that $r = 740\,\mathrm{kpc}$, $\mathrm{d}r/\mathrm{d}t = -125\,\mathrm{km\,s^{-1}}$, and $t = 13.7\,\mathrm{Gyr}$, the current age of the universe (eq. 1.77). Inserting these constraints in (1) gives a nonlinear equation for η, which is easily solved numerically to yield $\eta = 4.29$. Then equations (3.28) yield $a = 524\,\mathrm{kpc}$ and $T_r = 16.6\,\mathrm{Gyr}$, and equation (3.31) finally yields $M = 4.6 \times 10^{12}\,\mathcal{M}_\odot$ for the total mass of M31 and the Galaxy. The uncertainty in this result, assuming that our model is correct, is probably about a factor of 1.5.

This calculation assumes that the vacuum-energy density ρ_Λ is zero. Inclusion of non-zero ρ_Λ is simple (Problem 3.5); with parameters from equations (1.52) and (1.73), the required mass M increases by 15%.

The luminosity of the Galaxy in the R band is $3 \times 10^{10}\,L_\odot$ (Table 1.2) and M31 is about 1.5 times as luminous (BM Table 4.3); thus, if our mass estimate is correct, the mass-to-light ratio for the Local Group is $\Upsilon_V \simeq 60\Upsilon_\odot$. This is far larger than expected for any normal stellar population, and the total mass is far larger than the masses within the outer edges of the disks of these galaxies, as measured by circular-speed curves. Thus the Kahn–Woltjer timing argument provided the first direct evidence that most of the mass of the Local Group is composed of dark matter. For a review see Peebles (1996).

Box 3.2: The eccentricity vector for Kepler orbits

The orbit of a test particle in the Kepler potential can also be found using vector methods. Since the angular momentum per unit mass $\mathbf{L} = \mathbf{r} \times \mathbf{v}$ is constant in any central field $g(r)$, with the equation of motion (3.3) and the vector identity (B.9) we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{v} \times \mathbf{L}) = \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} \times \mathbf{L} = g(r)\hat{\mathbf{e}}_r \times (\mathbf{r} \times \mathbf{v}) = g(r)\left[(\hat{\mathbf{e}}_r \cdot \mathbf{v})\mathbf{r} - r\mathbf{v}\right]. \tag{1}$$

The time derivative of the unit radial vector is

$$\frac{\mathrm{d}\hat{\mathbf{e}}_r}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathbf{r}}{r}\right) = \frac{\mathbf{v}}{r} - \frac{\mathbf{r} \cdot \mathbf{v}}{r^3}\mathbf{r} = \frac{1}{r^2}\left[r\mathbf{v} - (\hat{\mathbf{e}}_r \cdot \mathbf{v})\mathbf{r}\right]. \tag{2}$$

Comparing equations (1) and (2) we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{v} \times \mathbf{L}) = -g(r)r^2\frac{\mathrm{d}\hat{\mathbf{e}}_r}{\mathrm{d}t}. \tag{3}$$

If and only if the field is Kepler, $g(r) = -GM/r^2$, this equation can be integrated to yield

$$\mathbf{v} \times \mathbf{L} = GM(\hat{\mathbf{e}}_r + \mathbf{e}), \tag{4}$$

where $\mathbf{e}$ is a vector constant, or integral of motion (see §3.1.1). Taking the dot product of $\mathbf{L}$ with equation (4), we find that $\mathbf{e} \cdot \mathbf{L} = 0$, so $\mathbf{e}$ lies in the orbital plane. Taking the dot product of $\mathbf{r}$ with equation (4) and using the vector identity (B.8), we have

$$L^2 = GM(r + \mathbf{e} \cdot \mathbf{r}). \tag{5}$$

If we now define ψ to be an azimuthal angle in the orbital plane, with $\mathbf{e}$ at azimuth ψ_0, then $\mathbf{e} \cdot \mathbf{r} = er\cos(\psi - \psi_0)$, where $e = |\mathbf{e}|$, and equation (5) can be rewritten

$$r = \frac{L^2}{GM}\frac{1}{1 + e\cos(\psi - \psi_0)}, \tag{6}$$

which is the same as equations (3.25b) and (3.26) for a Kepler orbit if we identify e with the eccentricity. It is therefore natural to call the vector constant $\mathbf{e}$ the **eccentricity vector**, also sometimes called the Laplace or Runge–Lenz vector. The eccentricity vector has length equal to the eccentricity and points from the central mass towards the pericenter. The direction of the eccentricity vector is called the **line of apsides**.

Orbits in other central fields have integrals of motion analogous to the scalar eccentricity, but they do not have vector integrals analogous to the eccentricity vector, because orbits in non-Kepler potentials are not closed.

It is convenient to define an auxiliary variable s by

$$s \equiv -\frac{GM}{b\Phi} = 1 + \sqrt{1 + \frac{r^2}{b^2}}. \tag{3.34}$$

Solving this equation for r, we find that

$$\frac{r^2}{b^2} = s^2\left(1 - \frac{2}{s}\right) \qquad (s \geq 2). \tag{3.35}$$

Given this one-to-one relationship between s and r, we may employ s as a radial coordinate in place of r. The integrals (3.17) and (3.18b) for T_r and $\Delta\psi$ both involve the infinitesimal quantity

$$\mathrm{d}I \equiv \frac{\mathrm{d}r}{\sqrt{2(E-\Phi) - L^2/r^2}}. \tag{3.36}$$

When we use equation (3.35) to eliminate r from this expression, we find

$$\mathrm{d}I = \frac{b(s-1)\mathrm{d}s}{\sqrt{2Es^2 - 2(2E - GM/b)s - 4GM/b - L^2/b^2}}. \tag{3.37}$$

As the star moves from pericenter r_1 to apocenter r_2, s varies from the smaller root s_1 of the quadratic expression in the denominator of equation (3.37) to the larger root s_2. Thus, combining equations (3.17) and (3.37), the radial period is

$$T_r = \frac{2b}{\sqrt{-2E}} \int_{s_1}^{s_2} \mathrm{d}s \, \frac{(s-1)}{\sqrt{(s_2 - s)(s - s_1)}} = \frac{2\pi b}{\sqrt{-2E}}\left[\tfrac{1}{2}(s_1 + s_2) - 1\right], \tag{3.38}$$

where we have assumed $E < 0$ since we are dealing with bound orbits. But from the denominator of equation (3.37) it follows that the roots s_1 and s_2 obey

$$s_1 + s_2 = 2 - \frac{GM}{Eb}, \tag{3.39a}$$

and so the radial period

$$T_r = \frac{2\pi GM}{(-2E)^{3/2}}, \tag{3.39b}$$

exactly as in the Kepler case (the limit of the isochrone as $b \to 0$). Note that T_r depends on the energy E but not on the angular momentum L—it is this unique property that gives the isochrone its name.

Equation (3.18b), for the increment $\Delta\psi$ in azimuthal angle per cycle in the radial direction, yields

$$\begin{aligned}\Delta\psi = 2L\int_{s_1}^{s_2} \frac{\mathrm{d}I}{r^2} &= \frac{2L}{b\sqrt{-2E}} \int_{s_1}^{s_2} \mathrm{d}s \, \frac{(s-1)}{s(s-2)\sqrt{(s_2 - s)(s - s_1)}} \\ &= \pi \, \mathrm{sgn}(L)\left(1 + \frac{|L|}{\sqrt{L^2 + 4GMb}}\right),\end{aligned} \tag{3.40}$$

where $\mathrm{sgn}(L) = \pm 1$ depending on the sign of L. From this expression we see that

$$\pi < |\Delta\psi| < 2\pi. \tag{3.41}$$

The only orbits for which $|\Delta\psi|$ approaches the value 2π characteristic of Kepler motion are those with $L^2 \gg 4GMb$. Such orbits never approach the core $r \lesssim b$ of the potential, and hence always move in a near-Kepler field. In the opposite limit, $L^2 \ll 4GMb$, $|\Delta\psi| \to \pi$; physically this implies that low angular-momentum orbits fly straight through the core of the potential. In fact, the behavior $|\Delta\psi| \to \pi$ as $L \to 0$ is characteristic of any spherical potential that is not strongly singular at $r = 0$—see Problem 3.19.

Inserting equations (3.39b) and (3.40) into equation (3.19), we have that the azimuthal period of an isochrone orbit is

$$T_\psi = \frac{4\pi GM}{(-2E)^{3/2}} \frac{\sqrt{L^2 + 4GMb}}{|L| + \sqrt{L^2 + 4GMb}}. \tag{3.42}$$

(d) Hyperbolic encounters In Chapter 7 we shall find that the dynamical evolution of globular clusters is largely driven by gravitational encounters between stars. These encounters are described by unbound Kepler orbits.

Let $(\mathbf{x}_M, \mathbf{v}_M)$ and $(\mathbf{x}_m, \mathbf{v}_m)$ be the positions and velocities of two point masses M and m, respectively; let $\mathbf{r} = \mathbf{x}_M - \mathbf{x}_m$ and $\mathbf{V} = \dot{\mathbf{r}}$. Then the separation vector $\mathbf{r}$ obeys equation (D.33),

$$\left(\frac{mM}{M+m}\right)\ddot{\mathbf{r}} = -\frac{GMm}{r^2}\,\hat{\mathbf{e}}_r \quad \text{or} \quad \mu\ddot{\mathbf{r}} = -\frac{G(M+m)\mu}{r^2}\,\hat{\mathbf{e}}_r. \tag{3.43}$$

This is the equation of motion of a fictitious particle, called the reduced particle, which has mass $\mu = Mm/(M+m)$ and travels in the Kepler potential of a fixed body of mass $M + m$ (see Appendix D.1). If $\Delta\mathbf{v}_m$ and $\Delta\mathbf{v}_M$ are the changes in the velocities of m and M during the encounter, we have

$$\Delta\mathbf{v}_M - \Delta\mathbf{v}_m = \Delta\mathbf{V}. \tag{3.44a}$$

Furthermore, since the velocity of the center of mass of the two bodies is unaffected by the encounter (eq. D.19), we also have

$$M\Delta\mathbf{v}_M + m\Delta\mathbf{v}_m = 0. \tag{3.44b}$$

Eliminating $\Delta\mathbf{v}_m$ between equations (3.44) we obtain $\Delta\mathbf{v}_M$ as

$$\Delta\mathbf{v}_M = \frac{m}{M+m}\Delta\mathbf{V}. \tag{3.45}$$

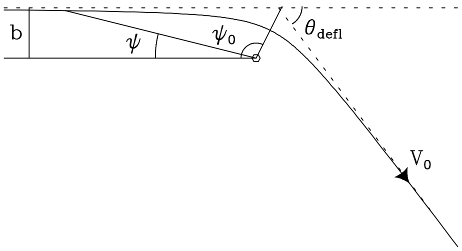


Figure 3.2 The motion of the reduced particle during a hyperbolic encounter.

We now evaluate $\Delta\mathbf{V}$.

Let the component of the initial separation vector that is perpendicular to the initial velocity vector $\mathbf{V}_0 = \mathbf{V}(t = -\infty)$ have length b (see Figure 3.2), the **impact parameter** of the encounter. Then the conserved angular momentum per unit mass associated with the motion of the reduced particle is

$$L = bV_0. \tag{3.46}$$

Equation (3.24), which relates the radius and azimuthal angle of a particle in a Kepler orbit, reads in this case,

$$\frac{1}{r} = C\cos(\psi - \psi_0) + \frac{G(M+m)}{b^2V_0^2}, \tag{3.47}$$

where the angle ψ is shown in Figure 3.2. The constants C and ψ_0 are determined by the initial conditions. Differentiating equation (3.47) with respect to time, we obtain

$$\begin{aligned}\frac{\mathrm{d}r}{\mathrm{d}t} &= Cr^2\dot{\psi}\sin(\psi - \psi_0)\\ &= CbV_0\sin(\psi - \psi_0),\end{aligned} \tag{3.48}$$

where the second line follows because $r^2\dot{\psi} = L$. If we define the direction $\psi = 0$ to point towards the particle as $t \to -\infty$, we find on evaluating equation (3.48) at $t = -\infty$,

$$-V_0 = CbV_0\sin(-\psi_0). \tag{3.49a}$$

On the other hand, evaluating equation (3.47) at this time we have

$$0 = C\cos\psi_0 + \frac{G(M+m)}{b^2V_0^2}. \tag{3.49b}$$

Eliminating C between these equations, we obtain

$$\tan\psi_0 = -\frac{bV_0^2}{G(M+m)}. \tag{3.50}$$

But from either (3.47) or (3.48) we see that the point of closest approach is reached when $\psi = \psi_0$. Since the orbit is symmetrical about this point, the angle through which the reduced particle's velocity is deflected is $\theta_{\rm defl} = 2\psi_0 - \pi$ (see Figure 3.2). It proves useful to define the **90° deflection radius** as the impact parameter at which $\theta_{\rm defl} = 90°$:

$$b_{90} \equiv \frac{G(M+m)}{V_0^2}. \tag{3.51}$$

Thus

$$\theta_{\rm defl} = 2\tan^{-1}\left(\frac{G(M+m)}{bV_0^2}\right) = 2\tan^{-1}(b_{90}/b). \tag{3.52}$$

By conservation of energy, the relative speed after the encounter equals the initial speed V_0. Hence the components $\Delta\mathbf{V}_\parallel$ and $\Delta\mathbf{V}_\perp$ of $\Delta\mathbf{V}$ parallel and perpendicular to the original relative velocity vector $\mathbf{V}_0$ are given by

$$\begin{aligned}|\Delta\mathbf{V}_\perp| &= V_0\sin\theta_{\rm defl} = V_0|\sin 2\psi_0| = \frac{2V_0|\tan\psi_0|}{1+\tan^2\psi_0} \\ &= \frac{2V_0(b/b_{90})}{1+b^2/b_{90}^2}, &\text{(3.53a)}\\ |\Delta\mathbf{V}_\parallel| &= V_0(1-\cos\theta_{\rm defl}) = V_0(1+\cos 2\psi_0) = \frac{2V_0}{1+\tan^2\psi_0} \\ &= \frac{2V_0}{1+b^2/b_{90}^2}. &\text{(3.53b)}\end{aligned}$$

$\Delta\mathbf{V}_\parallel$ always points in the direction opposite to $\mathbf{V}_0$. By equation (3.45) we obtain the components of $\Delta\mathbf{v}_M$ as

$$\begin{aligned}|\Delta\mathbf{v}_{M\perp}| &= \frac{2mV_0}{M+m}\frac{b/b_{90}}{1+b^2/b_{90}^2}, &\text{(3.54a)}\\ |\Delta\mathbf{v}_{M\parallel}| &= \frac{2mV_0}{M+m}\frac{1}{1+b^2/b_{90}^2}. &\text{(3.54b)}\end{aligned}$$

$\Delta\mathbf{v}_{M\parallel}$ always points in the direction opposite to $\mathbf{V}_0$. Notice that in the limit of large impact parameter b, $|\Delta\mathbf{v}_{M\perp}| = 2Gm/(bV_0)$, which agrees with the determination of the same quantity in equation (1.30).

3.1.1 Constants and integrals of the motion

Any stellar orbit traces a path in the six-dimensional space for which the coordinates are the position and velocity $\mathbf{x}, \mathbf{v}$. This space is called **phase space**.[1] A **constant of motion** in a given force field is any function

[1] In statistical mechanics phase space usually refers to position-momentum space rather than position-velocity space. Since all bodies have the same acceleration in a given gravitational field, mass is irrelevant, and position-velocity space is more convenient.

$C(\mathbf{x},\mathbf{v};t)$ of the phase-space coordinates and time that is constant along stellar orbits; that is, if the position and velocity along an orbit are given by $\mathbf{x}(t)$ and $\mathbf{v}(t) = \mathrm{d}\mathbf{x}/\mathrm{d}t$,

$$C[\mathbf{x}(t_1),\mathbf{v}(t_1);t_1] = C[\mathbf{x}(t_2),\mathbf{v}(t_2);t_2] \tag{3.55}$$

for any t_1 and t_2.

An **integral of motion** $I(\mathbf{x},\mathbf{v})$ is any function of the phase-space coordinates alone that is constant along an orbit:

$$I[\mathbf{x}(t_1),\mathbf{v}(t_1)] = I[\mathbf{x}(t_2),\mathbf{v}(t_2)]. \tag{3.56}$$

While every integral is a constant of the motion, the converse is not true. For example, on a circular orbit in a spherical potential the azimuthal coordinate ψ satisfies $\psi = \Omega t + \psi_0$, where Ω is the star's constant angular speed and ψ_0 is its azimuth at $t = 0$. Hence $C(\psi, t) \equiv t - \psi/\Omega$ is a constant of the motion, but it is not an integral because it depends on time as well as the phase-space coordinates.

Any orbit in any force field always has six independent constants of motion. Indeed, since the initial phase-space coordinates $(\mathbf{x}_0,\mathbf{v}_0) \equiv [\mathbf{x}(0),\mathbf{v}(0)]$ can always be determined from $[\mathbf{x}(t),\mathbf{v}(t)]$ by integrating the equations of motion backward, $(\mathbf{x}_0,\mathbf{v}_0)$ can be regarded as six constants of motion.

By contrast, orbits can have from zero to five integrals of motion. In certain important cases, a few of these integrals can be written down easily: in any static potential $\Phi(\mathbf{x})$, the Hamiltonian $H(\mathbf{x},\mathbf{v}) = \frac{1}{2}v^2 + \Phi$ is an integral of motion. If a potential $\Phi(R,z,t)$ is axisymmetric about the z axis, the z-component of the angular momentum is an integral, and in a spherical potential $\Phi(r,t)$ the three components of the angular-momentum vector $\mathbf{L} = \mathbf{x} \times \mathbf{v}$ constitute three integrals of motion. However, we shall find in §3.2 that even when integrals exist, analytic expressions for them are often not available.

These concepts and their significance for the geometry of orbits in phase space are nicely illustrated by the example of motion in a spherically symmetric potential. In this case the Hamiltonian H and the three components of the angular momentum per unit mass $\mathbf{L} = \mathbf{x} \times \mathbf{v}$ constitute four integrals. However, we shall find it more convenient to use $|\mathbf{L}|$ and the two independent components of the unit vector $\hat{\mathbf{n}} = \mathbf{L}/|\mathbf{L}|$ as integrals in place of $\mathbf{L}$. We have seen that $\hat{\mathbf{n}}$ defines the orbital plane within which the position vector $\mathbf{r}$ and the velocity vector $\mathbf{v}$ must lie. Hence we conclude that the two independent components of $\hat{\mathbf{n}}$ restrict the star's phase point to a four-dimensional region of phase space. Furthermore, the equations $H(\mathbf{x},\mathbf{v}) = E$ and $|\mathbf{L}(\mathbf{x},\mathbf{v})| = L$, where L is a constant, restrict the phase point to that two-dimensional surface in this four-dimensional region on which $v_r = \pm\sqrt{2[E - \Phi(r)] - L^2/r^2}$ and $v_\psi = L/r$. In §3.5.1 we shall see that this surface is a torus and that the sign ambiguity in v_r is analogous to the sign ambiguity in the z-coordinate of a point on the sphere $r^2 = 1$ when one specifies the point through its x and y coordinates. Thus, given E, L, and $\hat{\mathbf{n}}$, the star's position and velocity (up to its sign) can be specified by two quantities, for example r and ψ.

Is there a fifth integral of motion in a spherical potential? To study this question, we examine motion in the potential

$$\Phi(r) = -GM\left(\frac{1}{r} + \frac{a}{r^2}\right). \tag{3.57}$$

For this potential, equation (3.11b) becomes

$$\frac{\mathrm{d}^2u}{\mathrm{d}\psi^2} + \left(1 - \frac{2GMa}{L^2}\right)u = \frac{GM}{L^2}, \tag{3.58}$$

the general solution of which is

$$u = C\cos\left(\frac{\psi - \psi_0}{K}\right) + \frac{GMK^2}{L^2}, \tag{3.59a}$$

where

$$K \equiv \left(1 - \frac{2GMa}{L^2}\right)^{-1/2}. \tag{3.59b}$$

Hence

$$\psi_0 = \psi - K\,\mathrm{Arccos}\left[\frac{1}{C}\left(\frac{1}{r} - \frac{GMK^2}{L^2}\right)\right], \tag{3.60}$$

where $t =$ Arccos x is the multiple-valued solution of $x = \cos t$, and C can be expressed in terms of E and L by

$$E = \tfrac{1}{2}\frac{C^2L^2}{K^2} - \tfrac{1}{2}\left(\frac{GMK}{L}\right)^2. \tag{3.61}$$

If in equations (3.59b), (3.60) and (3.61) we replace E by $H(\mathbf{x},\mathbf{v})$ and L by $|\mathbf{L}(\mathbf{x},\mathbf{v})| = |\mathbf{x}\times\mathbf{v}|$, the quantity ψ_0 becomes a function of the phase-space coordinates which is constant as the particle moves along its orbit. Hence ψ_0 is a fifth integral of motion. (Since the function Arccos x is multiple-valued, a judicious choice of solution is necessary to avoid discontinuous jumps in ψ_0.) Now suppose that we know the numerical values of E, L, ψ_0, and the radial coordinate r. Since we have four numbers—three integrals and one coordinate—it is natural to ask how we might use these numbers to determine the azimuthal coordinate ψ. We rewrite equation (3.60) in the form

$$\psi = \psi_0 \pm K\cos^{-1}\left[\frac{1}{C}\left(\frac{1}{r} - \frac{GMK^2}{L^2}\right)\right] + 2nK\pi, \tag{3.62}$$

where $\cos^{-1}(x)$ is defined to be the value of Arccos (x) that lies between 0 and π, and n is an arbitrary integer. If K is irrational—as nearly all real numbers are—then by a suitable choice of the integer n, we can make ψ modulo 2π approximate any given number as closely as we please. Thus for any values of E and L, and any value of r between the pericenter and apocenter for the given E and L, an orbit that is known to have a given value of the integral ψ_0 can have an azimuthal angle as close as we please to any number between 0 and 2π.

On the other hand, if K is rational these problems do not arise. The simplest and most important case is that of the Kepler potential, when $a = 0$ and $K = 1$. Equation (3.62) now becomes

$$\psi = \psi_0 \pm \cos^{-1}\left[\frac{1}{C}\left(\frac{1}{r} - \frac{GM}{L^2}\right)\right] + 2n\pi, \tag{3.63}$$

which yields only two values of ψ modulo 2π for given E, L and r.

These arguments can be restated geometrically. The phase space has six dimensions. The equation $H(\mathbf{x}, \mathbf{v}) = E$ confines the orbit to a five-dimensional subspace. The vector equation $\mathbf{L}(\mathbf{x}, \mathbf{v}) = \mathit{constant}$ adds three further constraints, thereby restricting the orbit to a two-dimensional surface. Through the equation $\psi_0(\mathbf{x}, \mathbf{v}) = \mathit{constant}$ the fifth integral confines the orbit to a one-dimensional curve on this surface. Figure 3.1 can be regarded as a projection of this curve. In the Kepler case $K = 1$, the curve closes on itself, and hence does not cover the two-dimensional surface $H = \mathit{constant}$, $\mathbf{L} = \mathit{constant}$. But when K is irrational, the curve is endless and densely covers the surface of constant H and $\mathbf{L}$.

We can make an even stronger statement. Consider any volume of phase space, of any shape or size. Then if K is irrational, the fraction of the time that an orbit with given values of H and $\mathbf{L}$ spends in that volume does not depend on the value that ψ_0 takes on this orbit.

Integrals like ψ_0 for irrational K that do not affect the phase-space distribution of an orbit, are called **non-isolating integrals**. All other integrals are called **isolating integrals**. The examples of isolating integrals that we have encountered so far, namely, H, $\mathbf{L}$, and the function ψ_0 when $K = 1$, all confine stars to a five-dimensional region in phase space. However, there can also be isolating integrals that restrict the orbit to a six-dimensional subspace of phase space—see §3.7.3. Isolating integrals are of great practical and theoretical importance, whereas non-isolating integrals are of essentially no value for galactic dynamics.

3.2 Orbits in axisymmetric potentials

Few galaxies are even approximately spherical, but many approximate figures of revolution. Thus in this section we begin to explore the types of orbits that are possible in many real galaxies. As in Chapter 2, we shall usually employ a cylindrical coordinate system (R, ϕ, z) with origin at the galactic center, and shall align the z axis with the galaxy's symmetry axis.

Stars whose motions are confined to the equatorial plane of an axisymmetric galaxy have no way of perceiving that the potential in which they move is not spherically symmetric. Therefore their orbits will be identical with those we discussed in the last section; the radial coordinate R of a star on such an orbit oscillates between fixed extrema as the star revolves around the center, and the orbit again forms a rosette figure.

3.2.1 Motion in the meridional plane

The situation is much more complex and interesting for stars whose motions carry them out of the equatorial plane of the system. The study of such general orbits in axisymmetric galaxies can be reduced to a two-dimensional problem by exploiting the conservation of the z-component of angular momentum of any star. Let the potential, which we assume to be symmetric about the plane $z = 0$, be $\Phi(R, z)$. Then the motion is governed by the Lagrangian

$$\mathcal{L} = \tfrac{1}{2}\big[\dot{R}^2 + \big(R\dot{\phi}\big)^2 + \dot{z}^2\big] - \Phi(R, z). \tag{3.64}$$

The momenta are

$$p_R = \dot{R} \quad ; \quad p_\phi = R^2\dot{\phi} \quad ; \quad p_z = \dot{z}, \tag{3.65}$$

so the Hamiltonian is

$$H = \tfrac{1}{2}\Big(p_R^2 + \frac{p_\phi^2}{R^2} + p_z^2\Big) + \Phi(R, z). \tag{3.66}$$

From Hamilton's equations (D.54) we find that the equations of motion are

$$\dot{p}_R = \ddot{R} = \frac{p_\phi^2}{R^3} - \frac{\partial\Phi}{\partial R}, \tag{3.67a}$$

$$\dot{p}_\phi = \frac{\mathrm{d}}{\mathrm{d}t}\big(R^2\dot{\phi}\big) = 0, \tag{3.67b}$$

$$\dot{p}_z = \ddot{z} = -\frac{\partial\Phi}{\partial z}. \tag{3.67c}$$

Equation (3.67b) expresses conservation of the component of angular momentum about the z axis, $p_\phi = L_z$ (a constant), while equations (3.67a) and (3.67c) describe the coupled oscillations of the star in the R and z-directions.

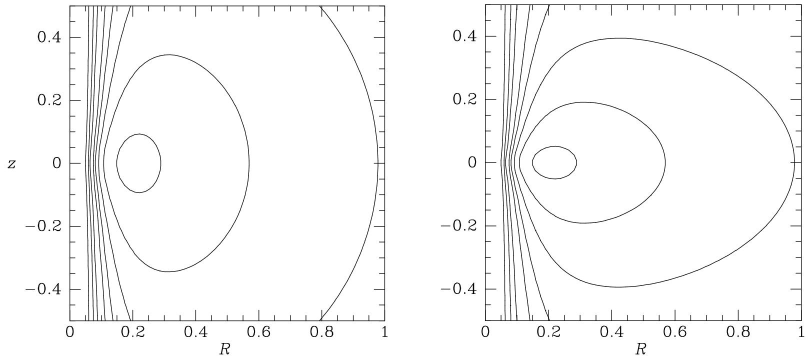


Figure 3.3 Level contours of the effective potential of equation (3.70) when $v_0 = 1$, $L_z = 0.2$. Contours are shown for $\Phi_{\rm eff} = -1$, -0.5, 0, 0.5, 1, 1.5, 2, 3, 5. The axis ratio is $q = 0.9$ in the left panel and $q = 0.5$ in the right.

After replacing p_ϕ in (3.67a) by its numerical value L_z, the first and last of equations (3.67) can be written

$$\ddot{R} = -\frac{\partial\Phi_{\rm eff}}{\partial R} \quad ; \quad \ddot{z} = -\frac{\partial\Phi_{\rm eff}}{\partial z}, \tag{3.68a}$$

where

$$\Phi_{\rm eff} \equiv \Phi(R,z) + \frac{L_z^2}{2R^2} \tag{3.68b}$$

is called the **effective potential**. Thus the three-dimensional motion of a star in an axisymmetric potential $\Phi(R,z)$ can be reduced to the two-dimensional motion of the star in the (R,z) plane (the **meridional plane**) under the Hamiltonian

$$H_{\rm eff} = \tfrac{1}{2}(p_R^2 + p_z^2) + \Phi_{\rm eff}(R,z). \tag{3.69}$$

Notice that $H_{\rm eff}$ differs from the full Hamiltonian (3.66) only in the substitution of the constant L_z for the azimuthal momentum p_ϕ. Consequently, the numerical value of $H_{\rm eff}$ is simply the orbit's total energy E. The difference $E - \Phi_{\rm eff}$ is the kinetic energy of motion in the (R,z) plane, equal to $\frac{1}{2}(p_R^2 + p_z^2)$. Since kinetic energy is non-negative, the orbit is restricted to the area in the meridional plane satisfying the inequality $E \geq \Phi_{\rm eff}$. The curve bounding this area is called the **zero-velocity curve**, since the orbit can only reach this curve if its velocity in the (R,z) plane is instantaneously zero.

Figure 3.3 shows contour plots of the effective potential

$$\Phi_{\rm eff} = \tfrac{1}{2}v_0^2 \ln\left(R^2 + \frac{z^2}{q^2}\right) + \frac{L_z^2}{2R^2}, \tag{3.70}$$

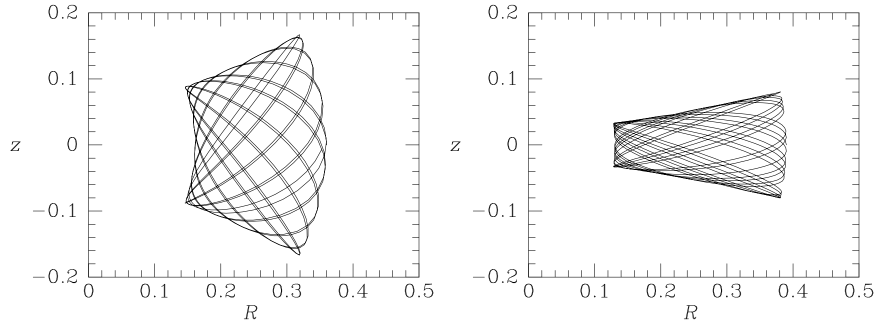


Figure 3.4 Two orbits in the potential of equation (3.70) with $q = 0.9$. Both orbits are at energy $E = -0.8$ and angular momentum $L_z = 0.2$, and we assume $v_0 = 1$.

for $v_0 = 1$, $L_z = 0.2$ and axial ratios $q = 0.9$ and 0.5. This resembles the effective potential experienced by a star in an oblate spheroidal galaxy that has a constant circular speed v_0 (§2.3.2). Notice that $\Phi_{\rm eff}$ rises very steeply near the z axis, as if the axis of symmetry were protected by a **centrifugal barrier**.

The minimum in $\Phi_{\rm eff}$ has a simple physical significance. The minimum occurs where

$$0 = \frac{\partial \Phi_{\rm eff}}{\partial R} = \frac{\partial \Phi}{\partial R} - \frac{L_z^2}{R^3} \quad ; \quad 0 = \frac{\partial \Phi_{\rm eff}}{\partial z}. \tag{3.71}$$

The second of these conditions is satisfied anywhere in the equatorial plane $z = 0$ on account of the assumed symmetry of Φ about this place, and the first is satisfied at the **guiding-center radius** $R_{\rm g}$ where

$$\left(\frac{\partial \Phi}{\partial R}\right)_{(R_{\rm g},0)} = \frac{L_z^2}{R_{\rm g}^3} = R_{\rm g}\dot{\phi}^2. \tag{3.72}$$

This is simply the condition for a circular orbit with angular speed $\dot{\phi}$. Thus the minimum of $\Phi_{\rm eff}$ occurs at the radius at which a circular orbit has angular momentum L_z, and the value of $\Phi_{\rm eff}$ at the minimum is the energy of this circular orbit.

Unless the gravitational potential Φ is of some special form, equations (3.68a) cannot be solved analytically. However, we may follow the evolution of $R(t)$ and $z(t)$ by integrating the equations of motion numerically, starting from a variety of initial conditions. Figure 3.4 shows the result of two such integrations for the potential (3.69) with $q = 0.9$ (see Richstone 1982). The orbits shown are of stars of the same energy and angular momentum, yet they look quite different in real space, and hence the stars on these orbits must move through different regions of phase space. Is this because the equations of motion admit a third isolating integral $I(R, z, p_R, p_z)$ in addition to E and L_z?

3.2.2 Surfaces of section

The phase space associated with the motion we are considering has four dimensions, R, z, p_R, and p_z, and the four-dimensional motion of the phase-space point of an individual star is too complicated to visualize. Nonetheless, we can determine whether orbits in the (R, z) plane admit an additional isolating integral by use of a simple graphical device. Since the Hamiltonian $H_{\rm eff}(R, z, p_R, p_z)$ is constant, we could plot the motion of the representative point in a three-dimensional reduced phase space, say (R, z, p_R), and then p_z would be determined (to within a sign) by the known value E of $H_{\rm eff}$. However, even three-dimensional spaces are difficult to draw, so we simply show the points where the star crosses some plane in the reduced phase space, say the plane $z = 0$; these points are called **consequents**. To remove the sign ambiguity in p_z, we plot the (R, p_R) coordinates only when $p_z > 0$. In other words, we plot the values of R and p_R every time the star crosses the equator going upward. Such plots were first used by Poincaré and are called **surfaces of section**.[2] The key feature of the surface of section is that, even though it is only two-dimensional, no two distinct orbits at the same energy can occupy the same point. Also, any orbit is restricted to an area in the surface of section defined by the constraint $H_{\rm eff} \geq \frac{1}{2}\dot{R}^2 + \Phi_{\rm eff}$; the curve bounding this area is often called the zero-velocity curve of the surface of section, since it can only be reached by an orbit with $p_z = 0$.

Figure 3.5 shows the (R, p_R) surface of section at the energy of the orbits of Figure 3.4: the full curve is the zero-velocity curve, while the dots show the consequents generated by the orbit in the left panel of Figure 3.4. The cross near the center of the surface of section, at $(R = 0.26, p_R = 0)$, is the single consequent of the **shell orbit**, in which the trajectory of the star is restricted to a two-dimensional surface. The shell orbit is the limit of orbits such as those shown in Figure 3.4 in which the distance between the inner and outer boundaries of the orbit shrinks to zero.

In Figure 3.5 the consequents of the orbit of the left panel of Figure 3.4 appear to lie on a smooth curve, called the **invariant curve** of the orbit. The existence of the invariant curve implies that some isolating integral I is respected by this orbit. The curve arises because the equation $I =$ *constant* restricts motion in the two-dimensional surface of section to a one-dimensional curve (or perhaps to a finite number of discrete points in exceptional cases). It is often found that for realistic galactic potentials, orbits do admit an integral of this type. Since I is in addition to the two classical integrals H and p_ϕ, it is called the **third integral**. In general there is no analytic expression for I as a function of the phase-space variables, so it is called a **non-classical integral**.

[2] A surface of section is defined by some arbitrarily chosen condition, here $z = 0, p_z > 0$. Good judgment must be used in the choice of this condition lest some important orbits never satisfy it, and hence do not appear on the surface of section.

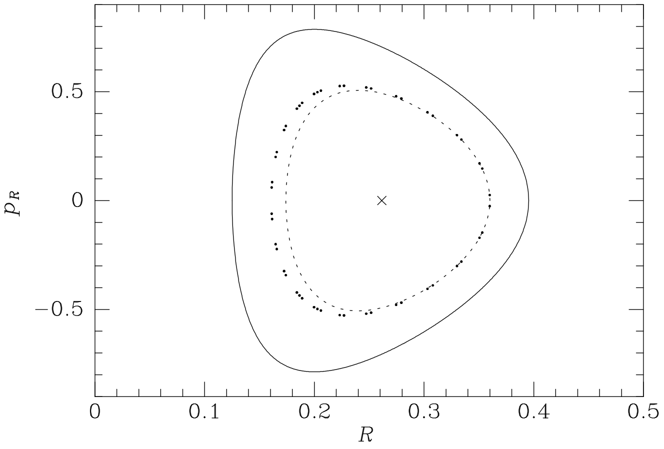


Figure 3.5 Points generated by the orbit of the left panel of Figure 3.4 in the (R, p_R) surface of section. If the total angular momentum L of the orbit were conserved, the points would fall on the dashed curve. The full curve is the zero-velocity curve at the energy of this orbit. The × marks the consequent of the shell orbit.

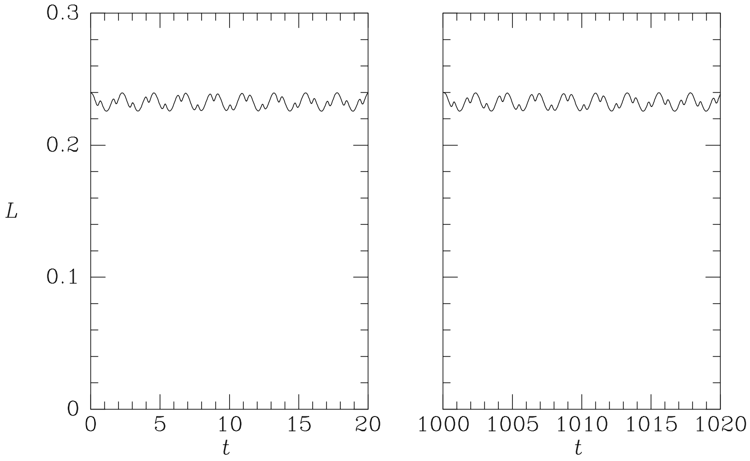


Figure 3.6 The total angular momentum is almost constant along the orbit shown in the left panel of Figure 3.5. For clarity $L(t)$ is plotted only at the beginning and end of a long integration.

We may form an intuitive picture of the nature of the third integral by considering two special cases. If the potential Φ is spherical, we know that the total angular momentum $|\mathbf{L}|$ is an integral. This suggests that for a nearly spherical potential—this one has axis ratio $q = 0.9$—the third integral may be approximated by $|\mathbf{L}|$. The dashed curve in Figure 3.5 shows the curve on which the points generated by the orbit of the left panel of Figure 3.4 would lie if the third integral were $|\mathbf{L}|$, and Figure 3.6 shows the actual time evolution of $|\mathbf{L}|$ along that orbit—notice that although $|\mathbf{L}|$ oscillates rapidly, its mean value does not change even over hundreds of orbital times. From these two figures we see that $|\mathbf{L}|$ is an approximately conserved quantity, even for orbits in potentials that are significantly flattened. We may think of these orbits as approximately planar and with more or less fixed peri- and apocenter radii. The approximate orbital planes have a fixed inclination to the z axis but precess about this axis, at a rate that gradually tends to zero as the potential becomes more and more nearly spherical.

The second special case is when the potential is separable in R and z:

$$\Phi(R,z) = \Phi_R(R) + \Phi_z(z). \tag{3.73}$$

Then the third integral can be taken to be the energy of vertical motion

$$H_z = \tfrac{1}{2}p_z^2 + \Phi_z(z). \tag{3.74}$$

Along nearly circular orbits in a thin disk, the potential is approximately separable, so equation (3.74) provides a useful expression for the third integral. In §3.6.2b we discuss a more sophisticated approximation to the third integral for orbits in thin disks.

3.2.3 Nearly circular orbits: epicycles and the velocity ellipsoid

In disk galaxies many stars are on nearly circular orbits, so it is useful to derive approximate solutions to equations (3.68a) that are valid for such orbits. We define

$$x \equiv R - R_{\rm g}, \tag{3.75}$$

where $R_{\rm g}(L_z)$ is the guiding-center radius for an orbit of angular momentum L_z (eq. 3.72). Thus $(x,z)=(0,0)$ are the coordinates in the meridional plane of the minimum in $\Phi_{\rm eff}$. When we expand $\Phi_{\rm eff}$ in a Taylor series about this point, we obtain

$$\Phi_{\rm eff} = \Phi_{\rm eff}(R_{\rm g},0)+\tfrac{1}{2}\left(\frac{\partial^2\Phi_{\rm eff}}{\partial R^2}\right)_{(R_{\rm g},0)}x^2+\tfrac{1}{2}\left(\frac{\partial^2\Phi_{\rm eff}}{\partial z^2}\right)_{(R_{\rm g},0)}z^2+{\rm O}(xz^2). \tag{3.76}$$

Note that the term that is proportional to xz vanishes because $\Phi_{\rm eff}$ is assumed to be symmetric about $z=0$. The equations of motion (3.68a) become very simple in the **epicycle approximation** in which we neglect all terms in $\Phi_{\rm eff}$ of order xz^2 or higher powers of x and z. We define two new quantities by

$$\kappa^2(R_{\rm g}) \equiv \left(\frac{\partial^2\Phi_{\rm eff}}{\partial R^2}\right)_{(R_{\rm g},0)} \quad ; \quad \nu^2(R_{\rm g}) \equiv \left(\frac{\partial^2\Phi_{\rm eff}}{\partial z^2}\right)_{(R_{\rm g},0)}, \tag{3.77}$$

for then equations (3.68a) become

$$\ddot{x} = -\kappa^2 x, \tag{3.78a}$$

$$\ddot{z} = -\nu^2 z. \tag{3.78b}$$

According to these equations, x and z evolve like the displacements of two harmonic oscillators, with frequencies κ and ν, respectively. The two frequencies κ and ν are called the **epicycle** or **radial frequency** and the **vertical frequency**. If we substitute from equation (3.68b) for $\Phi_{\rm eff}$ we obtain[3]

$$\kappa^2(R_{\rm g}) = \left(\frac{\partial^2\Phi}{\partial R^2}\right)_{(R_{\rm g},0)} + \frac{3L_z^2}{R_{\rm g}^4} = \left(\frac{\partial^2\Phi}{\partial R^2}\right)_{(R_{\rm g},0)} + \frac{3}{R_{\rm g}}\left(\frac{\partial\Phi}{\partial R}\right)_{(R_{\rm g},0)}, \tag{3.79a}$$

$$\nu^2(R_{\rm g}) = \left(\frac{\partial^2\Phi}{\partial z^2}\right)_{(R_{\rm g},0)}. \tag{3.79b}$$

Since the circular frequency is given by

$$\Omega^2(R) = \frac{1}{R}\left(\frac{\partial\Phi}{\partial R}\right)_{(R,0)} = \frac{L_z^2}{R^4}, \tag{3.79c}$$

equation (3.79a) may be written

$$\kappa^2(R_{\rm g}) = \left(R\frac{{\rm d}\Omega^2}{{\rm d}R} + 4\Omega^2\right)_{R_{\rm g}}. \tag{3.80}$$

Note that the radial and azimuthal periods (eqs. 3.17 and 3.19) are simply

$$T_r = \frac{2\pi}{\kappa} \quad ; \quad T_\psi = \frac{2\pi}{\Omega}. \tag{3.81}$$

Very near the center of a galaxy, where the circular speed rises approximately linearly with radius, Ω is nearly constant and $\kappa \simeq 2\Omega$. Elsewhere Ω declines with radius, though rarely faster than the Kepler falloff, $\Omega \propto R^{-3/2}$, which yields $\kappa = \Omega$. Thus, in general,

$$\Omega \lesssim \kappa \lesssim 2\Omega. \tag{3.82}$$

Using equations (3.19) and (3.81), it is easy to show that this range is consistent with the range of $\Delta\psi$ given by equation (3.41) for the isochrone potential.

[3] The formula for the ratio κ^2/Ω^2 from equations (3.79) was already known to Newton; see Proposition 45 of his *Principia*.

It is useful to define two functions

$$\begin{aligned} A(R) &\equiv \tfrac{1}{2}\left(\frac{v_\mathrm{c}}{R} - \frac{\mathrm{d}v_\mathrm{c}}{\mathrm{d}R}\right) = -\tfrac{1}{2}R\frac{\mathrm{d}\Omega}{\mathrm{d}R}, \\ B(R) &\equiv -\tfrac{1}{2}\left(\frac{v_\mathrm{c}}{R} + \frac{\mathrm{d}v_\mathrm{c}}{\mathrm{d}R}\right) = -\left(\Omega + \tfrac{1}{2}R\frac{\mathrm{d}\Omega}{\mathrm{d}R}\right), \end{aligned} \tag{3.83}$$

where $v_\mathrm{c}(R) = R\Omega(R)$ is the circular speed at radius R. These functions are related to the circular and epicycle frequencies by

$$\Omega = A - B \quad ; \quad \kappa^2 = -4B(A-B) = -4B\Omega. \tag{3.84}$$

The values taken by A and B at the solar radius can be measured directly from the kinematics of stars in the solar neighborhood (BM §10.3.3) and are called the **Oort constants**.[4] Taking values for these constants from Table 1.2, we find that the epicycle frequency at the Sun is $\kappa_0 = (37 \pm 3)\,\mathrm{km\,s^{-1}\,kpc^{-1}}$, and that the ratio κ_0/Ω_0 at the Sun is

$$\frac{\kappa_0}{\Omega_0} = 2\sqrt{\frac{-B}{A-B}} = 1.35 \pm 0.05. \tag{3.85}$$

Consequently the Sun makes about 1.3 oscillations in the radial direction in the time it takes to complete an orbit around the galactic center. Hence its orbit does not close on itself in an inertial frame, but forms a rosette figure like those discussed above for stars in spherically symmetric potentials.

The equations of motion (3.78) lead to two integrals, namely, the one-dimensional Hamiltonians

$$H_R \equiv \tfrac{1}{2}(\dot{x}^2 + \kappa^2 x^2) \quad ; \quad H_z \equiv \tfrac{1}{2}(\dot{z}^2 + \nu^2 z^2) \tag{3.86}$$

of the two oscillators. Thus if the star's orbit is sufficiently nearly circular that our truncation of the series for Φ_eff (eq. 3.76) is justified, then the orbit admits three integrals of motion: H_R, H_z, and p_ϕ. These are all isolating integrals.

From equations (3.75), (3.77), (3.78), and (3.86) we see that the Hamiltonian of such a star is made up of three parts:

$$H = H_R(R, p_R) + H_z(z, p_z) + \Phi_\mathrm{eff}(R_\mathrm{g}, 0). \tag{3.87}$$

[4] Jan Hendrik Oort (1900–1992) was Director of Leiden Observatory in the Netherlands from 1945 to 1970. In 1927 Oort confirmed Bertil Lindblad's hypothesis of galactic rotation with an analysis of the motions of nearby stars that established the mathematical framework for studying Galactic rotation. With his student H. van de Hulst, he predicted the 21-cm line of neutral hydrogen. Oort also established the Netherlands as a world leader in radio astronomy, and showed that many comets originate in a cloud surrounding the Sun at a distance $\sim 0.1\,\mathrm{pc}$, now called the Oort cloud.

Thus the three integrals of motion can equally be chosen as (H_R, H_z, p_ϕ) or (H, H_z, p_ϕ), and in the latter case H_z, which is a classical integral, is playing the role of the third integral.

We now investigate what the ratios of the frequencies κ, Ω and ν tell us about the properties of the Galaxy. At most points in a typical galactic disk (including the solar neighborhood) $v_\mathrm{c} \simeq$ *constant*, and from (3.80) it is easy to show that in this case $\kappa^2 = 2\Omega^2$. In cylindrical coordinates Poisson's equation for an axisymmetric galaxy reads

$$\begin{aligned} 4\pi G\rho &= \frac{1}{R}\frac{\partial}{\partial R}\Big(R\frac{\partial\Phi}{\partial R}\Big) + \frac{\partial^2\Phi}{\partial z^2} \\ &\simeq \frac{1}{R}\frac{\mathrm{d}v_\mathrm{c}^2}{\mathrm{d}R} + \nu^2, \end{aligned} \tag{3.88}$$

where in the second line we have approximated the right side by its value in the equatorial plane and used equation (3.79b). If the mass distribution were spherical, we would have $\Omega^2 \simeq GM/R^3 = \frac{4}{3}\pi G\overline{\rho}$, where M is the mass and $\overline{\rho}$ is the mean density within the sphere of radius R about the galactic center. From the plot of the circular speed of an exponential disk shown in Figure 2.17, we know that this relation is not far from correct even for a flat disk. Hence, at a typical point in a galaxy such as the Milky Way

$$\frac{\nu^2}{\kappa^2} \simeq \tfrac{3}{2}\rho/\overline{\rho}. \tag{3.89}$$

That is, the ratio ν^2/κ^2 is a measure of the degree to which the galactic material is concentrated towards the plane, and will be significantly greater than unity for a disk galaxy. From Table 1.1 we shall see that $\rho \simeq 0.1\,\mathcal{M}_\odot\,\mathrm{pc}^{-3}$, so the vertical period of small oscillations is $2\pi/\nu \simeq 87\,\mathrm{Myr}$. For $v_\mathrm{c} = 220\,\mathrm{km\,s^{-1}}$ and $R_0 = 8\,\mathrm{kpc}$ (Table 1.2) we find $\overline{\rho} = 0.039\,\mathcal{M}_\odot\,\mathrm{pc}^{-3}$. Equation (3.89) then yields $\nu/\kappa \simeq 2.0$.

From equation (3.88) it is clear that we expect $\Phi_\mathrm{eff} \propto z^2$ only for values of z small enough that $\rho_\mathrm{disk}(z) \simeq$ *constant*, i.e., for $z \ll 300\,\mathrm{pc}$ at R_0. For stars that do not rise above this height, equation (3.78b) yields

$$z = Z\cos(\nu t + \zeta), \tag{3.90}$$

where Z and ζ are arbitrary constants. However, the orbits of the majority of disk stars carry these stars further above the plane than 300 pc (Problem 4.23). Therefore the epicycle approximation does not provide a reliable guide to the motion of the majority of disk stars in the direction perpendicular to the disk. The great value of this approximation lies rather in its ability to describe the motions of stars *in* the disk plane. So far we have described only the radial component of this motion, so we now turn to the azimuthal motion.

Equation (3.78a), which governs the radial motion, has the general solution

$$x(t) = X\cos(\kappa t + \alpha), \tag{3.91}$$

where $X \geq 0$ and α are arbitrary constants. Now let $\Omega_{\rm g} = L_z/R_{\rm g}^2$ be the angular speed of the circular orbit with angular momentum L_z. Since $p_\phi = L_z$ is constant, we have

$$\begin{aligned}\dot{\phi} = \frac{p_\phi}{R^2} &= \frac{L_z}{R_{\rm g}^2}\left(1 + \frac{x}{R_{\rm g}}\right)^{-2} \\ &\simeq \Omega_{\rm g}\left(1 - \frac{2x}{R_{\rm g}}\right).\end{aligned} \tag{3.92}$$

Substituting for x from (3.91) and integrating, we obtain

$$\phi = \Omega_{\rm g}t + \phi_0 - \gamma\frac{X}{R_{\rm g}}\sin(\kappa t + \alpha), \tag{3.93a}$$

where

$$\gamma \equiv \frac{2\Omega_{\rm g}}{\kappa} = -\frac{\kappa}{2B}, \tag{3.93b}$$

where the second equality is derived using (3.84). The nature of the motion described by these equations can be clarified by erecting Cartesian axes (x, y, z) with origin at the **guiding center**, $(R, \phi) = (R_{\rm g}, \Omega_{\rm g}t + \phi_0)$. The x and z coordinates have already been defined, and the y coordinate is perpendicular to both and points in the direction of rotation.[5] To first order in the small parameter $X/R_{\rm g}$ we have

$$\begin{aligned}y &= -\gamma X\sin(\kappa t + \alpha) \\ &\equiv -Y\sin(\kappa t + \alpha).\end{aligned} \tag{3.94}$$

Equations (3.91) and (3.94) are the complete solution for an equatorial orbit in the epicycle approximation. The motion in the z-direction is independent of the motion in x and y. In the (x, y) plane the star moves on an ellipse called the **epicycle** around the guiding center (see Figure 3.7). The lengths of the semi-axes of the epicycle are in the ratio

$$\frac{X}{Y} = \gamma^{-1}. \tag{3.95}$$

For a harmonic oscillator potential $X/Y = 1$ and for a Kepler potential $X/Y = \frac{1}{2}$; the inequality (3.82) shows that in most galactic potentials

[5] In applications to the Milky Way, which rotates clockwise when viewed from the north Galactic pole, either $\hat{\mathbf{e}}_z$ is directed towards the south Galactic pole, or (x, y, z) is a left-handed coordinate system; we make the second choice in this book.

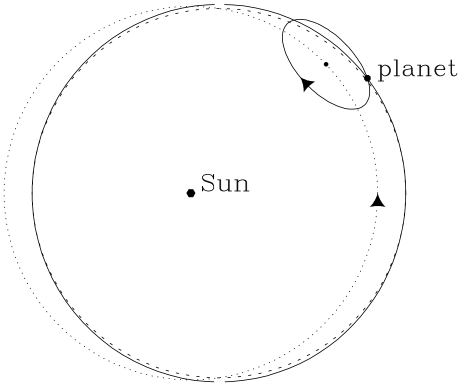


Figure 3.7 An elliptical Kepler orbit (dashed curve) is well approximated by the superposition of motion at angular frequency κ around a small ellipse with axis ratio $\frac{1}{2}$, and motion of the ellipse's center in the opposite sense at angular frequency Ω around a circle (dotted curve).

$Y > X$, so the epicycle is elongated in the tangential direction.[6] From equation (3.85), $X/Y \simeq 0.7$ in the solar neighborhood. The motion around the epicycle is in the opposite sense to the rotation of the guiding center around the galactic center, and the period of the epicycle motion is $2\pi/\kappa$, while the period of the guiding-center motion is $2\pi/\Omega_{\rm g}$.

Consider the motion of a star on an epicyclic orbit, as viewed by an astronomer who sits at the guiding center of the star's orbit. At different times in the orbit the astronomer's distance measurements range from a maximum value Y down to X. Since by equation (3.95), $X/Y = \kappa/(2\Omega_{\rm g})$, these measurements yield important information about the galactic potential. Of course, the epicycle period is much longer than an astronomer's lifetime, so we cannot in practice measure the distance to a given star as it moves around its epicycle. Moreover, in general we do not know the location of the guiding center of any given star. But we can measure v_R and $v_\phi(R_0) - v_{\rm c}(R_0)$ for a group of stars, each of which has its own guiding-center radius $R_{\rm g}$, as they pass near the Sun at radius R_0. We now show that from these measurements we can determine the ratio $2\Omega/\kappa$. We have

$$\begin{aligned}v_\phi(R_0) - v_{\rm c}(R_0) &= R_0(\dot\phi - \Omega_0) = R_0(\dot\phi - \Omega_{\rm g} + \Omega_{\rm g} - \Omega_0)\\ &\simeq R_0\left[(\dot\phi - \Omega_{\rm g}) - \left(\frac{{\rm d}\Omega}{{\rm d}R}\right)_{R_{\rm g}} x\right].\end{aligned} \tag{3.96a}$$

With equation (3.92) this becomes

$$v_\phi(R_0) - v_{\rm c}(R_0) \simeq -R_0 x \left(\frac{2\Omega}{R} + \frac{{\rm d}\Omega}{{\rm d}R}\right)_{R_{\rm g}}. \tag{3.96b}$$

[6] Epicycles were invented by the Greek astronomer Hipparchus (190–120 BC) to describe the motion of the planets about the Sun. Hipparchus also measured the distance to the Moon and discovered the precession of the Earth's spin axis. Epicycles—the first known perturbation expansion—were not very successful, largely because Hipparchus used circular epicycles with $X/Y = 1$. If only he had used epicycles with the proper axis ratio $X/Y = \frac{1}{2}$!

If we evaluate the coefficient of the small quantity x at R_0 rather than R_g, we introduce an additional error in $v_\phi(R_0)$ which is of order x^2 and therefore negligible. Making this approximation we find

$$v_\phi(R_0) - v_\mathrm{c}(R_0) \simeq -x\left(2\Omega + R\frac{\mathrm{d}\Omega}{\mathrm{d}R}\right)_{R_0}. \tag{3.96c}$$

Finally using equations (3.83) to introduce Oort's constants, we obtain

$$v_\phi(R_0) - v_\mathrm{c}(R_0) \simeq 2Bx = \frac{\kappa}{\gamma}x = \frac{\kappa}{\gamma}X\cos(\kappa t + \alpha). \tag{3.97}$$

Averaging over the phases α of stars near the Sun, we find

$$\overline{[v_\phi - v_\mathrm{c}(R_0)]^2} = \frac{\kappa^2 X^2}{2\gamma^2} = 2B^2X^2. \tag{3.98}$$

Similarly, we may neglect the dependence of κ on R_g to obtain with equation (3.84)

$$\overline{v_R^2} = \tfrac{1}{2}\kappa^2X^2 = -2B(A-B)X^2. \tag{3.99}$$

Taking the ratio of the last two equations we have

$$\frac{\overline{[v_\phi - v_\mathrm{c}(R_0)]^2}}{\overline{v_R^2}} \simeq \frac{-B}{A-B} = -\frac{B}{\Omega_0} = \frac{\kappa_0^2}{4\Omega_0^2} = \gamma^{-2} \simeq 0.46. \tag{3.100}$$

In §4.4.3 we shall re-derive this equation from a rather different point of view and compare its predictions with observational data.

Note that the ratio in equation (3.100) is the *inverse* of the ratio of the mean-square azimuthal and radial velocities relative to the guiding center: by (3.95)

$$\frac{\overline{\dot{y}^2}}{\overline{\dot{x}^2}} = \frac{\frac{1}{2}(\kappa Y)^2}{\frac{1}{2}(\kappa X)^2} = \gamma^2. \tag{3.101}$$

This counter-intuitive result arises because one measure of the RMS tangential velocity (eq. 3.101) is taken with respect to the guiding center of a single star, while the other (eq. 3.100) is taken with respect to the circular speed at the star's instantaneous radius.

This analysis also leads to an alternative expression for the integral of motion H_R defined in equation (3.86). Eliminating x using equation (3.97), we have

$$H_R = \tfrac{1}{2}\dot{x}^2 + \tfrac{1}{2}\gamma^2[v_\phi(R_0) - v_\mathrm{c}(R_0)]^2. \tag{3.102}$$

3.3 Orbits in planar non-axisymmetric potentials

Many, possibly most, galaxies have non-axisymmetric structures. These are evident near the centers of many disk galaxies, where one finds a luminous stellar bar—the Milky Way possesses just such a bar (BM §10.3). Non-axisymmetry is harder to detect in an elliptical galaxy, but we believe that many elliptical galaxies, especially the more luminous ones, are triaxial rather than axisymmetric (BM §4.2). Evidently we need to understand how stars orbit in a non-axisymmetric potential if we are to model galaxies successfully.

We start with the simplest possible problem, namely, planar motion in a non-rotating potential.[7] Towards the end of this section we generalize the discussion to two-dimensional motion in potentials whose figures rotate steadily, and in the next section we show how an understanding of two-dimensional motion can be exploited in problems involving three-dimensional potentials.

3.3.1 Two-dimensional non-rotating potential

Consider the logarithmic potential (cf. §2.3.2)

$$\Phi_L(x,y) = \tfrac{1}{2}v_0^2 \ln\left(R_\mathrm{c}^2 + x^2 + \frac{y^2}{q^2}\right) \quad (0 < q \le 1). \tag{3.103}$$

This potential has the following useful properties:

(i) The equipotentials have constant axial ratio q, so the influence of the non-axisymmetry is similar at all radii. Since $q \le 1$, the y axis is the minor axis.

(ii) For $R = \sqrt{x^2+y^2} \ll R_\mathrm{c}$, we may expand Φ_L in powers of R/R_c and find

$$\Phi_L(x,y) \simeq \frac{v_0^2}{2R_\mathrm{c}^2}\left(x^2 + \frac{y^2}{q^2}\right) + \text{constant} \qquad (R \ll R_\mathrm{c}), \tag{3.104}$$

which is just the potential of the two-dimensional harmonic oscillator. In §2.5 we saw that gravitational potentials of this form are generated by homogeneous ellipsoids. Thus for $R \lesssim R_\mathrm{c}$, Φ_L approximates the potential of a homogeneous density distribution.

(iii) For $R \gg R_\mathrm{c}$ and $q = 1$, $\Phi_L \simeq v_0^2 \ln R$, which yields a circular speed $v_\mathrm{c} \simeq v_0$ that is nearly constant. Thus the radial component of the force generated by Φ_L with $q \simeq 1$ is consistent with the flat circular-speed curves of many disk galaxies.

The simplest orbits in Φ_L are those that are confined to $R \ll R_\mathrm{c}$; when Φ_L is of the form (3.104), the orbit is the sum of independent harmonic motions

[7] This problem is equivalent to that of motion in the meridional plane of an axisymmetric potential when $L_z = 0$.

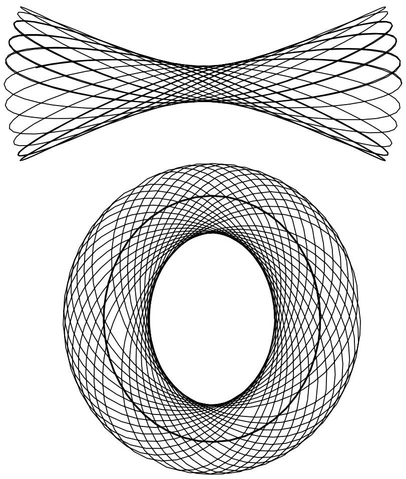


Figure 3.8 Two orbits of a common energy in the potential Φ_L of equation (3.103) when $v_0 = 1$, $q = 0.9$ and $R_\mathrm{c} = 0.14$: top, a box orbit; bottom, a loop orbit. The closed parent of the loop orbit is also shown. The energy, $E = -0.337$, is that of the isopotential surface that cuts the long axis at $x = 5R_\mathrm{c}$.

parallel to the x and y axes. The frequencies of these motions are $\omega_x = v_0/R_\mathrm{c}$ and $\omega_y = v_0/qR_\mathrm{c}$, and unless these frequencies are **commensurable** (i.e., unless $\omega_x/\omega_y = n/m$ for some integers n and m), the star eventually passes close to every point inside a rectangular box. These orbits are therefore known as **box orbits**.[8] Such orbits have no particular sense of circulation about the center and thus their time-averaged angular momentum is zero. They respect two integrals of the motion, which we may take to be the Hamiltonians of the independent oscillations parallel to the coordinate axes,

$$H_x = \tfrac{1}{2}v_x^2 + \tfrac{1}{2}v_0^2\frac{x^2}{R_\mathrm{c}^2} \qquad ; \qquad H_y = \tfrac{1}{2}v_y^2 + \tfrac{1}{2}v_0^2\frac{y^2}{q^2R_\mathrm{c}^2}. \tag{3.105}$$

To investigate orbits at larger radii $R \gtrsim R_\mathrm{c}$, we must use numerical integrations. Two examples are shown in Figure 3.8. Neither orbit fills the elliptical zero-velocity curve $\Phi_L = E$, so both orbits must respect a second integral in addition to the energy. The upper orbit is still called a box orbit because it can be thought of as a distorted form of a box orbit in the two-dimensional harmonic oscillator. Within the core the orbit's envelope runs approximately parallel to the long axis of the potential, while for $R \gg R_\mathrm{c}$ the envelope approximately follows curves of constant azimuth or radius.

In the lower orbit of Figure 3.8, the star circulates in a fixed sense about the center of the potential, while oscillating in radius. Orbits of this type are called **loop orbits**. Any star launched from $R \gg R_\mathrm{c}$ in the tangential direction with a speed of order v_0 will follow a loop orbit. If the star is launched at speed $\sim v_0$ at a large angle to the tangential direction, the annulus occupied by the orbit will be wide, while if the launch angle is small, the annulus is narrow. This dependence is analogous to the way in which

[8] The curve traced by a box orbit is sometimes called a **Lissajous figure** and is easily displayed on an oscilloscope.

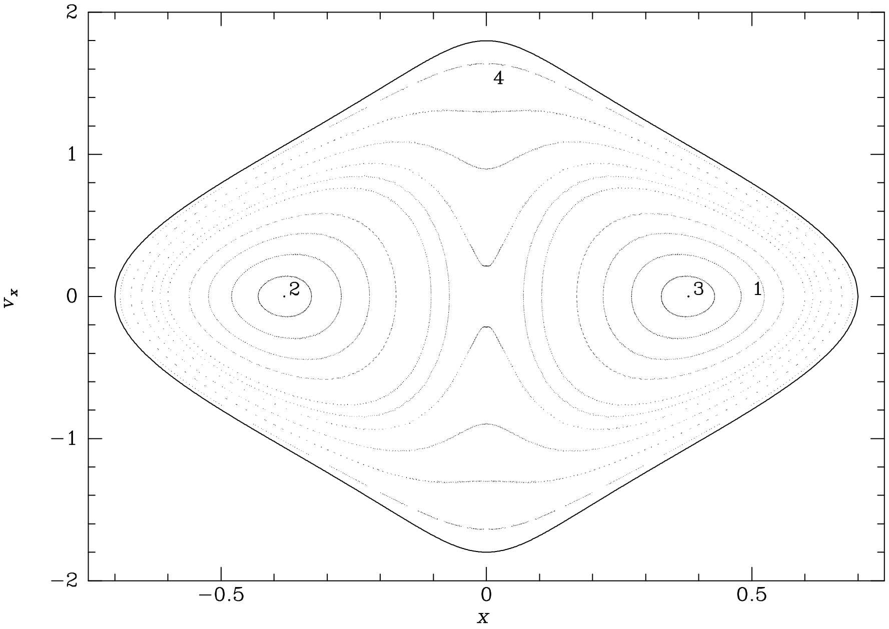


Figure 3.9 The $(x, \dot{x})$ surface of section formed by orbits in Φ_L of the same energy as the orbits depicted in Figure 3.8. The isopotential surface of this energy cuts the long axis at $x = 0.7$. The curves marked 4 and 1 correspond to the box and loop orbits shown in the top and bottom panels of Figure 3.8.

the thickness of the rosette formed by an orbit of given energy in a planar axisymmetric potential depends on its angular momentum. This analogy suggests that stars on loop orbits in Φ_L may respect an integral that is some sort of generalization of the angular momentum p_ϕ.

We may investigate these orbits further by generating a surface of section. Figure 3.9 is the surface of section $y = 0$, $\dot{y} > 0$ generated by orbits in Φ_L of the same energy as the orbits shown in Figure 3.8. The boundary curve in this figure arises from the energy constraint

$$\tfrac{1}{2}\dot{x}^2 + \Phi_L(x, 0) \le \tfrac{1}{2}(\dot{x}^2 + \dot{y}^2) + \Phi_L(x, 0) = H_{y=0}. \tag{3.106}$$

Each closed curve in this figure corresponds to a different orbit. All these orbits respect an integral I_2 in addition to the energy because each orbit is confined to a curve.

There are two types of closed curve in Figure 3.9, corresponding to the two basic types of orbit that we have identified. The lower panel of Figure 3.8 shows the spatial form of the loop orbit that generates the curve marked 1 in Figure 3.9. At a given energy there is a whole family of such orbits that differ in the width of the elliptical annuli within which they are confined—see Figure 3.10. The unique orbit of this family that circulates in

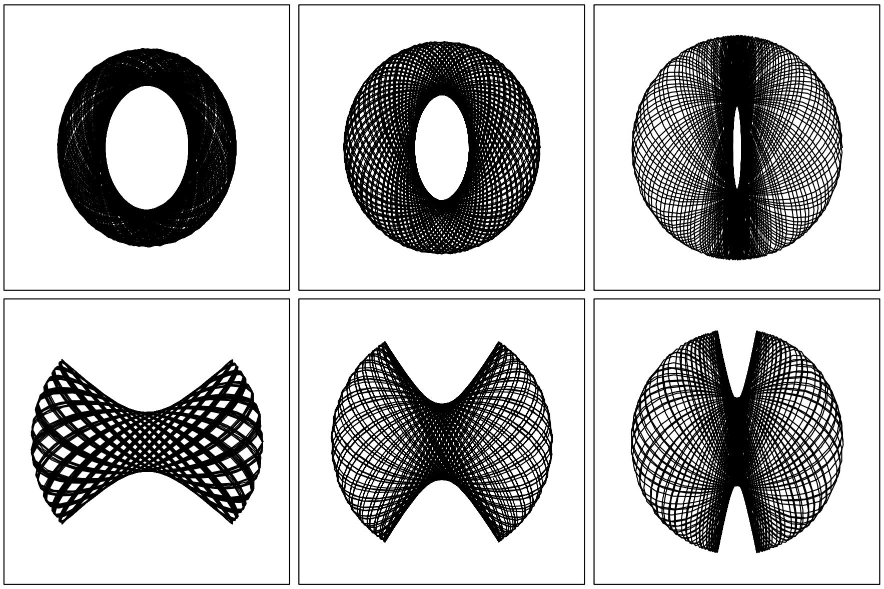


Figure 3.10 A selection of loop (top row) and box (bottom row) orbits in the potential $\Phi_L(q = 0.9, R_\mathrm{c} = 0.14)$ at the energy of Figures 3.8 and 3.9.

an anti-clockwise sense and closes on itself after one revolution is the **closed loop orbit**, which is also shown at the bottom of Figure 3.8. In the surface of section this orbit generates the single point 3. Orbits with non-zero annular widths generate the curves that loop around the point 3. Naturally, there are loop orbits that circulate in a clockwise sense in addition to the anti-clockwise orbits; in the surface of section their representative curves loop around the point 2.

The second type of closed curve in Figure 3.9 corresponds to box orbits. The box orbit shown at the top of Figure 3.8 generates the curve marked 4. All the curves in the surface of section that are symmetric about the origin, rather than centered on one of the points 2 or 3, correspond to box orbits. These orbits differ from loop orbits in two major ways: (i) in the course of time a star on any of them passes arbitrarily close to the center of the potential (in the surface of section their curves cross $x = 0$), and (ii) stars on these orbits have no unique sense of rotation about the center (in the surface of section their curves are symmetric about $x = 0$). The outermost curve in Figure 3.9 (the zero-velocity curve) corresponds to the orbit on which $y = \dot{y} = 0$; on this orbit the star simply oscillates back and forth along the x axis. We call this the **closed long-axis orbit**. The curves interior to this bounding curve that also center on the origin correspond to less and less elongated box orbits. The bottom row of Figure 3.10 shows this progression from left to right. Notice the strong resemblance of the most eccentric loop

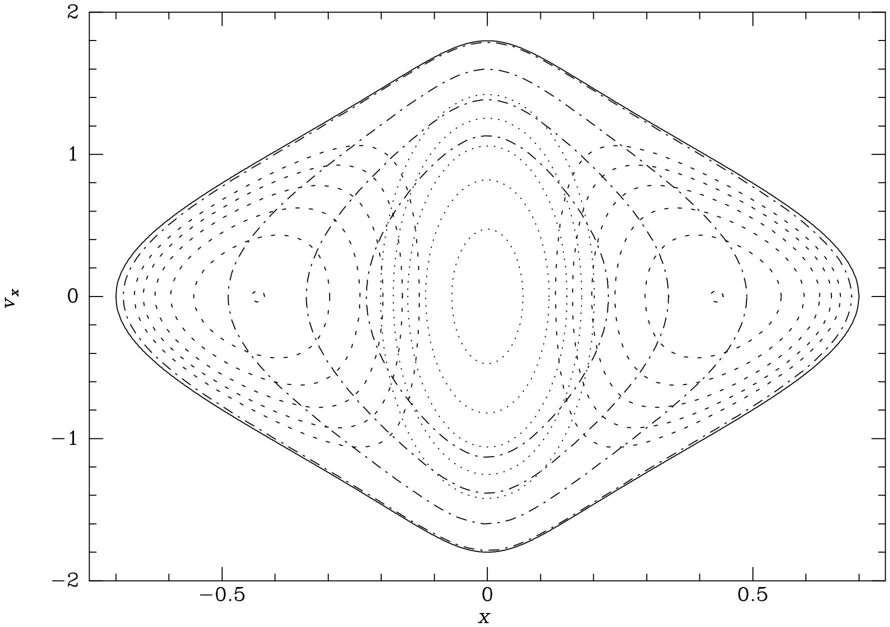


Figure 3.11 The appearance of the surface of section Figure 3.9 if orbits conserved (a) angular momentum (eq. 3.107; dashed curves), or (b) H_x (eq. 3.105; inner dotted curves), or (c) H'_x (eq. 3.108; outer dot-dashed curves).

orbit in the top right panel to the least elongated box orbit shown below it. The big difference between these orbits is that the loop orbit has a fixed sense of circulation about the center, while the box orbit does not.

It is instructive to compare the curves of Figure 3.9 with the curves generated by the integrals that we encountered earlier in this chapter. For example, if the angular momentum p_ϕ were an integral, the curves on the surface of section $y = 0$, $\dot{y} > 0$ would be given by the relation

$$(p_\phi)_{y=0} = x\dot{y} = x\sqrt{2[E - \Phi_L(x,0)] - \dot{x}^2}. \tag{3.107}$$

These curves are shown as dashed curves in Figure 3.11. They resemble the curves in Figure 3.9 near the closed loop orbits 2 and 3, thus supporting our suspicion that the integral respected by loop orbits is some generalization of angular momentum. However, the dashed curves do not reproduce the curves generated by box orbits. If the extra integral were the Hamiltonian H_x of the x-component of motion in the harmonic potential (3.105), the curves in Figure 3.9 would be the dotted ellipses near the center of Figure 3.11. They resemble the curves in Figure 3.9 that are generated by the box orbits only in that they are symmetrical about the x axis. Figure 3.11 shows that a better approximation to the invariant curves of box orbits is provided by contours of constant

$$H'_x \equiv \tfrac{1}{2}\dot{x}^2 + \Phi(x, 0). \tag{3.108}$$

H'_x may be thought of as the Hamiltonian associated with motion parallel to the potential's long axis. In a sense the integrals respected by box and loop orbits are analogous to H'_x and p_ϕ, respectively.

Figures 3.8 and 3.9 suggest an intimate connection between **closed orbits** and **families of non-closed orbits**. We say that the clockwise closed loop orbit is the **parent** of the family of clockwise loop orbits. Similarly, the closed long-axis orbit $y = 0$ is the parent of the box orbits.

The closed orbits that are the parents of orbit families are all **stable**, since members of their families that are initially close to them remain close at all times. In fact, we may think of any member of the family as engaged in stable oscillations about the parent closed orbit. A simple example of this state of affairs is provided by orbits in an axisymmetric potential. In a two-dimensional axisymmetric potential there are only two stable closed orbits at each energy—the clockwise and the anti-clockwise circular orbits.[9] All other orbits, having non-zero eccentricity, belong to families whose parents are these two orbits. The epicycle frequency (3.80) is simply the frequency of small oscillations around the parent closed orbit.

The relationship between stable closed orbits and families of non-closed orbits enables us to trace the evolution of the orbital structure of a potential as the energy of the orbits or the shape of the potential is altered, simply by tracing the evolution of the stable closed orbits. For example, consider how the orbital structure supported by Φ_L (eq. 3.103) evolves as we pass from the axisymmetric potential that is obtained when $q = 1$ to the barred potentials that are obtained when $q < 1$. When $q = 1$, p_ϕ is an integral, so the surface of section is qualitatively similar to the dashed curves in Figure 3.11. The only stable closed orbits are circular, and all orbits are loop orbits. When we make q slightly smaller than unity, the long-axis orbit becomes stable and parents a family of elongated box orbits that oscillate about the axial orbit. As q is diminished more and more below unity, a larger and larger portion of phase space comes to be occupied by box rather than loop orbits. Comparison of Figures 3.9 and 3.12 shows that this evolution manifests itself in the surface of section by the growth of the band of box orbits that runs around the outside of Figure 3.12 at the expense of the two bull's-eyes in that figure that are associated with the loop orbits. In real space the closed loop orbits become more and more elongated, with the result that less and less epicyclic motion needs to be added to one of these closed orbits to fill in the hole at its center and thus terminate the sequence of loop orbits. The erosion of the bull's-eyes in the surface of section is associated with this process.

The appearance of the surface of section also depends on the energy of its orbits. Figure 3.13 shows a surface of section for motion in $\Phi_L(q =$

[9] Special potentials such as the Kepler potential, in which all orbits are closed, must be excepted from this statement.

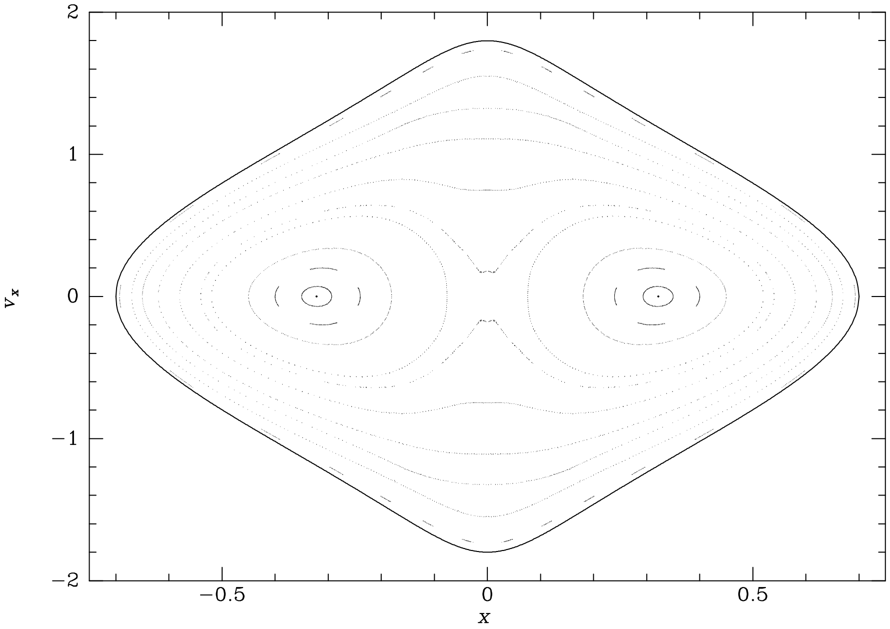


Figure 3.12 When the potential Φ_L is made more strongly barred by diminishing q, the proportion of orbits that are boxes grows at the expense of the loops: the figure shows the same surface of section as Figure 3.9 but for $q = 0.8$ rather than $q = 0.9$.

$0.9, R_c = 0.14$) at a lower energy than that of Figure 3.9. The changes in the surface of section are closely related to changes in the size and shape of the box and loop orbits. Box orbits that reach radii much greater than the core radius R_c have rather narrow waists (see Figure 3.10), and closed loop orbits of the same energy are nearly circular. If we consider box orbits and closed loop orbits of progressively smaller dimensions, the waists of the box orbits become steadily less narrow, and the closed orbits become progressively more eccentric as the dimensions of the orbits approach R_c. Eventually, at an energy E_c, the closed loop orbit degenerates into a line parallel to the short axis of the potential. Loop orbits do not exist at energies less than E_c. At $E < E_c$, all orbits are box orbits. The absence of loop orbits at $E < E_c$ is not unexpected since we saw above (eq. 3.105) that when $x^2 + y^2 \ll R_c^2$, the potential is essentially that of the two-dimensional harmonic oscillator, none of whose orbits are loops. At these energies the only closed orbits are the short- and the long-axis closed orbits, and we expect both of these orbits to be stable. In fact, the short-axis orbit becomes unstable at the energy E_c at which the loop orbits first appear. One says that the stable short-axis orbit of the low-energy regime **bifurcates** into the stable clockwise and anti-clockwise loop orbits at E_c. Stable closed orbits often appear in pairs like this.

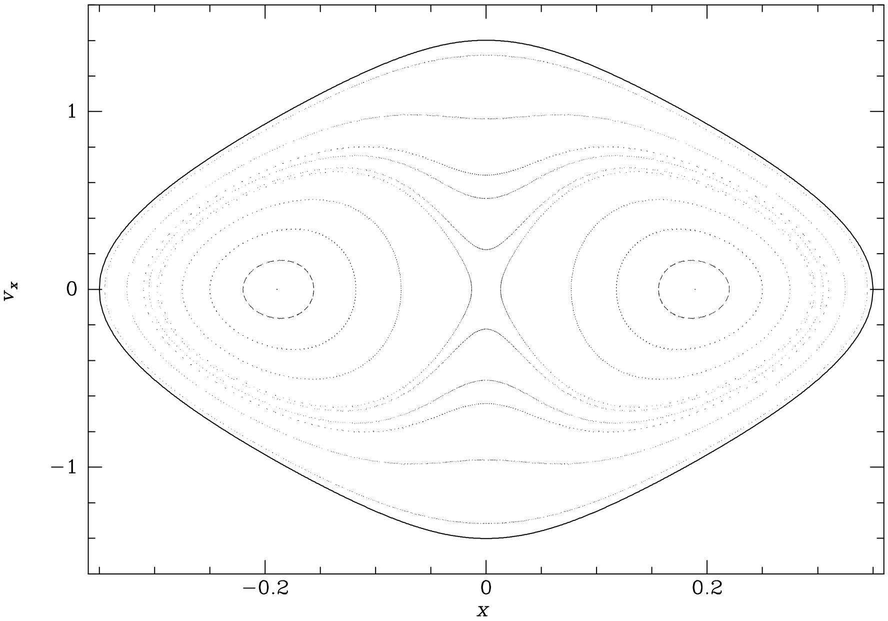


Figure 3.13 At low energies in a barred potential a large fraction of all orbits are boxes: the figure shows the same surface of section as Figure 3.9 but for the energy whose isopotential surface cuts the x axis at $x = 0.35$ rather than at $x = 0.7$ as in Figure 3.9.

Many two-dimensional barred potentials have orbital structures that resemble that of Φ_L. In particular:

(i) Most orbits in these potentials respect a second integral in addition to energy.

(ii) The majority of orbits in these potentials can be classified as either loop orbits or box orbits. The loop orbits have a fixed sense of rotation and never carry the star near the center, while the box orbits have no fixed sense of rotation and allow the star to pass arbitrarily close to the center.

(iii) When the axial ratio of the isopotential curves is close to unity, most of the phase space is filled with loop orbits, but as the axial ratio changes away from unity, box orbits fill a bigger fraction of phase space.

Although these properties are fairly general, in §3.7.3 we shall see that certain barred potentials have considerably more complex orbital structures.

3.3.2 Two-dimensional rotating potential

The figures of many non-axisymmetric galaxies rotate with respect to inertial space, so we now study orbits in rotating potentials. Let the frame of reference in which the potential Φ is static rotate steadily at angular velocity $\mathbf{\Omega}_{\rm b}$, often called the **pattern speed**. In this frame the velocity is $\dot{\mathbf{x}}$

and the corresponding velocity in an inertial frame is $\dot{\mathbf{x}} + \boldsymbol{\Omega}_\mathrm{b} \times \mathbf{x}$. Thus the Lagrangian is

$$\mathcal{L} = \tfrac{1}{2}\big|\dot{\mathbf{x}} + \boldsymbol{\Omega}_\mathrm{b} \times \mathbf{x}\big|^2 - \Phi(\mathbf{x}). \tag{3.109}$$

Consequently, the momentum is

$$\mathbf{p} = \frac{\partial \mathcal{L}}{\partial \dot{\mathbf{x}}} = \dot{\mathbf{x}} + \boldsymbol{\Omega}_\mathrm{b} \times \mathbf{x}, \tag{3.110}$$

which is just the momentum in the underlying inertial frame. The Hamiltonian is

$$\begin{aligned} H_\mathrm{J} &= \mathbf{p} \cdot \dot{\mathbf{x}} - \mathcal{L} \\ &= \mathbf{p} \cdot (\mathbf{p} - \boldsymbol{\Omega}_\mathrm{b} \times \mathbf{x}) - \tfrac{1}{2}p^2 + \Phi \\ &= \tfrac{1}{2}p^2 + \Phi - \boldsymbol{\Omega}_\mathrm{b} \cdot (\mathbf{x} \times \mathbf{p}), \end{aligned} \tag{3.111}$$

where we have used the vector identity (B.8). Since $\mathbf{p}$ coincides with the momentum in an inertial frame, $\mathbf{x} \times \mathbf{p} = \mathbf{L}$ is the angular momentum and $\frac{1}{2}p^2 + \Phi$ is the Hamiltonian H that governs the motion in the inertial frame. Hence, (3.111) can be written

$$H_\mathrm{J} = H - \boldsymbol{\Omega}_\mathrm{b} \cdot \mathbf{L}. \tag{3.112}$$

Since $\Phi(\mathbf{x})$ is constant in the rotating frame, H_J has no explicit time dependence, and its derivative along any orbit $\mathrm{d}H_\mathrm{J}/\mathrm{d}t = \partial H_\mathrm{J}/\partial t$ vanishes (eq. D.56). Thus H_J is an integral, called the **Jacobi integral**: in a rotating non-axisymmetric potential, neither H nor $\mathbf{L}$ is conserved, but the combination $H - \boldsymbol{\Omega}_\mathrm{b} \cdot \mathbf{L}$ *is* conserved. From (3.111) it is easy to show that the constant value of H_J may be written as

$$\begin{aligned} E_\mathrm{J} &= \tfrac{1}{2}|\dot{\mathbf{x}}|^2 + \Phi - \tfrac{1}{2}|\boldsymbol{\Omega}_\mathrm{b} \times \mathbf{x}|^2 \\ &= \tfrac{1}{2}|\dot{\mathbf{x}}|^2 + \Phi_\mathrm{eff}, \end{aligned} \tag{3.113}$$

where the effective potential

$$\begin{aligned} \Phi_\mathrm{eff}(\mathbf{x}) &\equiv \Phi(\mathbf{x}) - \tfrac{1}{2}|\boldsymbol{\Omega}_\mathrm{b} \times \mathbf{x}|^2 \\ &= \Phi(\mathbf{x}) - \tfrac{1}{2}\big[|\boldsymbol{\Omega}_\mathrm{b}|^2|\mathbf{x}|^2 - (\boldsymbol{\Omega}_\mathrm{b} \cdot \mathbf{x})^2\big]. \end{aligned} \tag{3.114}$$

In deriving the second line we have used the identity (B.10). The effective potential is the sum of the gravitational potential and a repulsive **centrifugal potential**. For $\boldsymbol{\Omega}_\mathrm{b} = \Omega_\mathrm{b}\hat{\mathbf{e}}_z$, this additional term is simply $-\frac{1}{2}\Omega^2 R^2$ in cylindrical coordinates.

With equation (3.111) Hamilton's equations become

$$\begin{aligned} \dot{\mathbf{p}} &= -\frac{\partial H_\mathrm{J}}{\partial \mathbf{x}} = -\boldsymbol{\nabla}\Phi - \boldsymbol{\Omega}_\mathrm{b} \times \mathbf{p} \\ \dot{\mathbf{x}} &= \frac{\partial H_\mathrm{J}}{\partial \mathbf{p}} = \mathbf{p} - \boldsymbol{\Omega}_\mathrm{b} \times \mathbf{x}, \end{aligned} \tag{3.115}$$

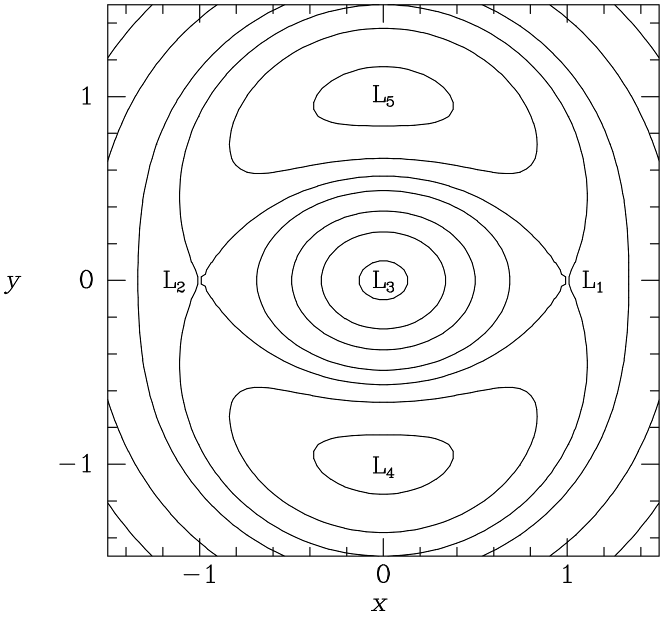


Figure 3.14 Contours of constant effective potential $\Phi_{\rm eff}$ when the potential is given by equation (3.103) with $v_0 = 1$, $q = 0.8$, $R_{\rm c} = 0.1$, and $\Omega_{\rm b} = 1$. The point marked L_3 is a minimum of $\Phi_{\rm eff}$, while those marked L_4 and L_5 are maxima. $\Phi_{\rm eff}$ has saddle points at L_1 and L_2.

where we have used the identity (B.40). Eliminating **p** between these equations we find

$$\begin{aligned}\ddot{\mathbf{x}} &= -\boldsymbol{\nabla}\Phi - 2\boldsymbol{\Omega}_{\rm b}\times\dot{\mathbf{x}} - \boldsymbol{\Omega}_{\rm b}\times(\boldsymbol{\Omega}_{\rm b}\times\mathbf{x})\\ &= -\boldsymbol{\nabla}\Phi - 2\boldsymbol{\Omega}_{\rm b}\times\dot{\mathbf{x}} + |\boldsymbol{\Omega}_{\rm b}|^2\mathbf{x} - \boldsymbol{\Omega}_{\rm b}(\boldsymbol{\Omega}_{\rm b}\cdot\mathbf{x}).\end{aligned}\tag{3.116}$$

Here $-2\boldsymbol{\Omega}_{\rm b}\times\dot{\mathbf{x}}$ is known as the **Coriolis force** and $-\boldsymbol{\Omega}_{\rm b}\times(\boldsymbol{\Omega}_{\rm b}\times\mathbf{x})$ is the **centrifugal force**. Taking the gradient of the last line of equation (3.114), we see that (3.116) can be written in the simpler form

$$\ddot{\mathbf{x}} = -\boldsymbol{\nabla}\Phi_{\rm eff} - 2\boldsymbol{\Omega}_{\rm b}\times\dot{\mathbf{x}}.\tag{3.117}$$

The surface $\Phi_{\rm eff} = E_{\rm J}$ is often called the **zero-velocity surface**. All regions in which $\Phi_{\rm eff} > E_{\rm J}$ are forbidden to the star. Thus, although the solution of the differential equations for the orbit in a rotating potential may be difficult, we can at least define forbidden regions into which the star cannot penetrate.

Figure 3.14 shows contours of $\Phi_{\rm eff}$ for the potential Φ_L of equation (3.103). $\Phi_{\rm eff}$ is characterized by five stationary points, marked L_1 to L_5, at which $\boldsymbol{\nabla}\Phi_{\rm eff} = 0$. These points are sometimes called **Lagrange points** after similar points in the restricted three-body problem (Figure 8.6). The central stationary point L_3 in Figure 3.14 is a minimum of the potential and is surrounded by a region in which the centrifugal potential $-\frac{1}{2}\Omega_{\rm b}^2 R^2$ makes only a small contribution to $\Phi_{\rm eff}$. At each of the four points L_1, L_2, L_4, and L_5, it is possible for a star to travel on a circular orbit while appearing to be stationary in the rotating frame, because the gravitational and centrifugal forces precisely balance. Such orbits are said to **corotate** with the potential. The stationary points L_1 and L_2 on the x axis (the long axis of the potential) are saddle points, while the stationary points L_4 and L_5 along the y axis are maxima of the effective potential. Stars with values of $E_{\rm J}$ smaller than the value $\Phi_{\rm c}$ taken by $\Phi_{\rm eff}$ at L_1 and L_2 cannot move from the center of the potential to infinity, or indeed anywhere outside the inner equipotential contour that runs through L_1 and L_2. By contrast, a star for which $E_{\rm J}$ exceeds $\Phi_{\rm c}$, or any star that is initially outside the contour through L_1 and L_2, can *in principle* escape to infinity. However, it cannot be assumed that a star of the latter class will *necessarily* escape, because the Coriolis force prevents stars from accelerating steadily in the direction of $-\boldsymbol{\nabla}\Phi_{\rm eff}$.

We now consider motion near each of the Lagrange points L_1 to L_5. These are stationary points of $\Phi_{\rm eff}$, so when we expand $\Phi_{\rm eff}$ around one of these points $\mathbf{x}_{\rm L} = (x_{\rm L}, y_{\rm L})$ in powers of $(x - x_{\rm L})$ and $(y - y_{\rm L})$, we have

$$\begin{aligned}\Phi_{\rm eff}(x,y) = \Phi_{\rm eff}(x_{\rm L}, y_{\rm L}) &+ \tfrac{1}{2}\left(\frac{\partial^2\Phi_{\rm eff}}{\partial x^2}\right)_{\mathbf{x}_{\rm L}}(x - x_{\rm L})^2 \\ + \left(\frac{\partial^2\Phi_{\rm eff}}{\partial x\partial y}\right)_{\mathbf{x}_{\rm L}}(x - x_{\rm L})(y - y_{\rm L}) &+ \tfrac{1}{2}\left(\frac{\partial^2\Phi_{\rm eff}}{\partial y^2}\right)_{\mathbf{x}_{\rm L}}(y - y_{\rm L})^2 + \cdots \ .\end{aligned} \tag{3.118}$$

Furthermore, for any bar-like potential whose principal axes lie along the coordinate axes, $\partial^2\Phi_{\rm eff}/\partial x\partial y = 0$ at $\mathbf{x}_{\rm L}$ by symmetry. Hence, if we retain only quadratic terms in equation (3.118) and define

$$\xi \equiv x - x_{\rm L} \quad ; \quad \eta \equiv y - y_{\rm L}, \tag{3.119}$$

and

$$\Phi_{xx} \equiv \left(\frac{\partial^2\Phi_{\rm eff}}{\partial x^2}\right)_{\mathbf{x}_{\rm L}} \quad ; \quad \Phi_{yy} \equiv \left(\frac{\partial^2\Phi_{\rm eff}}{\partial y^2}\right)_{\mathbf{x}_{\rm L}} , \tag{3.120}$$

the equations of motion (3.117) become for a star near $\mathbf{x}_{\rm L}$,

$$\ddot{\xi} = 2\Omega_{\rm b}\dot{\eta} - \Phi_{xx}\xi \quad ; \quad \ddot{\eta} = -2\Omega_{\rm b}\dot{\xi} - \Phi_{yy}\eta. \tag{3.121}$$

This is a pair of linear differential equations with constant coefficients. The general solution can be found by substituting $\xi = X\exp(\lambda t)$, $\eta = Y\exp(\lambda t)$, where X, Y, and λ are complex constants. With these substitutions, equations (3.121) become

$$(\lambda^2+\Phi_{xx})X-2\lambda\Omega_{\rm b}Y=0\ \ ;\ \ 2\lambda\Omega_{\rm b}X+(\lambda^2+\Phi_{yy})Y=0. \tag{3.122}$$

These simultaneous equations have a non-trivial solution for X and Y only if the determinant

$$\begin{vmatrix}\lambda^2+\Phi_{xx} & -2\lambda\Omega_{\rm b}\\ 2\lambda\Omega_{\rm b} & \lambda^2+\Phi_{yy}\end{vmatrix}=0. \tag{3.123}$$

Thus we require

$$\lambda^4+\lambda^2\left(\Phi_{xx}+\Phi_{yy}+4\Omega_{\rm b}^2\right)+\Phi_{xx}\Phi_{yy}=0. \tag{3.124}$$

This is the **characteristic equation** for λ. It has four roots, which may be either real or complex. If λ is a root, $-\lambda$ is also a root, so if there is any root that has non-zero real part $\mathrm{Re}(\lambda)=\gamma$, the general solution to equations (3.121) will contain terms that cause $|\xi|$ and $|\eta|$ to grow exponentially in time; $|\xi|\propto\exp(|\gamma|t)$ and $|\eta|\propto\exp(|\gamma|t)$. Under these circumstances essentially all orbits rapidly flee from the Lagrange point, and the approximation on which equations (3.121) rest breaks down. In this case the Lagrange point is said to be **unstable**.

When all the roots of equation (3.124) are pure imaginary, say $\lambda=\pm i\alpha$ or $\pm i\beta$, with $0\le\alpha\le\beta$ real, the general solution to equations (3.121) is

$$\begin{aligned}\xi&=X_1\cos(\alpha t+\phi_1)+X_2\cos(\beta t+\phi_2),\\ \eta&=Y_1\sin(\alpha t+\phi_1)+Y_2\sin(\beta t+\phi_2),\end{aligned} \tag{3.125}$$

and the Lagrange point is stable, since the perturbations ξ and η oscillate rather than growing. Substituting these equations into the differential equations (3.121), we find that X_1 and Y_1 and X_2 and Y_2 are related by

$$Y_1=\frac{\Phi_{xx}-\alpha^2}{2\Omega_{\rm b}\alpha}X_1=\frac{2\Omega_{\rm b}\alpha}{\Phi_{yy}-\alpha^2}X_1, \tag{3.126a}$$

and

$$Y_2=\frac{\Phi_{xx}-\beta^2}{2\Omega_{\rm b}\beta}X_2=\frac{2\Omega_{\rm b}\beta}{\Phi_{yy}-\beta^2}X_2. \tag{3.126b}$$

The following three conditions are necessary and sufficient for both roots λ^2 of the quadratic equation (3.124) in λ^2 to be real and negative, and hence for the Lagrange point to be stable:

$$\begin{aligned}&\text{(i)} && \lambda_1^2\lambda_2^2=\Phi_{xx}\Phi_{yy}>0,\\ &\text{(ii)} && \lambda_1^2+\lambda_2^2=-\left(\Phi_{xx}+\Phi_{yy}+4\Omega_{\rm b}^2\right)<0,\\ &\text{(iii)} && \lambda^2\text{ real }\Rightarrow(\Phi_{xx}+\Phi_{yy}+4\Omega_{\rm b}^2)^2>4\Phi_{xx}\Phi_{yy}.\end{aligned} \tag{3.127}$$

At saddle points of $\Phi_{\rm eff}$ such as L_1 and L_2, Φ_{xx} and Φ_{yy} have opposite signs, so these Lagrange points violate condition (i) and are always unstable. At a minimum of $\Phi_{\rm eff}$, such as L_3, Φ_{xx} and Φ_{yy} are both positive, so conditions (i) and (ii) are satisfied. Condition (iii) is also satisfied because it can be rewritten in the form

$$(\Phi_{xx} - \Phi_{yy})^2 + 8\Omega_{\rm b}^2 (\Phi_{xx} + \Phi_{yy}) + 16\Omega_{\rm b}^4 > 0, \qquad (3.128)$$

which is satisfied whenever both Φ_{xx} and Φ_{yy} are positive. Hence L_3 is stable.

For future use we note that when Φ_{xx} and Φ_{yy} are positive, we may assume $\Phi_{xx} < \Phi_{yy}$ (since the x axis is the major axis of the potential) and we have already assumed that $\alpha < \beta$, so we can show from (3.124) that

$$\alpha^2 < \Phi_{xx} < \Phi_{yy} < \beta^2. \qquad (3.129)$$

Also, when $\Omega_{\rm b}^2 \to 0$, α^2 tends to Φ_{xx}, and β^2 tends to Φ_{yy}.

The stability of the Lagrange points at maxima of $\Phi_{\rm eff}$, such as L_4 and L_5, depends on the details of the potential. For the potential Φ_L of equation (3.103) we have

$$\Phi_{\rm eff} = \tfrac{1}{2} v_0^2 \ln\left(R_{\rm c}^2 + x^2 + \frac{y^2}{q^2}\right) - \tfrac{1}{2}\Omega_{\rm b}^2 (x^2 + y^2), \qquad (3.130)$$

so L_4 and L_5 occur at $(0, \pm y_{\rm L})$, where

$$y_{\rm L} \equiv \sqrt{\frac{v_0^2}{\Omega_{\rm b}^2} - q^2 R_{\rm c}^2}, \qquad (3.131)$$

and we see that L_4, L_5 are present only if $\Omega_{\rm b} < v_0/(qR_{\rm c})$. Differentiating the effective potential again we find

$$\begin{aligned}\Phi_{xx}(0, y_{\rm L}) &= -\Omega_{\rm b}^2 (1 - q^2)\\ \Phi_{yy}(0, y_{\rm L}) &= -2\Omega_{\rm b}^2 \left[1 - q^2 \left(\frac{\Omega_{\rm b} R_{\rm c}}{v_0}\right)^2\right].\end{aligned} \qquad (3.132)$$

Hence $\Phi_{xx}\Phi_{yy}$ is positive if the Lagrange points exist, and stability condition (i) of (3.127) is satisfied. Deciding whether the other stability conditions hold is tedious in the general case, but straightforward in the limit of negligible core radius, $\Omega_{\rm b} R_{\rm c}/v_0 \ll 1$ (which applies, for example, to Figure 3.14). Then $\Phi_{xx} + \Phi_{yy} + 4\Omega_{\rm b}^2 = \Omega_{\rm b}^2(1 + q^2)$, so condition (ii) is satisfied. A straightforward calculation shows that condition (iii) holds—and thus that L_4 and L_5 are stable—providing $q^2 > \sqrt{32} - 5 \simeq (0.810)^2$. For future use we note that for small $R_{\rm c}$, and to leading order in the ellipticity $\epsilon = 1 - q$, we have

$$\alpha^2 = 2\epsilon\Omega_{\rm b}^2 = -\Phi_{xx} \quad ; \quad \beta^2 = 2(1 - 2\epsilon)\Omega_{\rm b}^2 = 2\Omega_{\rm b}^2 + {\rm O}(\epsilon). \qquad (3.133)$$

Equations (3.125) describing the motion about a stable Lagrange point show that each orbit is a superposition of motion at frequencies α and β around two ellipses. The shapes of these ellipses and the sense of the star's motion on them are determined by equations (3.126). For example, in the case of small R_c and ϵ, the α-ellipse around the point L_4 is highly elongated in the x- or ξ-direction (the tangential direction), while the β-ellipse has $Y_2 = -X_2/\sqrt{2}$. The star therefore moves around the β-ellipse in the sense opposite to that of the rotation of the potential. The β-ellipse is simply the familiar epicycle from §3.2.3, while the α-ellipse represents a slow tangential wallowing in the weak non-axisymmetric component of Φ_L.

Now consider motion about the central Lagrange point L_3. From equations (3.126) and the inequality (3.129), it follows that $Y_1/X_1 > 0$. Thus the star's motion around the α-ellipse has the same sense as the rotation of the potential; such an orbit is said to be **prograde** or **direct**. When $\Omega_\mathrm{b}^2 \ll |\Phi_{xx}|$, it is straightforward to show from equations (3.124) and (3.126) that $X_1 \gg Y_1$ and hence that this prograde motion runs almost parallel to the long axis of the potential—this is the long-axis orbit familiar to us from our study of non-rotating bars. Conversely the star moves around the β-ellipse in the sense opposite to that of the rotation of the potential (the motion is **retrograde**), and $|X_2| < |Y_2|$. When $\Omega_\mathrm{b}^2/|\Phi_{xx}|$ is small, the β-ellipse goes over into the short-axis orbit of a non-rotating potential. A general prograde orbit around L_3 is made up of motion on the β-ellipse around a guiding center that moves around the α-ellipse, and conversely for retrograde orbits.

We now turn to a numerical study of orbits in rotating potentials that are not confined to the vicinity of a Lagrange point. We adopt the logarithmic potential (3.103) with $q = 0.8$, $R_\mathrm{c} = 0.03$, $v_0 = 1$, and $\Omega_\mathrm{b} = 1$. This choice places the corotation annulus near $R_\mathrm{CR} = 30R_\mathrm{c}$. The Jacobi integral (eq. 3.112) now plays the role that energy played in our similar investigation of orbits in non-rotating potentials, and by a slight abuse of language we shall refer to its value E_J as the "energy". At radii $R \lesssim R_\mathrm{c}$ the two important sequences of stable closed orbits in the non-rotating case are the long- and the short-axis orbits. Figure 3.15 confirms the prediction of our analytic treatment that in the presence of rotation these become oval in shape. Orbits of both sequences are stable and therefore parent families of non-closed orbits.

Consider now the evolution of the orbital structure as we leave the core region. At an energy E_1, similar to that at which loop orbits first appeared in the non-rotating case, pairs of prograde orbits like those shown in Figure 3.16 appear. Only one member of the pair is stable. When it first appears, the stable orbit is highly elongated parallel to the short axis, but as the energy is increased it becomes more round. Eventually the decrease in the elongation of this orbit with increasing energy is reversed, the orbit again becomes highly elongated parallel to the short axis and finally disappears

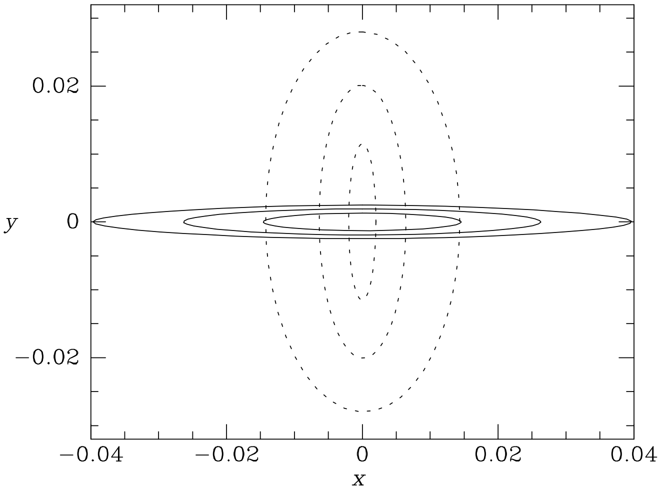


Figure 3.15 In the near-harmonic core of a rotating potential, the closed orbits are elongated ellipses. Stars on the orbits shown as full curves circulate about the center in the same sense as the potential's figure rotates. On the dashed orbits, stars circulate in the opposite sense. The x axis is the long axis of the potential.

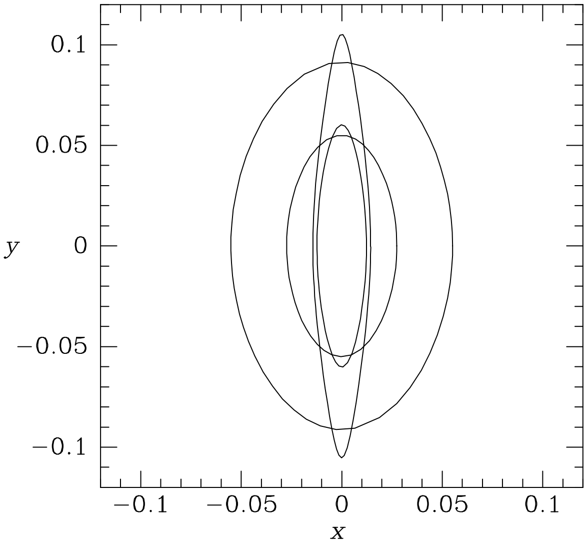


Figure 3.16 Closed orbits at two energies higher than those shown in Figure 3.15. Just outside the potential's near-harmonic core there are at each energy two prograde closed orbits aligned parallel to the potential's short axis. One of these orbits (the less elongated) is stable, while the other is unstable.

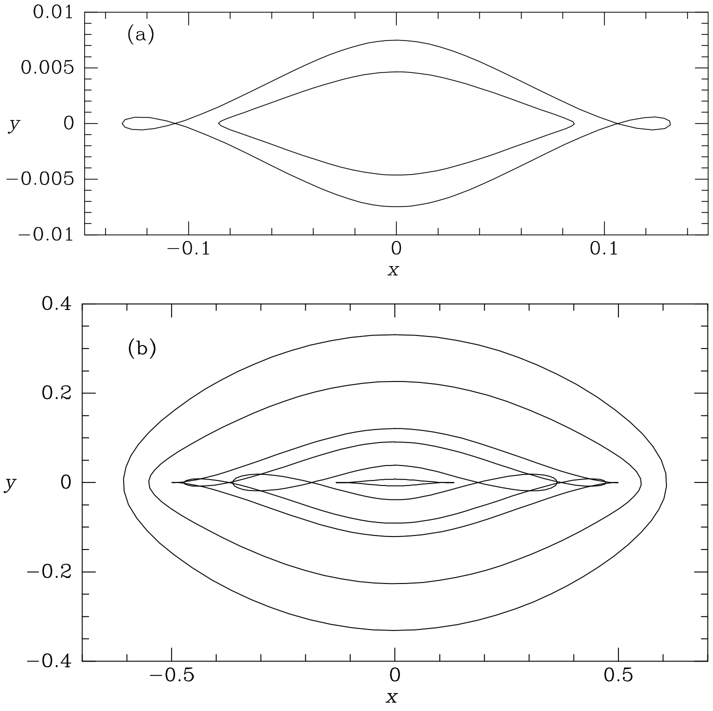


Figure 3.17 Near the energy at which the orbit pairs shown in Figure 3.16 appear, the closed long-axis orbits develop ears. Panel (a) shows orbits at energies just below and above this transition. Panel (b) shows the evolution of the closed long-axis orbits at higher energies. Notice that in panel (a) the x- and y-scales are different. The smallest orbit in panel (b) is the larger of the two orbits in panel (a).

along with its unstable companion orbit at an energy E_2.[10] In the notation of Contopoulos & Papayannopoulos (1980) these stable orbits are said to belong to the **sequence** $\mathbf{x}_2$, while their unstable companions are of the **sequence** $\mathbf{x}_3$.

The sequence of long-axis orbits (often called the **sequence** $\mathbf{x}_1$) suffers a significant transition near E_2. On the low-energy side of the transition the long-axis orbits are extremely elongated and lens shaped (smaller orbit in Figure 3.17a). On the high-energy side the orbits are self-intersecting (larger orbit in Figure 3.17a). As the energy continues to increase, the orbit's ears become first more prominent and then less prominent, vanishing to form a cusped orbit (Figure 3.17b). At still higher energies the orbits become approximately elliptical (largest orbit in Figure 3.17b), first growing rounder and then adopt progressively more complex shapes as they approach

[10] In the theory of weak bars, the energies E_1 and E_2 at which these prograde orbits appear and disappear are associated with the first and second inner Lindblad radii, respectively (eq. 3.150).

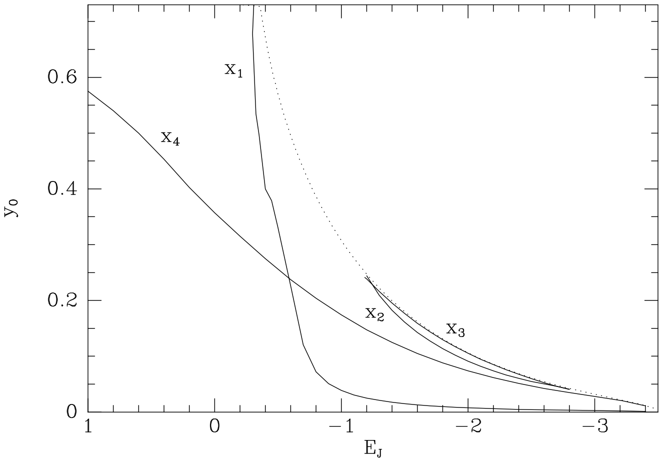


Figure 3.18 A plot of the Jacobi constant E_J of closed orbits in $\Phi_L(q = 0.8, R_c = 0.03, \Omega_b = 1)$ against the value of y at which the orbit cuts the potential's short axis. The dotted curve shows the relation $\Phi_{\rm eff}(0, y) = E_J$. The families of orbits x_1–x_4 are marked.

the corotation region in which the Lagrange points L_1, L_2, L_4, and L_5 are located.

In the vicinity of the corotation annulus, there are important sequences of closed orbits on which stars move around one of the Lagrange points L_4 or L_5, rather than about the center.

Essentially all closed orbits that carry stars well outside the corotation region are nearly circular. In fact, the potential's figure spins much more rapidly than these stars circulate on their orbits, so the non-axisymmetric forces on such stars tend to be averaged out. One finds that at large radii prograde orbits tend to align with the bar, while retrograde orbits align perpendicular to the bar.

These results are summarized in Figure 3.18. In this figure we plot against the value of E_J for each closed orbit the distance y at which it crosses the short axis of the potential. Each sequence of closed orbits generates a continuous curve in this diagram known as the **characteristic curve** of that sequence.

The stable closed orbits we have described are all associated with substantial families of non-closed orbits. Figure 3.19 shows two of these. As in the non-rotating case, a star on one of these non-closed orbits may be considered to be executing stable oscillations about one of the fundamental closed orbits. In potentials of the form (3.103) essentially all orbits belong to one of these families. This is not always true, however, as we explain in §3.7.

It is important to distinguish between orbits that enhance the elongation of the potential and those that oppose it. The overall mass distribution of a

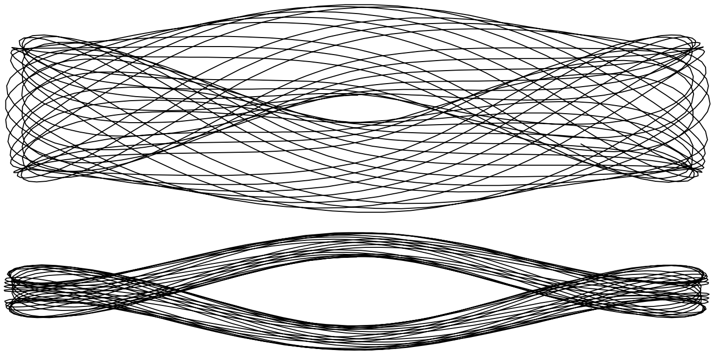


Figure 3.19 Two non-closed orbits of a common energy in the rotating potential Φ_L.

galaxy must be elongated in the same sense as the potential, which suggests that most stars are on orbits on which they spend the majority of their time nearer to the potential's long axis than to its short axis. Interior to the corotation radius, the only orbits that satisfy this criterion are orbits of the family parented by the long-axis orbits, which therefore must be the most heavily populated orbits in any bar that is confined by its own gravity. The shapes of these orbits range from butterfly-like at radii comparable to the core radius $R_{\rm c}$, to nearly rectangular between $R_{\rm c}$ and the **inner Lindblad radius** (see below), to oval between this radius and corotation.

To an observer in an inertial frame of reference, stars on orbits belonging to the long-axis family circulate about the center of the potential in the same sense as the potential rotates. One part of the circulation seen by such an observer is due to the rotation of the frame of reference in which the potential is static. A second component of circulation is due to the mean streaming motion of such stars when referred to the rotating frame of the potential. Both components of circulation diminish towards zero if the angular velocity of the potential is reduced to zero. Near corotation the dominant component arises from the rotation of the frame of reference of the potential, while at small radii the more important component is the mean streaming motion of the stars through the rotating frame of reference.

3.3.3 Weak bars

Before we leave the subject of orbits in planar non-axisymmetric potentials, we derive an analytic description of loop orbits in weak bars.

(a) Lindblad resonances We assume that the figure of the potential rotates at some steady pattern speed $\Omega_{\rm b}$, and we seek to represent a general loop orbit as a superposition of the circular motion of a guiding center and small oscillations around this guiding center. Hence our treatment of orbits in weak non-axisymmetric potentials will be closely related to the epicycle theory of nearly circular orbits in an axisymmetric potential (§3.2.3).

Let (R, φ) be polar coordinates in the frame that rotates with the potential, such that the line $\varphi = 0$ coincides with the long axis of the potential. Then the Lagrangian is

$$\mathcal{L} = \tfrac{1}{2}\dot{R}^2 + \tfrac{1}{2}[R(\dot{\varphi} + \Omega_\mathrm{b})]^2 - \Phi(R, \varphi), \tag{3.134}$$

so the equations of motion are

$$\ddot{R} = R(\dot{\varphi} + \Omega_\mathrm{b})^2 - \frac{\partial \Phi}{\partial R}, \tag{3.135a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}[R^2(\dot{\varphi} + \Omega_\mathrm{b})] = -\frac{\partial \Phi}{\partial \varphi}. \tag{3.135b}$$

Since we assume that the bar is weak, we may write

$$\Phi(R, \varphi) = \Phi_0(R) + \Phi_1(R, \varphi), \tag{3.136}$$

where $|\Phi_1/\Phi_0| \ll 1$. We divide R and φ into zeroth- and first-order parts

$$R(t) = R_0 + R_1(t) \quad ; \quad \varphi(t) = \varphi_0(t) + \varphi_1(t) \tag{3.137}$$

by substituting these expressions into equation (3.135) and requiring that the zeroth-order terms should sum to zero. Thus

$$R_0\,(\dot{\varphi}_0 + \Omega_\mathrm{b})^2 = \left(\frac{\mathrm{d}\Phi_0}{\mathrm{d}R}\right)_{R_0} \quad \text{and} \quad \dot{\varphi}_0 = \text{constant}. \tag{3.138}$$

This is the usual equation for centrifugal equilibrium at R_0. If we define $\Omega_0 \equiv \Omega(R_0)$, where

$$\Omega(R) \equiv \pm\sqrt{\frac{1}{R}\frac{\mathrm{d}\Phi_0}{\mathrm{d}R}} \tag{3.139}$$

is the circular frequency at R in the potential Φ_0, equation (3.138) for the angular speed of the guiding center (R_0, φ_0) becomes

$$\dot{\varphi}_0 = \Omega_0 - \Omega_\mathrm{b}, \tag{3.140}$$

where $\Omega_0 > 0$ for prograde orbits and $\Omega_0 < 0$ for retrograde ones. We choose the origin of time such that

$$\varphi_0(t) = (\Omega_0 - \Omega_\mathrm{b})t. \tag{3.141}$$

The first-order terms in the equations of motion (3.135) now yield

$$\ddot{R}_1 + \left(\frac{\mathrm{d}^2\Phi_0}{\mathrm{d}R^2} - \Omega^2\right)_{R_0} R_1 - 2R_0\Omega_0\dot{\varphi}_1 = -\left(\frac{\partial \Phi_1}{\partial R}\right)_{R_0}, \tag{3.142a}$$

$$\ddot{\varphi}_1 + 2\Omega_0 \frac{\dot{R}_1}{R_0} = -\frac{1}{R_0^2}\left(\frac{\partial \Phi_1}{\partial \varphi}\right)_{R_0}. \tag{3.142b}$$

To proceed further we must choose a specific form of Φ_1; we set

$$\Phi_1(R,\varphi) = \Phi_{\rm b}(R)\cos(m\varphi), \tag{3.143}$$

where m is a positive integer, since any potential that is an even function of φ can be expanded as a sum of terms of this form. In practice we are mostly concerned with the case $m = 2$ since the potential is then barred. If $\varphi = 0$ is to coincide with the long axis of the potential, we must have $\Phi_{\rm b} < 0$.

So far we have assumed only that the angular velocity $\dot{\varphi}_1$ is small, not that φ_1 is itself small. Allowing for large excursions in φ_1 will be important when we consider what happens at resonances in part (b) of this section, but for the moment we assume that $\varphi_1 \ll 1$ and hence that $\varphi(t)$ always remains close to $(\Omega_0 - \Omega_{\rm b})t$. With this assumption we may replace φ by φ_0 in the expressions for $\partial\Phi_1/\partial R$ and $\partial\Phi_1/\partial\varphi$ to yield

$$\ddot{R}_1 + \left(\frac{{\rm d}^2\Phi_0}{{\rm d}R^2} - \Omega^2\right)_{R_0} R_1 - 2R_0\Omega_0\dot{\varphi}_1 = -\left(\frac{{\rm d}\Phi_{\rm b}}{{\rm d}R}\right)_{R_0}\cos\left[m(\Omega_0 - \Omega_{\rm b})t\right], \tag{3.144a}$$

$$\ddot{\varphi}_1 + 2\Omega_0\frac{\dot{R}_1}{R_0} = \frac{m\Phi_{\rm b}(R_0)}{R_0^2}\sin\left[m(\Omega_0 - \Omega_{\rm b})t\right]. \tag{3.144b}$$

Integrating the second of these equations, we obtain

$$\dot{\varphi}_1 = -2\Omega_0\frac{R_1}{R_0} - \frac{\Phi_{\rm b}(R_0)}{R_0^2(\Omega_0 - \Omega_{\rm b})}\cos\left[m(\Omega_0 - \Omega_{\rm b})t\right] + \text{constant}. \tag{3.145}$$

We now eliminate $\dot{\varphi}_1$ from equation (3.144a) to find

$$\ddot{R}_1 + \kappa_0^2 R_1 = -\left[\frac{{\rm d}\Phi_{\rm b}}{{\rm d}R} + \frac{2\Omega\Phi_{\rm b}}{R(\Omega - \Omega_{\rm b})}\right]_{R_0}\cos\left[m(\Omega_0 - \Omega_{\rm b})t\right] + \text{constant}, \tag{3.146a}$$

where

$$\kappa_0^2 \equiv \left(\frac{{\rm d}^2\Phi_0}{{\rm d}R^2} + 3\Omega^2\right)_{R_0} = \left(R\frac{{\rm d}\Omega^2}{{\rm d}R} + 4\Omega^2\right)_{R_0} \tag{3.146b}$$

is the usual epicycle frequency (eq. 3.80). The constant in equation (3.146a) is unimportant since it can be absorbed by a shift $R_1 \rightarrow R_1 + \textit{constant}$.

Equation (3.146a) is the equation of motion of a harmonic oscillator of natural frequency κ_0 that is driven at frequency $m(\Omega_0 - \Omega_{\rm b})$. The general solution to this equation is

$$R_1(t) = C_1\cos(\kappa_0 t + \alpha) - \left[\frac{{\rm d}\Phi_{\rm b}}{{\rm d}R} + \frac{2\Omega\Phi_{\rm b}}{R(\Omega - \Omega_{\rm b})}\right]_{R_0}\frac{\cos\left[m(\Omega_0 - \Omega_{\rm b})t\right]}{\Delta}, \tag{3.147a}$$

where C_1 and α are arbitrary constants, and

$$\Delta \equiv \kappa_0^2 - m^2(\Omega_0 - \Omega_\mathrm{b})^2. \tag{3.147b}$$

If we use equation (3.141) to eliminate t from equation (3.147a), we find

$$R_1(\varphi_0) = C_1 \cos\left(\frac{\kappa_0 \varphi_0}{\Omega_0 - \Omega_\mathrm{b}} + \alpha\right) + C_2 \cos(m\varphi_0), \tag{3.148a}$$

where

$$C_2 \equiv -\frac{1}{\Delta}\left[\frac{\mathrm{d}\Phi_\mathrm{b}}{\mathrm{d}R} + \frac{2\Omega\Phi_\mathrm{b}}{R(\Omega - \Omega_\mathrm{b})}\right]_{R_0}. \tag{3.148b}$$

If $C_1 = 0$, $R_1(\varphi_0)$ becomes periodic in φ_0 with period $2\pi/m$, and thus the orbit that corresponds to $C_1 = 0$ is a closed loop orbit. The orbits with $C_1 \neq 0$ are the non-closed loop orbits that are parented by this closed loop orbit. In the following we set $C_1 = 0$ so that we may study the closed loop orbits.

The right side of equation (3.148a) for R_1 becomes singular at a number of values of R_0:

(i) **Corotation resonance**. When

$$\Omega_0 = \Omega_\mathrm{b}, \tag{3.149}$$

$\dot{\varphi}_0 = 0$, and the guiding center corotates with the potential.

(ii) **Lindblad resonances**. When

$$m(\Omega_0 - \Omega_\mathrm{b}) = \pm\kappa_0, \tag{3.150}$$

the star encounters successive crests of the potential at a frequency that coincides with the frequency of its natural radial oscillations. Radii at which such resonances occur are called **Lindblad radii** after the Swedish astronomer Bertil Lindblad (1895–1965). The plus sign in equation (3.150) corresponds to the case in which the star overtakes the potential, encountering its crests at the resonant frequency κ_0; this is called an **inner Lindblad resonance**. In the case of a minus sign, the crests of the potential sweep by the more slowly rotating star, and R_0 is said to be the radius of the **outer Lindblad resonance**.

There is a simple connection between these two types of resonance. A circular orbit has two natural frequencies. If the star is displaced radially, it oscillates at the epicycle frequency κ_0. On the other hand, if the star is displaced azimuthally in such a way that it is still on a circular orbit, then it will continue on a circular orbit displaced from the original one. Thus the star is neutrally stable to displacements of this form; in other words, its natural azimuthal frequency is zero. The two types of resonance arise when the

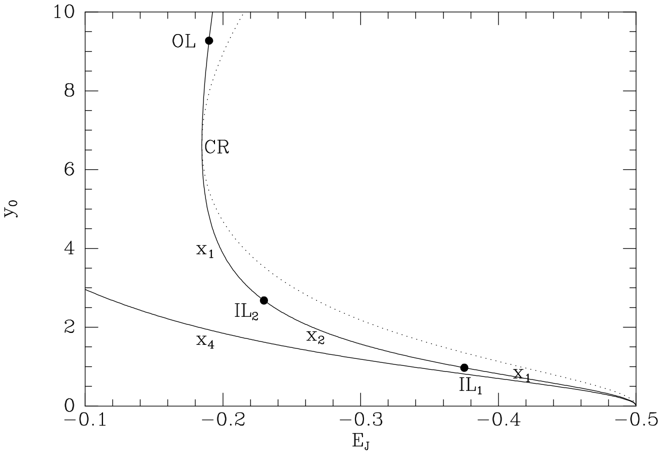


Figure 3.20 The full curves are the characteristic curves of the prograde (upper) and retrograde (lower) circular orbits in the isochrone potential (2.47) when a rotating frame of reference is employed. The dashed curve shows the relation $\Phi_{\rm eff}(0, y) = E_{\rm J}$, and the dots mark the positions of the Lindblad resonances when a small non-axisymmetric component is added to the potential.

forcing frequency seen by the star, $m(\Omega_0 - \Omega_{\rm b})$, equals one of the natural frequencies $\pm\kappa_0$ and 0.

Figure 6.11 shows plots of Ω, $\Omega + \frac{1}{2}\kappa$ and $\Omega - \frac{1}{2}\kappa$ for two circular-speed curves typical of galaxies. A galaxy may have 0, 1, 2, or more Lindblad resonances. The Lindblad and corotation resonances play a central role in the study of bars and spiral structure, and we shall encounter them again Chapter 6.

From equation (3.148a) it follows that for $m = 2$ the closed loop orbit is aligned with the bar whenever $C_2 > 0$, and is aligned perpendicular to the bar when $C_2 < 0$. When R_0 passes through a Lindblad or corotation resonance, the sign of C_2, and therefore the orientation of the closed loop orbits, changes.

It is interesting to relate the results of this analytic treatment to the orbital structure of a strong bar that we obtained numerically in the last subsection. In this connection it is helpful to compare Figure 3.18, which shows data for a barred potential, with Figure 3.20, which describes orbits in an axisymmetric potential viewed from a rotating frame. The full curves in Figure 3.20 show the relationship between the Jacobi constant $E_{\rm J}$ and the radii of prograde and retrograde circular orbits in the isochrone potential (2.47). As in Figure 3.18, the dotted curve marks the relation $\Phi_{\rm eff}(0, y) = E_{\rm J}$. There are no orbits in the region to the right of this curve, which touches the curve of the prograde circular orbits at the corotation resonance, marked CR in the figure. If in the given frame we were to add a small non-axisymmetric component to the potential, the orbits marked by large dots would lie at the Lindblad resonances (from right to left, the first and second inner Lindblad resonances and the outer Lindblad resonance marked OL). We call the radius of the first inner Lindblad resonance[11] $R_{\rm IL1}$, and similarly $R_{\rm IL2}$, $R_{\rm OL}$, and $R_{\rm CR}$ for the radii of the other Lindblad resonances and of corotation. Equations (3.148) with $C_1 = 0$ describe nearly circular orbits in a weakly barred potential. Comparing Figure 3.20 with Figure 3.18, we see that nearly circular retrograde orbits belong to the family x_4. Nearly circular prograde orbits belong to different families depending on their radius. Orbits that lie within $R_{\rm IL1}$ belong to the family x_1. In the radius range $R_{\rm IL1} < R < R_{\rm IL2}$ the families x_2 and x_3 exist and contain orbits that are more circular than those of x_1. We identify the orbits described by (3.148) with orbits of the family x_2 as indicated in Figure 3.20, since the family x_3 is unstable. In the radius range $R > R_{\rm IL2}$, equations (3.148) with $C_1 = 0$ describe orbits of the family x_1. Thus equations (3.148) describe only the families of orbits in a barred potential that are parented by a nearly circular orbit. However, when the non-axisymmetric component of the potential is very weak, most of phase space is occupied by such orbits. As the non-axisymmetry of the potential becomes stronger, families of orbits that are not described by equations (3.148) become more important.

(b) Orbits trapped at resonance When R_0 approaches the radius of either a Lindblad resonance or the corotation resonance, the value of R_1 that is predicted by equations (3.148) becomes large, and our linearized treatment of the equations of motion breaks down. However, one can modify the analysis to cope with these resonances. We now discuss the necessary modifications for the case of the corotation resonance. The case of the Lindblad resonances is described in Goldreich & Tremaine (1981).

The appropriate modification is suggested by our investigation of orbits near the Lagrange points L_4 and L_5 in the potential Φ_L (eq. 3.103), when the radius is large compared to the core radius $R_{\rm c}$ and the ellipticity $\epsilon = 1-q$ approaches zero. In this limit the non-axisymmetric part of the potential is proportional to ϵ, so we have an example of a weak bar when $\epsilon \to 0$. We found in §3.3.2 that a star's orbit was a superposition of motion at frequencies α and β around two ellipses. In the limit $\epsilon \to 0$, the β-ellipse represents the familiar epicyclic motion and will not be considered further. The α-ellipse is highly elongated in the azimuthal direction, with axis ratio $|Y_1/X_1| = \sqrt{2\epsilon}$, and its frequency is small, $\alpha = \sqrt{2\epsilon}\Omega_{\rm b}$.

These results suggest we consider the approximation in which R_1, $\dot{R}_1$, and $\dot{\varphi}_1$ are small but φ_1 is not. Specifically, if the bar strength Φ_1 is proportional to some small parameter that we may call ϵ, we assume that φ_1 is of order unity, R_1 is of order $\epsilon^{1/2}$, and the time derivative of any quantity is smaller than that quantity by of order $\epsilon^{1/2}$. Let us place the guiding

[11] Also called the **inner inner Lindblad resonance.**

Box 3.3: The donkey effect

An orbiting particle that is subject to weak tangential forces can exhibit unusual behavior. To illustrate this, suppose that the particle has mass m and is in a circular orbit of radius r, with angular speed $\dot{\phi} = \Omega(r)$ given by $r\Omega^2(r) = \mathrm{d}\Phi/\mathrm{d}r$ (eq. 3.7a). Now let us imagine that the particle experiences a small force, F, directed parallel to its velocity vector. Since the force is small, the particle remains on a circular orbit, which slowly changes in radius in response to the force. To determine the rate of change of radius, we note that the angular momentum is $L(r) = mr^2\Omega$ and the torque is $N = rF = \dot{L}$. Thus

$$\dot{r} = \frac{\mathrm{d}r}{\mathrm{d}L}\dot{L} = \frac{F/m}{2\Omega + r\,\mathrm{d}\Omega/\mathrm{d}r} = -\frac{F}{2mB}; \tag{1}$$

where $B(r) = -\Omega - \frac{1}{2}r\,\mathrm{d}\Omega/\mathrm{d}r$ is the function defined by equation (3.83). The azimuthal angle accelerates at a rate

$$\ddot{\phi} = \frac{\mathrm{d}\Omega}{\mathrm{d}r}\dot{r} = -\frac{2A\dot{r}}{r}, \tag{2}$$

where $A(r) = -\frac{1}{2}r\mathrm{d}\Omega/\mathrm{d}r$ (eq. 3.83). Combining these results,

$$r\ddot{\phi} = \frac{A}{mB}F. \tag{3}$$

This acceleration in azimuthal angle can be contrasted to the acceleration of a free particle under the same force, $\ddot{x} = F/m$. Thus the particle behaves as if it had an inertial mass mB/A, which is negative whenever

$$-2 < \frac{\mathrm{d}\ln\Omega}{\mathrm{d}\ln R} < 0. \tag{4}$$

Almost all galactic potentials satisfy this inequality. Thus the orbiting particle behaves as if it had negative inertial mass, accelerating in the opposite direction to the applied force.

There are many examples of this phenomenon in galactic dynamics, which has come to be called the **donkey effect**: to quote Lynden–Bell & Kalnajs (1972), who introduced the term, "in azimuth stars behave like donkeys, slowing down when pulled forwards and speeding up when held back."

The simplest example of the donkey effect is an Earth satellite subjected to atmospheric drag: the satellite sinks gradually into a lower orbit with a larger circular speed and shorter orbital period, so the drag force speeds up the angular passage of the satellite across the sky.

center at L_5 $[\Omega(R_0) = \Omega_\mathrm{b};\ \varphi_0 = \pi/2]$ and use equation (3.146b) to write the equations of motion (3.142) as

$$\ddot{R}_1 + \left(\kappa_0^2 - 4\Omega_0^2\right) R_1 - 2R_0\Omega_0\dot{\varphi}_1 = -\frac{\partial\Phi_1}{\partial R}, \tag{3.151a}$$

$$\ddot{\varphi}_1 + 2\Omega_0\frac{\dot{R}_1}{R_0} = -\frac{1}{R_0^2}\frac{\partial\Phi_1}{\partial\varphi}. \tag{3.151b}$$

According to our ordering, the terms on the left side of the first line are of order $\epsilon^{3/2}$, $\epsilon^{1/2}$, and $\epsilon^{1/2}$, respectively, while the term on the right side is of order ϵ. All the terms on the second line are of order ϵ. Hence we may simplify the first line by keeping only the terms of order $\epsilon^{1/2}$:

$$\left(\kappa_0^2 - 4\Omega_0^2\right) R_1 - 2R_0\Omega_0\dot{\varphi}_1 = 0. \tag{3.152}$$

Substituting equation (3.152) into equation (3.151b) to eliminate R_1, we find

$$\ddot{\varphi}_1\left(\frac{\kappa_0^2}{\kappa_0^2 - 4\Omega_0^2}\right) = -\frac{1}{R_0^2}\frac{\partial\Phi_1}{\partial\varphi}\bigg|_{(R_0,\varphi_0+\varphi_1)}. \tag{3.153}$$

Substituting from equation (3.143) for Φ_1 we obtain with $m = 2$

$$\ddot{\varphi}_1 = -\frac{2\Phi_\mathrm{b}}{R_0^2}\left(\frac{4\Omega_0^2 - \kappa_0^2}{\kappa_0^2}\right)\sin\left[2(\varphi_0 + \varphi_1)\right]. \tag{3.154}$$

By inequality (3.82) we have that $4\Omega_0^2 > \kappa_0^2$. Also we have $\Phi_\mathrm{b} < 0$ and $\varphi_0 = \pi/2$, and so equation (3.154) becomes

$$\frac{\mathrm{d}^2\psi}{\mathrm{d}t^2} = -p^2\sin\psi, \tag{3.155a}$$

where

$$\psi \equiv 2\varphi_1 \quad \text{and} \quad p^2 \equiv \frac{4}{R_0^2}\left|\Phi_\mathrm{b}(R_0)\right|\frac{4\Omega_0^2 - \kappa_0^2}{\kappa_0^2}. \tag{3.155b}$$

Equation (3.155a) is simply the equation of a pendulum. Notice that the singularity in R_1 that appeared at corotation in equations (3.148) has disappeared in this more careful analysis. Notice also the interesting fact that the stable equilibrium point of the pendulum, $\varphi_1 = 0$, is at the *maximum*, not the minimum, of the potential Φ_1 (Box 3.3). If the integral of motion

$$E_\mathrm{p} = \tfrac{1}{2}\dot{\psi}^2 - p^2\cos\psi \tag{3.156}$$

is less than p^2, the star oscillates slowly or **librates** about the Lagrange point, whereas if $E_p > p^2$, the star is not trapped by the bar but **circulates** about the center of the galaxy. For small-amplitude librations, the libration frequency is p, consistent with our assumption that the oscillation frequency is of order $\epsilon^{1/2}$ when Φ_b is of order ϵ. Large-amplitude librations of this kind may account for the rings of material often seen in barred galaxies (page 538).

We may obtain the shape of the orbit from equation (3.152) by using equation (3.156) to eliminate $\dot\varphi_1 = \frac{1}{2}\dot\psi$:

$$R_1 = -\frac{2R_0\Omega_0\dot\varphi_1}{4\Omega_0^2-\kappa_0^2} = \pm\frac{2^{1/2}R_0\Omega_0}{4\Omega_0^2-\kappa_0^2}\sqrt{E_\mathrm{p}+p^2\cos(2\varphi_1)}. \tag{3.157}$$

We leave as an exercise the demonstration that when $E_p \gg p^2$, equation (3.157) describes the same orbits as are obtained from (3.148a) with $C_1 = 0$ and $\Omega \neq \Omega_\mathrm{b}$.

The analysis of this subsection complements the analysis of motion near the Lagrange points in §3.3.2. The earlier analysis is valid for small oscillations around a Lagrange point of an arbitrary two-dimensional rotating potential, while the present analysis is valid for excursions of any amplitude in azimuth around the Lagrange points L_4 and L_5, but only if the potential is nearly axisymmetric.

3.4 Numerical orbit integration

In most stellar systems, orbits cannot be computed analytically, so effective algorithms for numerical orbit integration are among the most important tools for stellar dynamics. The orbit-integration problems we have to address vary in complexity from following a single particle in a given, smooth galactic potential, to tens of thousands of interacting stars in a globular cluster, to billions of dark-matter particles in a simulation of cosmological clustering. In each of these cases, the dynamics is that of a Hamiltonian system: with N particles there are $3N$ coordinates that form the components of a vector $\mathbf{q}(t)$, and $3N$ components of the corresponding momentum $\mathbf{p}(t)$. These vectors satisfy Hamilton's equations,

$$\dot{\mathbf{q}} = \frac{\partial H}{\partial \mathbf{p}} \quad ; \quad \dot{\mathbf{p}} = -\frac{\partial H}{\partial \mathbf{q}}, \tag{3.158}$$

which can be written as

$$\frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = \mathbf{f}(\mathbf{w},t), \tag{3.159}$$

where $\mathbf{w} \equiv (\mathbf{q},\mathbf{p})$ and $\mathbf{f} \equiv (\partial H/\partial\mathbf{p}, -\partial H/\partial\mathbf{q})$. For simplicity we shall assume in this section that the Hamiltonian has the form $H(\mathbf{q},\mathbf{p}) = \frac{1}{2}p^2+\Phi(\mathbf{q})$, although many of our results can be applied to more general Hamiltonians. Given a phase-space position $\mathbf{w}$ at time t, and a **timestep** h, we require an algorithm—an **integrator**—that generates a new position $\mathbf{w}'$ that approximates the true position at time $t' = t+h$. Formally, the problem to be solved is the same whether we are following the motion of a single star in a given potential, or the motion of 10^{10} particles under their mutual gravitational attraction.

The best integrator to use for a given problem is determined by several factors:

- How smooth is the potential? The exploration of orbits in an analytic model of a galaxy potential places fewer demands on the integrator than following orbits in an open cluster, where the stars are buffeted by close encounters with their neighbors.
- How cheaply can we evaluate the gravitational field? At one extreme, evaluating the field by direct summation in simulations of globular cluster with $\gtrsim 10^5$ particles requires $\mathrm{O}(N^2)$ operations, and thus is quite expensive compared to the $\mathrm{O}(N)$ cost of orbit integrations. At the other extreme, tree codes, spherical-harmonic expansions, or particle-mesh codes require $\mathrm{O}(N \ln N)$ operations and thus are comparable in cost to the integration. So the integrator used in an N-body simulation of a star cluster should make the best possible use of each expensive but accurate force evaluation, while in a cosmological simulation it is better to use a simple integrator and evaluate the field more frequently.
- How much memory is available? The most accurate integrators use the position and velocity of a particle at several previous timesteps to help predict its future position. When simulating a star cluster, the number of particles is small enough ($N \lesssim 10^5$) that plenty of memory should be available to store this information. In a simulation of galaxy dynamics or a cosmological simulation, however, it is important to use as many particles as possible, so memory is an important constraint. Thus for such simulations the optimal integrator predicts the future phase-space position using only the current position and gravitational field.
- How long will the integration run? The answer can range from a few crossing times for the simulation of a galaxy merger to 10^5 crossing times in the core of a globular cluster. Long integrations require that the integrator does not introduce any systematic drift in the energy or other integrals of motion.

Useful references include Press et al. (1986), Hairer, Lubich, & Wanner (2002), and Aarseth (2003).

3.4.1 Symplectic integrators

(a) Modified Euler integrator Let us replace the original Hamiltonian $H(\mathbf{q},\mathbf{p}) = \frac{1}{2}p^2 + \Phi(\mathbf{q})$ by the time-dependent Hamiltonian

$$H_h(\mathbf{q},\mathbf{p},t) = \tfrac{1}{2}p^2 + \Phi(\mathbf{q})\delta_h(t), \quad \text{where} \quad \delta_h(t) \equiv h \sum_{j=-\infty}^{\infty} \delta(t-jh) \qquad (3.160)$$

is an infinite series of delta functions (Appendix C.1). Averaged over a time interval that is long compared to h, $\langle H_h \rangle \simeq H$, so the trajectories determined by H_h should approach those determined by H as $h \to 0$.

Hamilton's equations for H_h read

$$\dot{\mathbf{q}} = \frac{\partial H_h}{\partial \mathbf{p}} = \mathbf{p} \quad ; \quad \dot{\mathbf{p}} = -\frac{\partial H_h}{\partial \mathbf{q}} = -\boldsymbol{\nabla}\Phi(\mathbf{q})\delta_h(t). \tag{3.161}$$

We now integrate these equations from $t = -\epsilon$ to $t = h - \epsilon$, where $0 < \epsilon \ll h$. Let the system have coordinates $(\mathbf{q}, \mathbf{p})$ at time $t = -\epsilon$, and first ask for its coordinates $(\overline{\mathbf{q}}, \overline{\mathbf{p}})$ at $t = +\epsilon$. During this short interval $\mathbf{q}$ changes by a negligible amount, and $\mathbf{p}$ suffers a kick governed by the second of equations (3.161). Integrating this equation from $t = -\epsilon$ to ϵ is trivial since $\mathbf{q}$ is fixed, and we find

$$\overline{\mathbf{q}} = \mathbf{q} \quad ; \quad \overline{\mathbf{p}} = \mathbf{p} - h\boldsymbol{\nabla}\Phi(\mathbf{q}); \tag{3.162a}$$

this is called a **kick step** because the momentum changes but the position does not. Next, between $t = +\epsilon$ and $t = h - \epsilon$, the value of the delta function is zero, so the system has constant momentum, and Hamilton's equations yield for the coordinates at $t = h - \epsilon$

$$\mathbf{q}' = \overline{\mathbf{q}} + h\overline{\mathbf{p}} \quad ; \quad \mathbf{p}' = \overline{\mathbf{p}}; \tag{3.162b}$$

this is called a **drift** step because the position changes but the momentum does not. Combining these results, we find that over a timestep h starting at $t = -\epsilon$ the Hamiltonian H_h generates a map $(\mathbf{q}, \mathbf{p}) \to (\mathbf{q}', \mathbf{p}')$ given by

$$\mathbf{p}' = \mathbf{p} - h\boldsymbol{\nabla}\Phi(\mathbf{q}) \quad ; \quad \mathbf{q}' = \mathbf{q} + h\mathbf{p}'. \tag{3.163a}$$

Similarly, starting at $t = +\epsilon$ yields the map

$$\mathbf{q}' = \mathbf{q} + h\mathbf{p} \quad ; \quad \mathbf{p}' = \mathbf{p} - h\boldsymbol{\nabla}\Phi(\mathbf{q}'). \tag{3.163b}$$

These maps define the "kick-drift" or "drift-kick" **modified-Euler integrator**. The performance of this integrator in a simple galactic potential is shown in Figure 3.21.

The map induced by any Hamiltonian is a canonical or symplectic map (page 803), so it can be derived from a generating function. It is simple to confirm using equations (D.93) that the generating function $S(\mathbf{q}, \mathbf{p}') = \mathbf{q}\cdot\mathbf{p}' + \frac{1}{2}hp'^2 + h\Phi(\mathbf{q})$ yields the kick-drift modified-Euler integrator (3.163a).

According to the modified-Euler integrator, the position after timestep h is

$$\mathbf{q}' = \mathbf{q} + h\mathbf{p}' = \mathbf{q} + h\mathbf{p} - h^2\boldsymbol{\nabla}\Phi(\mathbf{q}), \tag{3.164}$$

while the exact result may be written as a Taylor series,

$$\mathbf{q}' = \mathbf{q} + h\dot{\mathbf{q}}(t=0) + \tfrac{1}{2}h^2\ddot{\mathbf{q}}(t=0) + \mathrm{O}(h^3) = \mathbf{q} + h\mathbf{p} - \tfrac{1}{2}h^2\boldsymbol{\nabla}\Phi(\mathbf{q}) + \mathrm{O}(h^3). \tag{3.165}$$

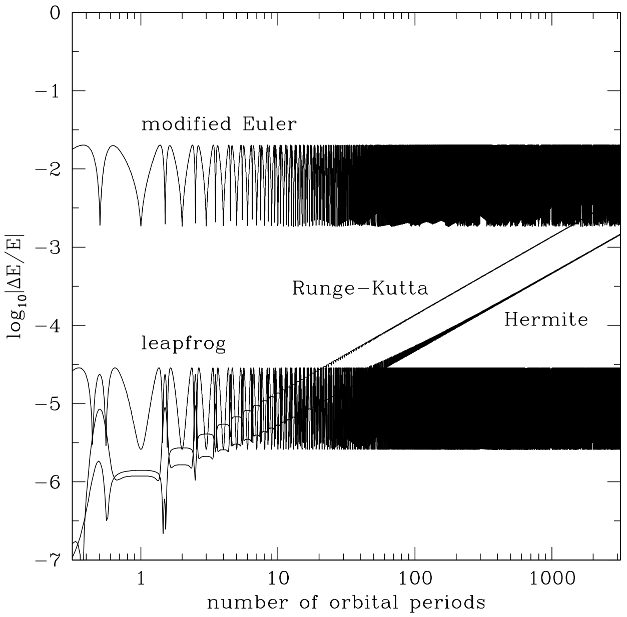


Figure 3.21 Fractional energy error as a function of time for several integrators, following a particle orbiting in the logarithmic potential $\Phi(r) = \ln r$. The orbit is moderately eccentric (apocenter twice as big as pericenter). The timesteps are fixed, and chosen so that there are 300 evaluations of the force or its derivatives per period for all of the integrators. The integrators shown are kick-drift modified-Euler (3.163a), leapfrog (3.166a), Runge–Kutta (3.168), and Hermite (3.172a–d). Note that (i) over moderate time intervals, the errors are smallest for the fourth-order integrators (Runge–Kutta and Hermite), intermediate for the second-order integrator (leapfrog), and largest for the first-order integrator (modified-Euler); (ii) the energy error of the symplectic integrators does not grow with time.

The error after a single step of the modified-Euler integrator is seen to be $\mathrm{O}(h^2)$, so it is said to be a **first-order** integrator.

Since the mappings (3.163) are derived from the Hamiltonian (3.160), they are symplectic, so either flavor of the modified-Euler integrator is a **symplectic integrator**. Symplectic integrators conserve phase-space volume and Poincaré invariants (Appendix D.4.2). Consequently, if the integrator is used to advance a series of particles that initially lie on a closed curve in the (q_i, p_i) phase plane, the curve onto which it moves the particles has the same line integral $\oint p_i \mathrm{d}q_i$ around it as the original curve. This conservation property turns out to constrain the allowed motions in phase space so strongly that the usual tendency of numerical orbit integrations to drift in energy (sometimes called **numerical dissipation**, even through the energy can either decay or grow) is absent in symplectic integrators (Hairer, Lubich, & Wanner 2002).

Leapfrog integrator By alternating kick and drift steps in more elaborate sequences, we can construct higher-order integrators (Yoshida 1993); these are automatically symplectic since they are the composition of maps (the kick and drift steps) that are symplectic. The simplest and most widely used of these is the **leapfrog** or **Verlet** integrator in which we drift for $\frac{1}{2}h$, kick for h and then drift for $\frac{1}{2}h$:

$$\mathbf{q}_{1/2} = \mathbf{q} + \tfrac{1}{2}h\mathbf{p} \ ; \ \mathbf{p}' = \mathbf{p} - h\boldsymbol{\nabla}\Phi(\mathbf{q}_{1/2}) \ ; \ \mathbf{q}' = \mathbf{q} + \tfrac{1}{2}h\mathbf{p}'. \qquad (3.166\text{a})$$

This algorithm is sometimes called "drift-kick-drift" leapfrog; an equally good form is "kick-drift-kick" leapfrog:

$$\mathbf{p}_{1/2} = \mathbf{p} - \tfrac{1}{2}h\boldsymbol{\nabla}\Phi(\mathbf{q}) \ ; \ \mathbf{q}' = \mathbf{q} + h\mathbf{p}_{1/2} \ ; \ \mathbf{p}' = \mathbf{p} - \tfrac{1}{2}h\boldsymbol{\nabla}\Phi(\mathbf{q}'). \qquad (3.166\text{b})$$

Drift-kick-drift leapfrog can also be derived by considering motion in the Hamiltonian (3.160) from $t = -\frac{1}{2}h$ to $t = \frac{1}{2}h$.

The leapfrog integrator has many appealing features: (i) In contrast to the modified-Euler integrator, it is second- rather than first-order accurate, in that the error in phase-space position after a single timestep is $\mathrm{O}(h^3)$ (Problem 3.26). (ii) Leapfrog is **time reversible** in the sense that if leapfrog advances the system from $(\mathbf{q}, \mathbf{p})$ to $(\mathbf{q}', \mathbf{p}')$ in a given time, it will also advance it from $(\mathbf{q}', -\mathbf{p}')$ to $(\mathbf{q}, -\mathbf{p})$ in the same time. Time-reversibility is a constraint on the phase-space flow that, like symplecticity, suppresses numerical dissipation, since dissipation is not a time-reversible phenomenon (Roberts & Quispel 1992; Hairer, Lubich, & Wanner 2002). (iii) A sequence of n leapfrog steps can be regarded as a drift step for $\frac{1}{2}h$, then n kick-drift steps of the modified-Euler integrator, then a drift step for $-\frac{1}{2}h$; thus if $n \gg 1$ the leapfrog integrator requires negligibly more work than the same number of steps of the modified-Euler integrator. (iv) Leapfrog also needs no storage of previous timesteps, so is economical of memory.

Because of all these advantages, most codes for simulating collisionless stellar systems use the leapfrog integrator. Time-reversible, symplectic integrators of fourth and higher orders, derived by combining multiple kick and drift steps, are described in Problem 3.27 and Yoshida (1993).

One serious limitation of symplectic integrators is that they work well only with fixed timesteps, for the following reason. Consider an integrator with fixed timestep h that maps phase-space coordinates $\mathbf{w}$ to $\mathbf{w}' = \mathbf{W}(\mathbf{w}, h)$. The integrator is symplectic if the function $\mathbf{W}$ satisfies the symplectic condition (D.78), which involves the Jacobian matrix $g_{\alpha\beta} = \partial W_\alpha / \partial w_\beta$. Now suppose that the timestep is varied, by choosing it to be some function $h(\mathbf{w})$ of location in phase space, so $\mathbf{w}' = \mathbf{W}[\mathbf{w}, h(\mathbf{w})] \equiv \widetilde{\mathbf{W}}(\mathbf{w})$. The Jacobian matrix of $\widetilde{\mathbf{W}}$ is not equal to the Jacobian matrix of $\mathbf{W}$, and in general will not satisfy the symplectic condition; in words, a symplectic integrator with fixed timestep is generally no longer symplectic once the timestep is varied.

Fortunately, the geometric constraints on phase-space flow imposed by time-reversibility are also strong, so the leapfrog integrator retains its good behavior if the timestep is adjusted in a time-reversible manner, even though the resulting integrator is no longer symplectic. Here is one way to do this: suppose that the appropriate timestep h is given by some function $\tau(\mathbf{w})$ of the phase-space coordinates. Then we modify equations (3.166a) to

$$\begin{aligned} \mathbf{q}_{1/2} = \mathbf{q} + \tfrac{1}{2}h\mathbf{p} \quad &; \quad \mathbf{p}_{1/2} = \mathbf{p} - \tfrac{1}{2}h\boldsymbol{\nabla}\Phi(\mathbf{q}_{1/2}), \\ t' &= t + \tfrac{1}{2}(h + h'), \\ \mathbf{p}' = \mathbf{p}_{1/2} - \tfrac{1}{2}h'\boldsymbol{\nabla}\Phi(\mathbf{q}_{1/2}) \quad &; \quad \mathbf{q}' = \mathbf{q}_{1/2} + \tfrac{1}{2}h'\mathbf{p}'. \end{aligned} \tag{3.167}$$

Here h' is determined from h by solving the equation $u(h, h') = \tau(\mathbf{q}_{1/2}, \mathbf{p}_{1/2})$, where $\tau(\mathbf{q}, \mathbf{p})$ is the desired timestep at $(\mathbf{q}, \mathbf{p})$ and $u(h, h')$ is any symmetric function of h and h' such that $u(h, h) = h$; for example, $u(h, h') = \frac{1}{2}(h + h')$ or $u(h, h') = 2hh'/(h + h')$.

3.4.2 Runge–Kutta and Bulirsch–Stoer integrators

To follow the motion of particles in a given smooth gravitational potential $\Phi(\mathbf{q})$ for up to a few hundred crossing times, the fourth-order Runge–Kutta integrator provides reliable transportation. The algorithm is

$$\begin{aligned} \mathbf{k}_1 = h\mathbf{f}(\mathbf{w}, t) \quad &; \quad \mathbf{k}_2 = h\mathbf{f}(\mathbf{w} + \tfrac{1}{2}\mathbf{k}_1, t + \tfrac{1}{2}h), \\ \mathbf{k}_3 = h\mathbf{f}(\mathbf{w} + \tfrac{1}{2}\mathbf{k}_2, t + \tfrac{1}{2}h) \quad &; \quad \mathbf{k}_4 = h\mathbf{f}(\mathbf{w} + \mathbf{k}_3, t + h), \\ \mathbf{w}' = \mathbf{w} + \tfrac{1}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4) \quad &; \quad t' = t + h. \end{aligned} \tag{3.168}$$

The Runge–Kutta integrator is neither symplectic nor reversible, and it requires considerably more memory than the leapfrog integrator because memory has to be allocated to $\mathbf{k}_1, \ldots, \mathbf{k}_4$. However, it is easy to use and provides fourth-order accuracy.

The **Bulirsch–Stoer** integrator is used for the same purposes as the Runge–Kutta integrator; although more complicated to code, it often surpasses the Runge–Kutta integrator in performance. The idea behind this integrator is to estimate $\mathbf{w}(t + h)$ from $\mathbf{w}(t)$ using first one step of length h, then two steps of length $h/2$, then four steps of length $h/4$, etc., up to 2^K steps of length $h/2^K$ for some predetermined number K. Then one extrapolates this sequence of results to the coordinates that would be obtained in the limit $K \to \infty$. Like the Runge–Kutta integrator, this integrator achieves speed and accuracy at the cost of the memory required to hold intermediate results. Like all high-order integrators, the Runge–Kutta and Bulirsch–Stoer integrators work best when following motion in smooth gravitational fields.

3.4.3 Multistep predictor-corrector integrators

We now discuss more complex integrators that are widely used in simulations of star clusters. We have a trajectory that has arrived at some phase-space position $\mathbf{w}_0$ at time t_0, and we wish to predict its position $\mathbf{w}_1$ at t_1. The general idea is to assume that the trajectory $\mathbf{w}(t)$ is a polynomial function of time $\mathbf{w}^{\text{poly}}(t)$, called the **interpolating polynomial**. The interpolating polynomial is determined by fitting to some combination of the present position $\mathbf{w}_0$, the past positions, $\mathbf{w}_{-1}, \mathbf{w}_{-2}, \ldots$ at times $t_{-1}, t_{-2}, \ldots$, and the present and past phase-space velocities, which are known through $\dot{\mathbf{w}}_j = \mathbf{f}(\mathbf{w}_j, t_j)$. There is no requirement that $\mathbf{f}$ is derived from Hamilton's equations, so these methods can be applied to any first-order differential equations; on the other hand they are not symplectic.

If the interpolating polynomial has order k, then the error after a small time interval h is given by the first term in the Taylor series for $\mathbf{w}(t)$ not represented in the polynomial, which is $\mathrm{O}(h^{k+1})$. Thus the order of the integrator is k.[12]

The **Adams–Bashforth** multistep integrator takes $\mathbf{w}^{\text{poly}}$ to be the unique kth-order polynomial that passes through $\mathbf{w}_0$ at t_0 and through the k points $(t_{-k+1}, \dot{\mathbf{w}}_{-k+1}), \ldots, (t_0, \dot{\mathbf{w}}_0)$.

Explicit formulae for the Adams–Bashforth integrators are easy to find by computer algebra; however, the formulae are too cumbersome to write here except in the special case of equal timesteps, $t_{j+1} - t_j = h$ for all j. Then the first few Adams–Bashforth integrators are

$$\mathbf{w}_1 = \mathbf{w}_0 + h \begin{cases} \dot{\mathbf{w}}_0 & (k=1) \\ \frac{3}{2}\dot{\mathbf{w}}_0 - \frac{1}{2}\dot{\mathbf{w}}_{-1} & (k=2) \\ \frac{23}{12}\dot{\mathbf{w}}_0 - \frac{4}{3}\dot{\mathbf{w}}_{-1} + \frac{5}{12}\dot{\mathbf{w}}_{-2} & (k=3) \\ \frac{55}{24}\dot{\mathbf{w}}_0 - \frac{59}{24}\dot{\mathbf{w}}_{-1} + \frac{37}{24}\dot{\mathbf{w}}_{-2} - \frac{3}{8}\dot{\mathbf{w}}_{-3} & (k=4). \end{cases} \tag{3.169}$$

The case $k = 1$ is called **Euler's integrator**.

The **Adams–Moulton** integrator differs from Adams–Bashforth only in that it computes the interpolating polynomial from the position $\mathbf{w}_0$ and the phase-space velocities $\dot{\mathbf{w}}_{-k+2}, \ldots, \dot{\mathbf{w}}_1$. For equal timesteps, the first few Adams–Moulton integrators are

$$\mathbf{w}_1 = \mathbf{w}_0 + h \begin{cases} \dot{\mathbf{w}}_1 & (k=1) \\ \frac{1}{2}\dot{\mathbf{w}}_1 + \frac{1}{2}\dot{\mathbf{w}}_0 & (k=2) \\ \frac{5}{12}\dot{\mathbf{w}}_1 + \frac{2}{3}\dot{\mathbf{w}}_0 - \frac{1}{12}\dot{\mathbf{w}}_{-1} & (k=3) \\ \frac{3}{8}\dot{\mathbf{w}}_1 + \frac{19}{24}\dot{\mathbf{w}}_0 - \frac{5}{24}\dot{\mathbf{w}}_{-1} + \frac{1}{24}\dot{\mathbf{w}}_{-2} & (k=4). \end{cases} \tag{3.170}$$

[12] Unfortunately, the term "order" is used both for the highest power retained in the Taylor series for $\mathbf{w}(t)$, t^k, and the dependence of the one-step error on the timestep, h^{k+1}; fortunately, both orders are the same.

Since $\dot{\mathbf{w}}_1$ is determined by the unknown phase-space position $\mathbf{w}_1$ through $\dot{\mathbf{w}}_1 = \mathbf{f}(\mathbf{w}_1, t_1)$, equations (3.170) are nonlinear equations for $\mathbf{w}_1$ that must be solved iteratively. The Adams–Moulton integrator is therefore said to be **implicit**, in contrast to Adams–Bashforth, which is **explicit**.

The strength of the Adams–Moulton integrator is that it determines $\mathbf{w}_1$ by *interpolating* the phase-space velocities, rather than by extrapolating them, as with Adams–Bashforth. This feature makes it a more reliable and stable integrator; the cost is that a nonlinear equation must be solved at every timestep.

In practice the Adams–Bashforth and Adams–Moulton integrators are used together as a **predictor-corrector** integrator. Adams–Bashforth is used to generate a preliminary value $\mathbf{w}_1$ (the prediction or P step), which is then used to generate $\dot{\mathbf{w}}_1 = \mathbf{f}(\mathbf{w}_1, t_1)$ (the evaluation or E step), which is used in the Adams–Moulton integrator (the corrector or C step). This three-step sequence is abbreviated as PEC. In principle one can then iterate the Adams–Moulton integrator to convergence through the sequence PECEC$\cdots$; however, this is not cost-effective, since the Adams–Moulton formula, even if solved exactly, is only an approximate representation of the differential equation we are trying to solve. Thus one usually stops with PEC (stop the iteration after evaluating $\mathbf{w}_1$ twice) or PECE (stop the iteration after evaluating $\dot{\mathbf{w}}_1$ twice).

When these methods are used in orbit integrations, the equations of motion usually have the form $\dot{\mathbf{x}} = \mathbf{v}$, $\dot{\mathbf{v}} = -\boldsymbol{\nabla}\Phi(\mathbf{x}, t)$. In this case it is best to apply the integrator only to the second equation, and to generate the new position $\mathbf{x}_1$ by analytically integrating the interpolating polynomial for $\mathbf{v}(t)$—this gives a formula for $\mathbf{x}_1$ that is more accurate by one power of h.

Analytic estimates (Makino 1991) suggest that the one-step error in the Adams–Bashforth–Moulton predictor-corrector integrator is smaller than the error in the Adams–Bashforth integrator by a factor of 5 for $k = 2$, 9 for $k = 3$, 13 for $k = 4$, etc. These analytic results, or the difference between the predicted and corrected values of $\mathbf{w}_1$, can be used to determine the longest timestep that is compatible with a prescribed target accuracy—see §3.4.5.

Because multistep integrators require information from the present time and $k - 1$ past times, a separate startup integrator, such as Runge–Kutta, must be used to generate the first $k - 1$ timesteps. Multistep integrators are not economical of memory because they store the coefficients of the entire interpolating polynomial rather than just the present phase-space position.

3.4.4 Multivalue integrators

By differentiating the equations of motion $\dot{\mathbf{w}} = \mathbf{f}(\mathbf{w})$ with respect to time, we obtain an expression for $\ddot{\mathbf{w}}$, which involves second derivatives of the potential, $\partial^2\Phi/\partial q_i \partial q_j$. If our Poisson solver delivers reliable values for these second derivatives, it can be advantageous to use $\ddot{\mathbf{w}}$ or even higher time derivatives of $\mathbf{w}$ to determine the interpolating polynomial $\mathbf{w}^{\rm poly}(t)$. Algorithms that employ the second and higher derivatives of $\mathbf{w}$ are called **multivalue integrators**.

In the simplest case we set $\mathbf{w}^{\rm poly}(t)$ to the kth-order polynomial that matches $\mathbf{w}$ and its first k time derivatives at t_0; this provides $k+1$ constraints for the $k+1$ polynomial coefficients and corresponds to predicting $\mathbf{w}(t)$ by its Taylor series expansion around t_0. A more satisfactory approach is to determine $\mathbf{w}^{\rm poly}(t)$ from the values taken by $\mathbf{w}$, $\dot{\mathbf{w}}$, $\ddot{\mathbf{w}}$, etc., at both t_0 and t_1. Specifically, for even k only, we make $\mathbf{w}^{\rm poly}(t)$ the kth-order polynomial that matches $\mathbf{w}$ at t_0 and its first $\frac{1}{2}k$ time derivatives at both t_0 and t_1—once again this provides $1+2\times\frac{1}{2}k=k+1$ constraints and hence determines the $k+1$ coefficients of the interpolating polynomial. The first few integrators of this type are

$$\mathbf{w}_1=\mathbf{w}_0+\begin{cases}\frac{1}{2}h(\dot{\mathbf{w}}_0+\dot{\mathbf{w}}_1) & (k=2)\\ \frac{1}{2}h(\dot{\mathbf{w}}_0+\dot{\mathbf{w}}_1)+\frac{1}{12}h^2(\ddot{\mathbf{w}}_0-\ddot{\mathbf{w}}_1) & (k=4)\\ \frac{1}{2}h(\dot{\mathbf{w}}_0+\dot{\mathbf{w}}_1)+\frac{1}{10}h^2(\ddot{\mathbf{w}}_0-\ddot{\mathbf{w}}_1) & \\ \qquad\qquad+\frac{1}{120}h^3(\dddot{\mathbf{w}}_0+\dddot{\mathbf{w}}_1) & (k=6).\end{cases} \tag{3.171}$$

Like the Adams–Moulton integrator, all of these integrators are implicit, and in fact the first of these formulae is the same as the second-order Adams–Moulton integrator in equation (3.170). Because these integrators employ information from only t_0 and t_1, there are two significant simplifications compared to multistep integrators: no separate startup procedure is needed, and the formulae look the same even if the timestep is variable.

Multivalue integrators are sometimes called **Obreshkov** (or Obrechkoff) or **Hermite** integrators, the latter name arising because they are based on Hermite interpolation, which finds a polynomial that fits specified values of a function and its derivatives (Butcher 1987).

Makino & Aarseth (1992) and Makino (2001) recommend a fourth-order multivalue predictor-corrector integrator for star-cluster simulations. Their predictor is a single-step, second-order multivalue integrator, that is, a Taylor series including terms of order h^2. Writing $\mathrm{d}\mathbf{v}/\mathrm{d}t=\mathbf{g}$, where $\mathbf{g}$ is the gravitational field, their predicted velocity is

$$\mathbf{v}_{\rm p,1}=\mathbf{v}_0+h\mathbf{g}_0+\tfrac{1}{2}h^2\dot{\mathbf{g}}_0. \tag{3.172a}$$

The predicted position is obtained by analytically integrating the interpolating polynomial for $\mathbf{v}$,

$$\mathbf{x}_{\rm p,1}=\mathbf{x}_0+h\mathbf{v}_0+\tfrac{1}{2}h^2\mathbf{g}_0+\tfrac{1}{6}h^3\dot{\mathbf{g}}_0. \tag{3.172b}$$

The predicted position and velocity are used to compute the gravitational field and its time derivative at time t_1, $\mathbf{g}_1$ and $\dot{\mathbf{g}}_1$. These are used to correct the velocity using the fourth-order formula (3.171):

$$\mathbf{v}_1=\mathbf{v}_0+\tfrac{1}{2}h(\mathbf{g}_0+\mathbf{g}_1)+\tfrac{1}{12}h^2(\dot{\mathbf{g}}_0-\dot{\mathbf{g}}_1); \tag{3.172c}$$

in words, $\mathbf{v}_1$ is determined by the fourth-order interpolating polynomial $\mathbf{v}^{\rm poly}(t)$ that satisfies the five constraints $\mathbf{v}^{\rm poly}(t_0) = \mathbf{v}_0$, $\dot{\mathbf{v}}^{\rm poly}(t_i) = \mathbf{g}_i$, $\ddot{\mathbf{v}}^{\rm poly}(t_i) = \dot{\mathbf{g}}_i$ for $i = 0, 1$.

To compute the corrected position, the most accurate procedure is to integrate analytically the interpolating polynomial for $\mathbf{v}$, which yields:

$$\mathbf{x}_1 = \mathbf{x}_0 + h\mathbf{v}_0 + \tfrac{1}{20}h^2(7\mathbf{g}_0 + 3\mathbf{g}_1) + \tfrac{1}{60}h^3(3\dot{\mathbf{g}}_0 - 2\dot{\mathbf{g}}_1). \tag{3.172d}$$

The performance of this integrator, often simply called the Hermite integrator, is illustrated in Figure 3.21.

3.4.5 Adaptive timesteps

Except for the simplest problems, any integrator should have an **adaptive timestep**, that is, an automatic procedure that continually adjusts the timestep to achieve some target level of accuracy. Choosing the right timestep is one of the most challenging tasks in designing a numerical integration scheme. Many sophisticated procedures are described in publicly available integration packages and numerical analysis textbooks. Here we outline a simple approach.

Let us assume that our goal is that the error in $\mathbf{w}$ after some short time τ should be less than $\epsilon|\mathbf{w}_0|$, where $\epsilon \ll 1$ and $\mathbf{w}_0$ is some reference phase-space position. We first move from $\mathbf{w}$ to $\mathbf{w}_2$ by taking two timesteps of length $h \ll \tau$. Then we return to $\mathbf{w}$ and take one step of length $2h$ to reach $\mathbf{w}_1$. Suppose that the correct position after an interval $2h$ is $\mathbf{w}'$, and that our integrator has order k. Then the errors in $\mathbf{w}_1$ and $\mathbf{w}_2$ may be written

$$\mathbf{w}_1 - \mathbf{w}' \simeq (2h)^{k+1}\mathbf{E} \quad ; \quad \mathbf{w}_2 - \mathbf{w}' \simeq 2h^{k+1}\mathbf{E}, \tag{3.173}$$

where $\mathbf{E}$ is an unknown error vector. Subtracting these equations to eliminate $\mathbf{w}'$, we find $\mathbf{E} \simeq (\mathbf{w}_1 - \mathbf{w}_2)/[2(2^k - 1)h^{k+1}]$. Now if we advance for a time τ, using $n \equiv \tau/h'$ timesteps of length h', the error will be

$$\mathbf{\Delta} = nh'^{k+1}\mathbf{E} = (\mathbf{w}_1 - \mathbf{w}_2)\frac{\tau h'^k}{2(2^k - 1)h^{k+1}}. \tag{3.174}$$

Our goal that $|\mathbf{\Delta}| \lesssim \epsilon|\mathbf{w}_0|$ will be satisfied if

$$h' < h_{\rm max} \equiv \left(2(2^k - 1)\frac{h}{\tau}\frac{\epsilon|\mathbf{w}_0|}{|\mathbf{w}_1 - \mathbf{w}_2|}\right)^{1/k} h. \tag{3.175}$$

If we are using a predictor-corrector scheme, a similar analysis can be used to deduce $h_{\rm max}$ from the difference of the phase-space positions returned by the predictor and the corrector, without repeating the entire predictor-corrector sequence.

3.4.6 Individual timesteps

The density in many stellar systems varies by several orders of magnitude between the center and the outer parts, and as a result the crossing time of orbits near the center is much smaller than the crossing time in the outer envelope. For example, in a typical globular cluster the crossing time at the center is $\lesssim 1\,\mathrm{Myr}$, while the crossing time near the tidal radius is $\sim 100\,\mathrm{Myr}$. Consequently, the timestep that can be safely used to integrate the orbits of stars is much smaller at the center than the edge. It is extremely inefficient to integrate *all* of the cluster stars with the shortest timestep needed for *any* star, so integrators must allow individual timesteps for each star.

If the integrator employs an interpolating polynomial, the introduction of individual timesteps is in principle fairly straightforward. To advance a given particle, one uses the most recent interpolating polynomials of all the other particles to predict their locations at whatever times the integrator requires, and then evaluates the forces between the given particle and the other particles.

This procedure makes sense if the Poisson solver uses direct summation (§2.9.1). However, with other Poisson solvers there is a much more efficient approach. Suppose, for example, that we are using a tree code (§2.9.2). Then before a single force can be evaluated, *all* particles have to be sorted into a tree. Once that has been done, it is comparatively inexpensive to evaluate large numbers of forces; hence to minimize the computational work done by the Poisson solver, it is important to evaluate the forces on many particles simultaneously. A **block timestep** scheme makes this possible whilst allowing different timesteps for different particles, by quantizing the timesteps. We now describe how one version of this scheme works with the leapfrog integrator.

We assign each particle to one of $K+1$ classes, such that particles in class k are to be advanced with timestep $h_k \equiv 2^k h$ for $k = 0, 1, 2, \ldots, K$. Thus h is the shortest timestep (class 0) and $2^K h$ is the longest (class K). The Poisson solver is used to evaluate the gravitational field at the initial time t_0, and each particle is kicked by the impulse $-\frac{1}{2}h_k\boldsymbol{\nabla}\Phi$, corresponding to the first part of the kick-drift-kick leapfrog step (3.166b). In Figure 3.22 the filled semicircles on the left edge of the diagram symbolize these kicks; they are larger at the top of the diagram to indicate that the strength of the kicks increases as 2^k. Then every particle is drifted through time h, and the Poisson solver is used only to find the forces on the particles in class 0, so these particles can be kicked by $-h\boldsymbol{\nabla}\Phi$, which is the sum of the kicks at the end of their first leapfrog step and the start of their second.

Next we drift all particles through h a second time, and use the Poisson solver to find the forces on the particles in both class 0 and class 1. The particles of class 0 are kicked by $-h\boldsymbol{\nabla}\Phi$, and the particles of class 1 are kicked by $-h_1\boldsymbol{\nabla}\Phi = -2h\boldsymbol{\nabla}\Phi$. After an interval $3h$ the particles in class 0 are kicked, after $4h$ the particles in classes 0, 1 and 2 are kicked, etc. This

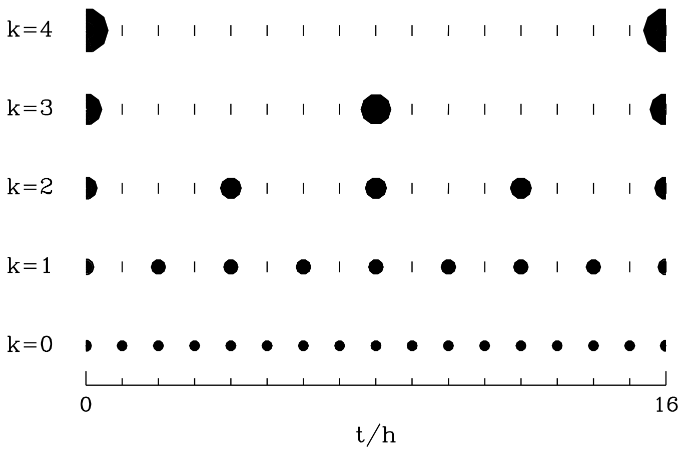


Figure 3.22 Schematic of the block timestep scheme, for a system with 5 classes of particles, having timestep h (class $k = 0$), $2h, \dots, 16h$ (class $k = K = 4$). The particles are integrated for a total time of $16h$. Each filled circle or half-circle marks the time at which particles in a given class are kicked. Each vertical bar marks a time at which particles in a class are paused in their drift step, without being kicked, in order to calculate their contribution to the kick given to particles in lower classes. The kicks at the start and end of the integration, $t = 0$ and $t = 16h$, are half as strong as the other kicks, and so are denoted by half-circles.

process continues until all particles are due for a kick, after a time $h_{\rm K} = 2^K h$. The final kick for particles in class k is $-\frac{1}{2}h_k\boldsymbol{\nabla}\Phi$, which completes 2^{K-k} leapfrog steps for each particle. At this point it is prudent to reconsider how the particles are assigned to classes in case some need smaller or larger timesteps.

A slightly different block timestep scheme works well with a particle-mesh Poisson solver (§2.9.3) when parts of the computational domain are covered by finer meshes than others, with each level of refinement being by a factor of two in the number of mesh points per unit length (Knebe, Green, & Binney 2001). Then particles are assigned timesteps according to the fineness of the mesh they are in: particles in the finest mesh have timestep $\Delta t = h$, while particles in the next coarser mesh have $\Delta t = 2h$, and so on. Particles on the finest mesh are drifted through time $\frac{1}{2}h$ before the density is determined on this mesh, and the Poisson solver is invoked to determine the forces on this mesh. Then the particles on this mesh are kicked through time h and drifted through time $\frac{1}{2}h$. Then the same drift-kick-drift sequence is used to advance particles on the next coarser mesh through time $2h$. Now these particles are ahead in time of the particles on the finest mesh. This situation is remedied by again advancing the particles on the finest mesh by h with the drift-kick-drift sequence. Once the particles on the two finest meshes have been advanced through time $2h$, we are ready to advance by $\Delta t = 4h$ the particles that are the next coarser mesh, followed by a repeat of the operations that were used to advance the particles on the two finest meshes by $2h$. The key point about this algorithm is that at each level k, particles are first advanced ahead of particles on the next coarser mesh, and then the latter particles jump ahead of the particles on level k so the next time the particles on level k are advanced, they are catching up with the particles of the coarser mesh. Errors arising from moving particles in a gravitational field from the surroundings that is out-of-date are substantially canceled by errors arising from moving particles in an ambient field that has run ahead of itself.

3.4.7 Regularization

In any simulation of a star cluster, sooner or later two particles will suffer an encounter having a very small impact parameter. In the limiting case in which the impact parameter is exactly zero (a **collision orbit**), the equation of motion for the distance r between the two particles is (eq. D.33)

$$\ddot{r} = -GM/r^2, \tag{3.176}$$

where M is the sum of the masses of the two particles. This equation is singular at $r = 0$, and a conscientious integrator will attempt to deal with the singularity by taking smaller and smaller timesteps as r diminishes, thereby bringing the entire N-body integration grinding to a halt. Even in a near-collision orbit, the integration through pericenter will be painfully slow. This problem is circumvented by transforming to a coordinate system in which the two-body problem has no singularity—this procedure is called **regularization** (Stiefel & Schiefele 1971; Mikkola 1997; Heggie & Hut 2003; Aarseth 2003). Standard integrators can then be used to solve the equations of motion in the regularized coordinates.

(a) Burdet–Heggie regularization The simplest approach to regularization is time transformation. We write the equations of motion for the two-body problem as

$$\ddot{\mathbf{r}} = -GM\frac{\mathbf{r}}{r^3} + \mathbf{g}, \tag{3.177}$$

where $\mathbf{g}$ is the gravitational field from the other $N-2$ bodies in the simulation, and change to a fictitious time τ that is defined by

$$\mathrm{d}t = r\,\mathrm{d}\tau. \tag{3.178}$$

Denoting derivatives with respect to τ by a prime we find

$$\dot{\mathbf{r}} = \frac{\mathrm{d}\tau}{\mathrm{d}t}\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}\tau} = \frac{1}{r}\mathbf{r}' \quad ; \quad \ddot{\mathbf{r}} = \frac{\mathrm{d}\tau}{\mathrm{d}t}\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{1}{r}\mathbf{r}' = \frac{1}{r^2}\mathbf{r}'' - \frac{r'}{r^3}\mathbf{r}'. \tag{3.179}$$

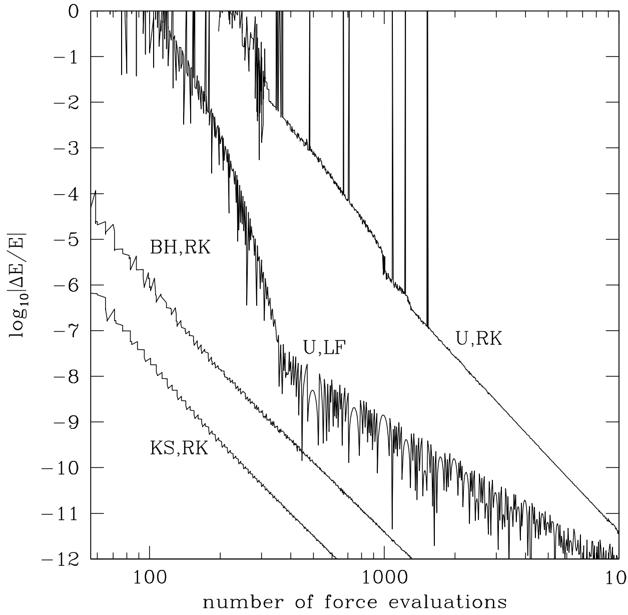


Figure 3.23 Fractional energy error from integrating one pericenter passage of a highly eccentric orbit in a Keplerian potential, as a function of the number of force evaluations. The orbit has semi-major axis $a = 1$ and eccentricity $e = 0.99$, and is integrated from $r = 1, \dot{r} < 0$ to $r = 1, \dot{r} > 0$. Curves labeled by "RK" are followed using a fourth-order Runge–Kutta integrator (3.168) with adaptive timestep control as described by Press et al. (1986). The curve labeled "U" for "unregularized" is integrated in Cartesian coordinates, the curve "BH" uses Burdet–Heggie regularization, and the curve "KS" uses Kustaanheimo–Stiefel regularization. The curve labeled "U,LF" is followed in Cartesian coordinates using a leapfrog integrator with timestep proportional to radius (eq. 3.167). The horizontal axis is the number of force evaluations used in the integration.

Substituting these results into the equation of motion, we obtain

$$\mathbf{r}'' = \frac{r'}{r}\mathbf{r}' - GM\frac{\mathbf{r}}{r} + r^2\mathbf{g}. \tag{3.180}$$

The eccentricity vector $\mathbf{e}$ (eq. 4 of Box 3.2) helps us to simplify this equation. We have

$$\begin{aligned}\mathbf{e} &= \mathbf{v} \times (\mathbf{r} \times \mathbf{v}) - GM\hat{\mathbf{e}}_r \\ &= |\mathbf{r}'|^2\frac{\mathbf{r}}{r^2} - \frac{r'}{r}\mathbf{r}' - GM\frac{\mathbf{r}}{r},\end{aligned} \tag{3.181}$$

where we have used $\mathbf{v} = \dot{\mathbf{r}} = \mathbf{r}'/r$ and the vector identity (B.9). Thus equation (3.180) can be written

$$\mathbf{r}'' = |\mathbf{r}'|^2\frac{\mathbf{r}}{r^2} - 2GM\frac{\mathbf{r}}{r} - \mathbf{e} + r^2\mathbf{g}. \tag{3.182}$$

The energy of the two-body orbit is

$$E_2 = \tfrac{1}{2}v^2 - \frac{GM}{r} = \frac{|\mathbf{r}'|^2}{2r^2} - \frac{GM}{r}, \tag{3.183}$$

so we arrive at the regularized equation of motion

$$\mathbf{r}'' - 2E_2\mathbf{r} = -\mathbf{e} + r^2\mathbf{g}, \tag{3.184}$$

in which the singularity at the origin has disappeared. This must be supplemented by equations for the rates of change of E_2, $\mathbf{e}$, and t with fictitious time τ,

$$E_2' = \mathbf{g}\cdot\mathbf{r}' \quad ; \quad \mathbf{e}' = 2\mathbf{r}(\mathbf{r}'\cdot\mathbf{g}) - \mathbf{r}'(\mathbf{r}\cdot\mathbf{g}) - \mathbf{g}(\mathbf{r}\cdot\mathbf{r}') \quad ; \quad t' = r. \tag{3.185}$$

When the external field $\mathbf{g}$ vanishes, the energy E_2 and eccentricity vector $\mathbf{e}$ are constants, the equation of motion (3.184) is that of a harmonic oscillator that is subject to a constant force $-\mathbf{e}$, and the fictitious time τ is proportional to the eccentric anomaly (Problem 3.29).

Figure 3.23 shows the fractional energy error that arises in the integration of one pericenter passage of an orbit in a Kepler potential with eccentricity $e = 0.99$. The error is plotted as a function of the number of force evaluations; this is the correct economic model if force evaluations dominate the computational cost, as is true for N-body integrations with $N \gg 1$. Note that even with $\gtrsim 1000$ force evaluations per orbit, a fourth-order Runge–Kutta integrator with adaptive timestep is sometimes unable to follow the orbit. Using the same integrator, Burdet–Heggie regularization reduces the energy error by almost five orders of magnitude.

This figure also shows the energy error that arises when integrating the same orbit using leapfrog with adaptive timestep (eq. 3.167) in unregularized coordinates. Even though leapfrog is only second-order, it achieves an accuracy that substantially exceeds that of the fourth-order Runge–Kutta integrator in unregularized coordinates, and approaches the accuracy of Burdet–Heggie regularization. Thus a time-symmetric leapfrog integrator provides much of the advantage of regularization without coordinate or time transformations.

(b) Kustaanheimo–Stiefel (KS) regularization An alternative regularization procedure, which involves the transformation of the coordinates in addition to time, can be derived using the symmetry group of the Kepler problem, the theory of quaternions and spinors, or several other methods (Stiefel & Schiefele 1971; Yoshida 1982; Heggie & Hut 2003). Once again we use the fictitious time τ defined by equation (3.178). We also define a four-vector $\mathbf{u} = (u_1, u_2, u_3, u_4)$ that is related to the position $\mathbf{r} = (x, y, z)$ by

$$\begin{aligned} u_1^2 &= \tfrac{1}{2}(x+r)\cos^2\psi \qquad & u_2 &= \frac{yu_1 + zu_4}{x+r} \\ u_4^2 &= \tfrac{1}{2}(x+r)\sin^2\psi \qquad & u_3 &= \frac{zu_1 - yu_4}{x+r}, \end{aligned} \tag{3.186}$$

where ψ is an arbitrary parameter. The inverse relations are

$$x = u_1^2 - u_2^2 - u_3^2 + u_4^2 \ ; \ y = 2(u_1u_2 - u_3u_4) \ ; \ z = 2(u_1u_3 + u_2u_4). \quad (3.187)$$

Note that $r = u_1^2 + u_2^2 + u_3^2 + u_4^2$. Let Φ_e be the potential that generates the external field $\mathbf{g} = -\boldsymbol{\nabla}\Phi_\mathrm{e}$. Then in terms of the new variables the equation of motion (3.177) reads

$$\begin{aligned} \mathbf{u}'' - \tfrac{1}{2}E\mathbf{u} &= -\tfrac{1}{4}\frac{\partial}{\partial\mathbf{u}}\left(|\mathbf{u}|^2\Phi_\mathrm{e}\right), \\ E = \tfrac{1}{2}v^2 - \frac{GM}{r} + \Phi_\mathrm{e} &= 2\frac{|\mathbf{u}'|^2}{|\mathbf{u}|^2} - \frac{GM}{|\mathbf{u}|^2} + \Phi_\mathrm{e}, \\ E' = |\mathbf{u}|^2\frac{\partial\Phi_\mathrm{e}}{\partial t} \quad &; \quad t' = |\mathbf{u}|^2, \end{aligned} \qquad (3.188)$$

When the external force vanishes, the first of equations (3.188) is the equation of motion for a four-dimensional harmonic oscillator.

Figure 3.23 shows the fractional energy error that arises in the integration of an orbit with eccentricity $e = 0.99$ using KS regularization. Using the same integrator, the energy error is more than an order of magnitude smaller than the error using Burdet–Heggie regularization.

3.5 Angle-action variables

In §3.1 we introduced the concept of an integral of motion and we saw that every spherical potential admitted at least four integrals I_i, namely, the Hamiltonian and the three components of angular momentum. Later we found that orbits in flattened axisymmetric potentials frequently admit three integrals, the classical integrals H and p_ϕ, and the non-classical third integral. Finally in §3.3 we found that many orbits in planar non-axisymmetric potentials admitted a non-classical integral in addition to the Hamiltonian.

In this section we explore the advantages of using integrals as coordinates for phase space. Since elementary Newtonian or Lagrangian mechanics restricts our choice of coordinates to ones that are rarely integrals, we work in the more general framework of Hamiltonian mechanics (Appendix D). For definiteness, we shall assume that there are three independent coordinates (so phase space is six-dimensional) and that we have three analytic isolating integrals $I_i(\mathbf{x}, \mathbf{v})$. We shall focus on a particular set of canonical coordinates, called **angle-action** variables; the three momenta are integrals, called "actions", and the conjugate coordinates are called "angles". An orbit fortunate enough to possess angle-action variables is called a **regular orbit**.

We start with a number of general results that apply to any system of angle-action variables. Then in a series of subsections we obtain explicit expressions for these variables in terms of ordinary phase-space coordinates for spherical potentials, flattened axisymmetric potentials and planar, non-axisymmetric potentials. The section ends with a description of how actions enable us to solve problems in which the gravitational potential evolves slowly.

Angle-action variables cannot be defined for many potentials of practical importance for galactic dynamics. Nonetheless, the conceptual framework of angle-action variables proves extremely useful for understanding the complex phenomena that arise in potentials that do not admit them.

The discussion below is heuristic and non-rigorous; for a precise and elegant account see Arnold (1989).

3.5.1 Orbital tori

Let us denote the angle-action variables by $(\boldsymbol{\theta}, \mathbf{J})$. We assume that the momenta $\mathbf{J} = (J_1, J_2, J_3)$ are integrals of motion. Then Hamilton's equations (D.54) for the motion of the J_i read

$$0 = \dot{J}_i = -\frac{\partial H}{\partial \theta_i}. \tag{3.189}$$

Therefore, the Hamiltonian must be independent of the coordinates $\boldsymbol{\theta}$, that is $H = H(\mathbf{J})$. Consequently, we can trivially solve Hamilton's equations for the θ_i as functions of time:

$$\dot{\theta}_i = \frac{\partial H}{\partial J_i} \equiv \Omega_i(\mathbf{J}), \quad \text{a constant} \quad \Rightarrow \quad \theta_i(t) = \theta_i(0) + \Omega_i t. \tag{3.190}$$

So everything lies at our feet if we can install three integrals of motion as the momenta of a system of canonical coordinates.[13]

We restrict our attention to bound orbits. In this case, the Cartesian coordinates x_i cannot increase without limit as the θ_i do (eq. 3.190). From this we infer that the x_i are periodic functions of the θ_i. We can scale θ_i so that $\mathbf{x}$ returns to its original value after θ_i has increased by 2π. Then we can expand $\mathbf{x}$ in a Fourier series (Appendix B.4)

$$\mathbf{x}(\boldsymbol{\theta}, \mathbf{J}) = \sum_{\mathbf{n}} \mathbf{X}_{\mathbf{n}}(\mathbf{J}) \mathrm{e}^{\mathrm{i}\mathbf{n}\cdot\boldsymbol{\theta}}, \tag{3.191}$$

where the sum is over all vectors $\mathbf{n}$ with integer components. When we eliminate the θ_i using equation (3.190), we find that the spatial coordinates are

[13] To be able to use the J_i as a set of momenta, they must satisfy the canonical commutation relations (D.71), so we require $[J_i, J_j] = 0$; functions satisfying this condition are said to be **in involution**. For example, the components of angular momenta are not in involution: $[L_x, L_y] = L_z$, etc.

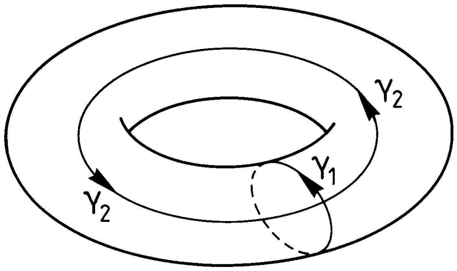


Figure 3.24 Two closed paths on a torus that cannot be deformed into one another, nor contracted to single points.

Fourier series in time, in which every frequency is a sum of integer multiples of the three **fundamental frequencies** $\Omega_i(\mathbf{J})$ that are defined by equation (3.190). Such a time series is said to be **conditionally periodic** or **quasiperiodic**.[14] For example, in spherical potentials (§3.1) the periods T_r and T_ψ are inverses of such fundamental frequencies: $T_i = 2\pi/\Omega_i$. The third fundamental frequency is zero because the orbital plane is fixed in space—see §3.5.2.

An orbit is said to be **resonant** when its fundamental frequencies satisfy a relation of the form $\mathbf{n}\cdot\mathbf{\Omega} = 0$ for some integer triple $\mathbf{n} \neq \mathbf{0}$. Usually this implies that two of the frequencies are commensurable, that is the ratio Ω_i/Ω_j is a rational number $(-n_j/n_i)$.

Consider the three-surface (i.e., volume) of fixed $\mathbf{J}$ and varying $\boldsymbol{\theta}$. This is a cube of side-length 2π, and points on opposite sides must be identified since we have seen that incrementing, say, θ_1 by 2π while leaving θ_2, θ_3 fixed brings one back to the same point in phase space. A cube with faces identified in this way is called a three-torus by analogy with the connection between a rectangle and a two-torus: if we sew together opposite edges of a rectangular sheet of rubber, we generate the doughnut-shaped inner tube of a bicycle tire.

We shall find that these three-tori are in many respects identical with orbits, so it is important to have a good scheme for labeling them. The best set of labels proves to be the Poincaré invariants (Appendix D.4.2)

$$J_i' \equiv \frac{1}{2\pi}\iint \mathrm{d}\mathbf{q}\cdot\mathrm{d}\mathbf{p} = \frac{1}{2\pi}\iint \sum_{j=1,3} \mathrm{d}q_j \mathrm{d}p_j, \tag{3.192}$$

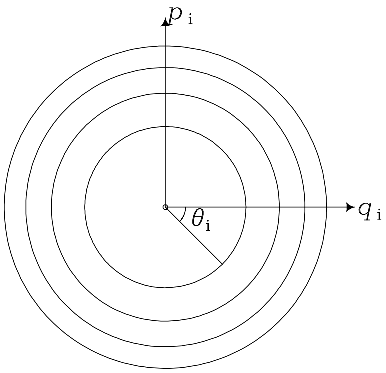
where the integral is over any surface that is bounded by the path γ_i on which θ_i increases from 0 to 2π while everything else is held constant (Figure 3.24). Since angle-action variables are canonical, $\mathrm{d}\mathbf{q}\cdot\mathrm{d}\mathbf{p} = \mathrm{d}\boldsymbol{\theta}\cdot\mathrm{d}\mathbf{J}$ (eq. D.84), so

$$J_i' = \frac{1}{2\pi}\iint_{\text{interior of } \gamma_i} \mathrm{d}\boldsymbol{\theta}\cdot\mathrm{d}\mathbf{J} = \frac{1}{2\pi}\iint_{\text{interior of } \gamma_i} \mathrm{d}\theta_i \mathrm{d}J_i. \tag{3.193}$$

[14] Observers of binary stars use the term quasiperiod more loosely. Our usage is equivalent to what is meant by a quasicrystal: a structure whose Fourier transform is discrete, but in which there are more fundamental frequencies than independent variables (in our case one, t, in a quasicrystal three x, y, z).

Box 3.4: Angle-action variables as polar coordinates

The figure shows the intersection with a coordinate plane of some of the nested orbital tori of a two-dimensional harmonic oscillator. The coordinates q_i, p_i have been scaled such that the tori appear as circles. The values of the action J_i on successive tori are chosen to be $0, 1, 2\ldots$ (in some suitable units), so, by equation (3.192), the areas inside successive tori are 0, 2π, 4π, ... Hence, the radii $r = (q_i^2 + p_i^2)^{1/2}$ of successive circles are $\sqrt{2} \times (0, 1, \sqrt{2}, \sqrt{3}, \ldots)$. In general the radius of the circle associated with the torus on which J_i takes the value J' is $r = \sqrt{2J'}$. In this plane, the angle variable θ_i is closely analogous to the usual azimuthal angle. Hence, angle-action variables are closely analogous to plane polar coordinates, the major difference being that coordinate circles are labeled not by radius but by $\sqrt{2}$ times the area they enclose. The generating function for the transformation from (θ_i, J_i) to (q_i, p_i) is given in Problem 3.31.



As Box 3.4 explains, angle-action variables are a kind of polar coordinates for phase space, and have a coordinate singularity within the domain of integration. We must exclude this from the domain of integration before we use Green's theorem to convert the surface integral in (3.193) into a line integral. The value of our surface integral is unchanged by excluding this point, but when we use Green's theorem (eq. B.61) on the original domain less the excluded point, we obtain two line integrals, one along the curve γ_i and one along the boundary that surrounds the excluded point—along this second boundary, J_i takes some definite value, J_i^c, say, and θ_i takes all values in the range $(0, 2\pi)$. Thus we have

$$J_i' = \frac{1}{2\pi}\left(\oint_{\gamma_i} J_i \mathrm{d}\theta_i - \oint_{J_i=J_i^c} J_i \mathrm{d}\theta_i\right) = J_i - J_i^c. \qquad (3.194)$$

This equation shows that the label J_i' defined by equation (3.192) will be identical with our original action coordinate J_i providing we set $J_i = 0$ at the coordinate singularity that marks the center of the angle-action coordinate system. We shall henceforth assume that this choice has been made.

In practical applications we often evaluate the integral of equation (3.192) using phase-space coordinates that have no singularity within the domain of integration. Then we can replace the surface integral with a line integral that is easier to evaluate:

$$J_i = \frac{1}{2\pi} \oint_{\gamma_i} \mathbf{p} \cdot \mathrm{d}\mathbf{q}. \tag{3.195}$$

(a) Time-averages theorem In Chapter 4 we shall make extensive use a result that we can now prove.

Time averages theorem *When a regular orbit is non-resonant, the average time that the phase point of a star on that orbit spends in any region D of its torus is proportional to the integral $V(D) = \int_D \mathrm{d}^3\boldsymbol{\theta}$.*
Proof: Let f_D be the function such that $f_D(\boldsymbol{\theta}) = 1$ when the point $\boldsymbol{\theta}$ lies in D, and is zero otherwise. We may expand f_D in a Fourier series (cf. eq. 3.191)

$$f_D(\boldsymbol{\theta}) = \sum_{\mathbf{n}=-\infty}^{\infty} F_{\mathbf{n}} \exp(\mathrm{i}\mathbf{n} \cdot \boldsymbol{\theta}). \tag{3.196}$$

Now

$$\int_{\text{torus}} \mathrm{d}^3\boldsymbol{\theta}\, f_D(\boldsymbol{\theta}) = \int_D \mathrm{d}^3\boldsymbol{\theta} = V(D). \tag{3.197a}$$

With equation (3.196) we therefore have

$$V(D) = \int_{\text{torus}} \mathrm{d}^3\boldsymbol{\theta}\, f_D(\boldsymbol{\theta}) = \sum_{\mathbf{n}=-\infty}^{\infty} F_{\mathbf{n}} \prod_{k=1}^{3} \int_0^{2\pi} \mathrm{d}\psi\, \exp(\mathrm{i}n_k\psi) = (2\pi)^3 F_{\mathbf{0}}. \tag{3.197b}$$

On the other hand, the fraction of the interval $(0, T)$ during which the star's phase point lies in D is

$$\tau_T(D) = \frac{1}{T} \int_0^T \mathrm{d}t\, f_D[\boldsymbol{\theta}(t)], \tag{3.198}$$

where $\boldsymbol{\theta}(t)$ is the position of the star's phase point at time t. With equations (3.190) and (3.196), equation (3.198) becomes

$$\begin{aligned} \tau_T(D) &= \frac{1}{T} \sum_{\mathbf{n}} \mathrm{e}^{\mathrm{i}\mathbf{n}\cdot\boldsymbol{\theta}(0)} \int_0^T \mathrm{d}t\, F_{\mathbf{n}} \mathrm{e}^{\mathrm{i}(\mathbf{n}\cdot\boldsymbol{\Omega})t} \\ &= F_{\mathbf{0}} + \frac{1}{T} \sum_{\mathbf{n}\neq 0} \mathrm{e}^{\mathrm{i}\mathbf{n}\cdot\boldsymbol{\theta}(0)} F_{\mathbf{n}} \frac{\mathrm{e}^{\mathrm{i}(\mathbf{n}\cdot\boldsymbol{\Omega})T} - 1}{\mathrm{i}\mathbf{n} \cdot \boldsymbol{\Omega}}. \end{aligned} \tag{3.199}$$

Thus

$$\lim_{T\to\infty} \tau_T(D) = F_{\mathbf{0}} = \frac{V(D)}{(2\pi)^3}, \tag{3.200}$$

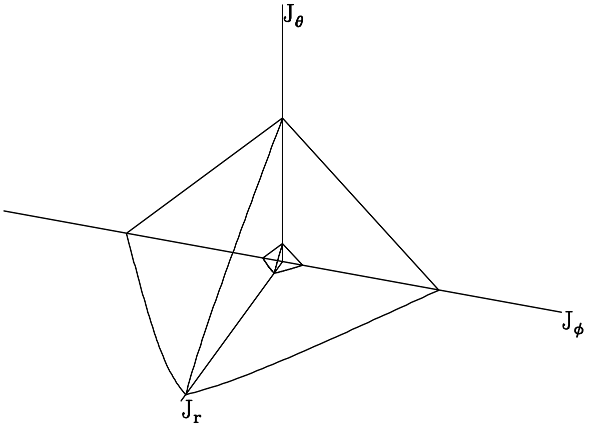


Figure 3.25 The action space of an axisymmetric potential. Two constant-energy surfaces are shown for the spherical isochrone potential (2.47). The surfaces $H = -0.5(GM/2b)$ and $H = -0.03(GM/2b)$ are shown (eq. 3.226) with the axes all scaled to length $5\sqrt{GMb}$.

which completes the proof.◁

Note that if the orbit is resonant, $\mathbf{n}\cdot\mathbf{\Omega}$ vanishes for some $\mathbf{n} \neq 0$ and the second equality in (3.199) becomes invalid, so the theorem cannot be proved. In fact, if $\Omega_i : \Omega_j = m : n$, say, then by equations (3.190) $I_4 \equiv n\theta_i - m\theta_j$ becomes an isolating integral that confines the star to a spiral on the torus. We shall see below that motion in a spherical potential provides an important example of this phenomenon.

(b) Action space In Chapter 4 we shall develop the idea that galaxies are made up of orbits, and we shall find it helpful to think of whole orbits as single points in an abstract space. Any isolating integrals can serve as coordinates for such a representation, but the most advantageous coordinates are the actions. We define **action space** to be the imaginary space whose Cartesian coordinates are the actions. Figure 3.25 shows the action space of an axisymmetric potential, when the actions can be taken to be generalizations of the actions for spherical potentials listed in Table 3.1 below. Points on the axes represent orbits for which only one of the integrals (3.192) is non-zero. These are the closed orbits. The origin represents the orbit of a star that just sits at the center of the potential. In each octant, surfaces of constant energy are approximate planes; by equation (3.190) the local normal to this surface is parallel to the vector $\mathbf{\Omega}$. Every point in the positive quadrant $J_r, J_\vartheta \geq 0$, all the way to infinity, represents a bound orbit.

A region R_3 in action space represents a group of orbits. Let the volume of R_3 be V_3. The volume of six-dimensional phase space occupied by the orbits is

$$V_6 = \int_{R_6} \mathrm{d}^3\mathbf{x}\,\mathrm{d}^3\mathbf{v}, \tag{3.201}$$

where R_6 is the region of phase space visited by stars on the orbits of R_3. Since the coordinate set $(\mathbf{J}, \boldsymbol{\theta})$ is canonical, $\mathrm{d}^3\mathbf{x}\mathrm{d}^3\mathbf{v} = \mathrm{d}^3\mathbf{J}\mathrm{d}^3\boldsymbol{\theta}$ (see eq. D.81) and thus

$$V_6 = \int_{R_6} \mathrm{d}^3\mathbf{J}\mathrm{d}^3\boldsymbol{\theta}. \tag{3.202}$$

But for any orbit the angle variables cover the range $(0, 2\pi)$, so we may immediately integrate over the angles to find

$$V_6 = (2\pi)^3 \int_{R_3} \mathrm{d}^3\mathbf{J} = (2\pi)^3 V_3. \tag{3.203}$$

Thus the volume of a region of action space is directly proportional to the volume of phase space occupied by its orbits.

(c) Hamilton–Jacobi equation The transformation between any two sets of canonical coordinates can be effected with a generating function (Appendix D.4.6). Let $S(\mathbf{q}, \mathbf{J})$ be the (unknown) generating function of the transformation between angle-action variables and ordinary phase space coordinates $(\mathbf{q}, \mathbf{p})$ such as $\mathbf{q} = \mathbf{x}$, $\mathbf{p} = \mathbf{v}$. Then (eq. D.93)

$$\boldsymbol{\theta} = \frac{\partial S}{\partial \mathbf{J}} \quad ; \quad \mathbf{p} = \frac{\partial S}{\partial \mathbf{q}}, \tag{3.204}$$

where $\mathbf{p}$ and $\boldsymbol{\theta}$ are now to be considered functions of $\mathbf{q}$ and $\mathbf{J}$. We can use $S(\mathbf{q}, \mathbf{J})$ to eliminate $\mathbf{p}$ from the usual Hamiltonian function $H(\mathbf{q}, \mathbf{p})$ and thus express H as a function

$$H\Big(\mathbf{q}, \frac{\partial S}{\partial \mathbf{q}}(\mathbf{q}, \mathbf{J})\Big).$$

of $(\mathbf{q}, \mathbf{J})$. By moving along an orbit, we can vary the q_i while holding constant the J_i. As we vary the q_i in this way, H must remain constant at the energy E of the orbit in question. This suggests that we investigate the partial differential equation

$$H\Big(\mathbf{q}, \frac{\partial S}{\partial \mathbf{q}}\Big) = E \quad \text{at fixed } \mathbf{J}. \tag{3.205}$$

If we can solve this **Hamilton–Jacobi equation**, our solution should contain some arbitrary constants K_i—we shall see below that we usually solve the equation by the method of separation of variables (e.g., §2.4) and the constants are separation constants. We identify the K_i with functions of the actions as follows. Eliminating $\mathbf{p}$ from equation (3.195) we have

$$J_i = \frac{1}{2\pi} \oint_{\gamma_i} \frac{\partial S}{\partial \mathbf{q}} \cdot \mathrm{d}\mathbf{q} = \frac{\Delta S(\mathbf{K})}{2\pi}. \tag{3.206}$$

This equation states that J_i is proportional to the increment in the generating function when one passes once around the torus on the ith path—S, like the magnetic scalar potential around a current-carrying wire, is a multivalued function. The increment in S, and therefore J_i, depends on the integration constants that appear in S, so these are functions of the actions.

Once the Hamilton–Jacobi equation has been solved and the integrals in (3.206) have been evaluated, S becomes a known function $S(\mathbf{q},\mathbf{J})$ and we can henceforth use equations (3.204) to transform between angle-action variables and ordinary phase-space coordinates. In particular, we can integrate orbits trivially by transforming the initial conditions into angle-action variables, incrementing the angles, and transforming back to ordinary phase-space coordinates.

Let us see how this process works in a simple example. The Hamiltonian of a two-dimensional harmonic oscillator is

$$H(\mathbf{x},\mathbf{p}) = \tfrac{1}{2}(p_x^2 + p_y^2 + \omega_x^2 x^2 + \omega_y^2 y^2). \tag{3.207}$$

Substituting in $p_x = \partial S/\partial x$, $p_y = \partial S/\partial y$ (eq. 3.204), the Hamilton–Jacobi equation reads

$$\left(\frac{\partial S}{\partial x}\right)^2 + \left(\frac{\partial S}{\partial y}\right)^2 + \omega_x^2 x^2 + \omega_y^2 y^2 = 2E, \tag{3.208}$$

where S is a function of x, y and $\mathbf{J}$. We solve this partial differential equation by the method of separation of variables.[15] We write $S(x,y,\mathbf{J}) = S_x(x,\mathbf{J}) + S_y(y,\mathbf{J})$ and rearrange the equation to

$$\left(\frac{\partial S_x}{\partial x}\right)^2 + \omega_x^2 x^2 = 2E - \left(\frac{\partial S_y}{\partial y}\right)^2 - \omega_y^2 y^2. \tag{3.209}$$

The left side does not depend on y and the right side does not depend on x. Consequently, each side can only be a function of $\mathbf{J}$, which we call $K^2(\mathbf{J})$ because it is evidently non-negative:

$$K^2 \equiv \left(\frac{\partial S_x}{\partial x}\right)^2 + \omega_x^2 x^2. \tag{3.210}$$

It follows that

$$S_x(x,\mathbf{J}) = K \int^x \mathrm{d}x'\, \epsilon \sqrt{1 - \frac{\omega_x^2 x'^2}{K^2}},$$

[15] When this method is applied in quantum mechanics and in potential theory (e.g., §2.4) one usually assumes that the dependent variable is a *product* of functions of one variable, rather than a sum of such functions as here.

where ϵ is chosen to be ± 1 so that $S_x(x,\mathbf{J})$ increases continuously along a path over the orbital torus. Changing the variable of integration, we have

$$\begin{aligned} S_x(x,\mathbf{J}) &= \frac{K^2}{\omega_x}\int \mathrm{d}\psi'\,\sin^2\psi' \quad \text{where} \quad x = -\frac{K}{\omega_x}\cos\psi \\ &= \frac{K^2}{2\omega_x}\bigl(\psi - \tfrac{1}{2}\sin 2\psi\bigr). \end{aligned} \tag{3.211}$$

Moreover, $p_x = \partial S/\partial x = \epsilon K\sqrt{1-\omega_x^2x^2/K^2} = K\sin\psi$, so both x and p_x return to their old values when ψ is incremented by 2π. We infer that incrementing ψ by 2π carries us around the path γ_x that is associated with J_x through equation (3.192). Thus equation (3.206) now yields

$$J_x = \frac{\Delta S}{2\pi} = \frac{\Delta S_x}{2\pi}. \tag{3.212}$$

Equation (3.211) tells us that when ψ is incremented by 2π, S_x increases by $K^2\pi/\omega_x$. Hence,

$$J_x(x,p_x) = \frac{K^2}{2\omega_x} = \frac{p_x^2+\omega_x^2x^2}{2\omega_x}, \tag{3.213}$$

where the last equality follows from (3.210) with $\partial S_x/\partial x$ replaced by p_x. The solution for $J_y(y,p_y)$ proceeds analogously and yields

$$J_y(y,p_y) = \frac{2E-K^2}{2\omega_y} = \frac{p_y^2+\omega_y^2y^2}{2\omega_y}. \tag{3.214}$$

Comparing with equation (3.207), we find that

$$H(\mathbf{J}) = \omega_xJ_x + \omega_yJ_y. \tag{3.215}$$

Notice from (3.215) that $\Omega_x \equiv \partial H/\partial J_x = \omega_x$ and similarly for Ω_y.

Finally we determine the angle variables from the second of equations (3.204). The obvious procedure is to eliminate both K and ψ from equation (3.211) in favor of J_x and x. In practice it is expedient to leave ψ in and treat it as a function of J_x and x:

$$S_x(x,\mathbf{J}) = J_x(\psi - \tfrac{1}{2}\sin 2\psi) \quad \text{where} \quad \cos\psi = -\sqrt{\frac{\omega_x}{2J_x}}x, \tag{3.216}$$

so

$$\begin{aligned} \theta_x = \frac{\partial S}{\partial J_x} &= \psi - \tfrac{1}{2}\sin 2\psi + J_x(1-\cos 2\psi)\frac{\partial\psi}{\partial J_x} \\ &= \psi - \tfrac{1}{2}\sin 2\psi + \sin^2\psi\cot\psi \\ &= \psi. \end{aligned} \tag{3.217}$$

Thus the variable ψ that we introduced for convenience in doing an integral is, in fact, the angle variable conjugate to J_x. Problem 3.33 explains an alternative, and sometimes simpler, route to the angle variables.

3.5.2 Angle-action variables for spherical potentials

We now derive angle-action variables for any spherical potential. These are useful not only for strictly spherical systems, but also for axisymmetric disks, and serve as the starting point for perturbative analyses of mildly aspherical potentials. To minimize confusion between ordinary spherical polar coordinates and angle variables, in this section we reserve ϑ for the usual polar angle, and continue to use θ_i for the variable conjugate to J_i.

The Hamilton–Jacobi equation (3.205) for the potential $\Phi(r)$ is

$$\begin{aligned} E &= \tfrac{1}{2}|\boldsymbol{\nabla} S|^2 + \Phi(r) \\ &= \tfrac{1}{2}\left[\left(\frac{\partial S}{\partial r}\right)^2 + \left(\frac{1}{r}\frac{\partial S}{\partial \vartheta}\right)^2 + \left(\frac{1}{r\sin\vartheta}\frac{\partial S}{\partial \phi}\right)^2\right] + \Phi(r), \end{aligned} \tag{3.218}$$

where we have used equation (B.38) for the gradient operator in spherical polar coordinates. We write the generating function as $S(\mathbf{x},\mathbf{J}) = S_r(r,\mathbf{J}) + S_\vartheta(\vartheta,\mathbf{J}) + S_\phi(\phi,\mathbf{J})$ and solve (3.218) by separation of variables. With the help of equation (3.204) we find

$$L_z^2 = \left(\frac{\partial S_\phi}{\partial \phi}\right)^2 = p_\phi^2, \tag{3.219a}$$

$$L^2 - \frac{L_z^2}{\sin^2\vartheta} = \left(\frac{\partial S_\vartheta}{\partial \vartheta}\right)^2 = p_\vartheta^2, \tag{3.219b}$$

$$2E - 2\Phi(r) - \frac{L^2}{r^2} = \left(\frac{\partial S_r}{\partial r}\right)^2 = p_r^2. \tag{3.219c}$$

Here we have introduced two separation constants, L and L_z. We assume that $L > 0$ and choose the sign of L_z so that $L_z = p_\phi$; with these conventions L and L_z prove to be the magnitude and z-component of the angular-momentum vector (Problem 3.20). Taking the square root of each equation and integrating, we obtain a formula for S:

$$\begin{aligned} S(\mathbf{x},\mathbf{J}) = \int_0^\phi \mathrm{d}\phi\, L_z &+ \int_{\pi/2}^\vartheta \mathrm{d}\vartheta\, \epsilon_\vartheta \sqrt{L^2 - \frac{L_z^2}{\sin^2\vartheta}} \\ &+ \int_{r_{\min}}^r \mathrm{d}r\, \epsilon_r \sqrt{2E - 2\Phi(r) - \frac{L^2}{r^2}}, \end{aligned} \tag{3.220}$$

where ϵ_ϑ and ϵ_r are chosen to be ± 1 such that the integrals in which they appear increase monotonically along a path over the orbital torus. The lower limits of these integrals specify some point on the orbital torus, and are arbitrary. It is convenient to take $r_{\min}$ to be the orbit's pericenter radius.

To obtain the actions from equation (3.206) we have to evaluate the change in S as we go round the orbital torus along curves on which only one of the coordinates is incremented. The case of changing ϕ is easy: on the relevant curve, ϕ increases by 2π, so (3.220) states that $\Delta S = 2\pi L_z$ and

$$J_\phi = L_z. \tag{3.221}$$

We call J_ϕ the **azimuthal action**. Consider next the case of changing ϑ. Let $\vartheta_{\rm min}$ be the smallest value that ϑ attains on the orbit, given by

$$\sin\vartheta_{\rm min} = \frac{|L_z|}{L}, \tag{3.222}$$

$\vartheta_{\rm min} \le \pi/2$. Then we start at $\pi/2$, where the integrand peaks, and integrate to $\pi - \vartheta_{\rm min}$, where it vanishes. We have now integrated over a quarter period of the integrand, so the whole integral is four times the value from this leg,

$$J_\vartheta = \frac{2}{\pi}\int_{\pi/2}^{\pi-\vartheta_{\rm min}} {\rm d}\vartheta\,\sqrt{L^2 - \frac{L_z^2}{\sin^2\vartheta}} = L - |L_z|. \tag{3.223}$$

We call J_ϑ the **latitudinal action**. The evaluation of J_r from equations (3.206) and (3.220) proceeds similarly and yields

$$J_r = \frac{1}{\pi}\int_{r_{\rm min}}^{r_{\rm max}} {\rm d}r\,\sqrt{2E - 2\Phi(r) - \frac{L^2}{r^2}}, \tag{3.224}$$

where $r_{\rm max}$ is the radius of the apocenter—$r_{\rm min}$ and $r_{\rm max}$ are the two roots of the radical—and J_r is the **radial action**.

An important example is that of the isochrone potential (2.47), which encompasses both the Kepler and spherical harmonic potentials as limiting cases. One finds that (Problem 3.41)

$$J_r = \frac{GM}{\sqrt{-2E}} - \tfrac{1}{2}\left(L + \tfrac{1}{2}\sqrt{L^2 - 4GMb}\right). \tag{3.225}$$

If we rewrite this expression as an equation for the Hamiltonian $H_{\rm I} = E$ as a function of the actions, we obtain

$$H_{\rm I}(\mathbf{J}) = -\frac{(GM)^2}{2\left[J_r + \frac{1}{2}\left(L + \sqrt{L^2 + 4GMb}\right)\right]^2} \qquad (L = J_\theta + |J_\phi|). \tag{3.226a}$$

Differentiating this expression with respect to the actions, we find the frequencies (eq. 3.190):

$$\begin{aligned} \Omega_r &= \frac{(GM)^2}{\left[J_r + \frac{1}{2}(L + \sqrt{L^2 + 4GMb})\right]^3} \\ \Omega_\vartheta &= \tfrac{1}{2}\left(1 + \frac{L}{\sqrt{L^2 + 4GMb}}\right)\Omega_r \quad ; \quad \Omega_\phi = {\rm sgn}(J_\phi)\Omega_\vartheta. \end{aligned} \tag{3.226b}$$

It is straightforward to check that these results are consistent with the radial and azimuthal periods determined in §3.1c.[16] In the limit $b \to 0$, the isochrone potential becomes the Kepler potential and all three frequencies become equal. The corresponding results for the spherical harmonic oscillator are obtained by examining the limit $b \to \infty$ (Problem 3.36).

J_θ and J_ϕ occur in equations (3.226) only in the combination $L = J_\theta + |J_\phi|$, and in fact, the Hamiltonian for any spherical potential is a function $H(J_r, L)$. Therefore we elevate L to the status of an action by making the canonical transformation that is defined by the generating function (eq. D.93)

$$S' = \theta_\phi J_1 + \theta_\vartheta (J_2 - |J_1|) + \theta_r J_3, \tag{3.227}$$

where (J_1, J_2, J_3) are new angle-action coordinates. Differentiating with respect to the old angles, we discover the connection between the new and old actions:

$$\begin{aligned}
J_\phi &= \frac{\partial S'}{\partial \theta_\phi} = J_1 \quad \Rightarrow \quad J_1 = L_z, \\
J_\vartheta &= \frac{\partial S'}{\partial \theta_\vartheta} = J_2 - |J_1| \quad \Rightarrow \quad J_2 = J_\vartheta + |J_\phi| = L, \\
J_r &= \frac{\partial S'}{\partial \theta_r} = J_3.
\end{aligned} \tag{3.228}$$

Thus the new action J_2 is L as desired. Differentiating S' with respect to the new actions we find that the new angles are

$$\theta_1 = \theta_\phi - \mathrm{sgn}(J_1)\theta_\vartheta \quad ; \quad \theta_2 = \theta_\vartheta \quad ; \quad \theta_3 = \theta_r. \tag{3.229}$$

Equation (3.224) can be regarded as an implicit equation for the Hamiltonian $H(\mathbf{J}) = E$ in terms of $J_3 = J_r$ and $J_2 = L$. Since J_1 does not appear in this equation, the Hamiltonian of *any* spherical potential must be of the form $H(J_2, J_3)$. Thus $\Omega_1 = \partial H/\partial J_1 = 0$ for all spherical potentials, and the corresponding angle θ_1 is an integral of motion. In §3.1 we saw that any spherical potential admits four isolating integrals. Here we have recovered this result from a different point of view: three of the integrals are the actions (J_1, J_2, J_3), and the fourth is the angle θ_1.

From Figure 3.26 we see that for orbits with $L_z > 0$ the inclination of the orbital plane $i = \frac{1}{2}\pi - \vartheta_{\min}$, while when $L_z < 0$, $i = \frac{1}{2}\pi + \vartheta_{\min}$. Combining these equations with (3.222) we find that

$$i = \cos^{-1}(L_z/L) = \cos^{-1}(J_1/J_2). \tag{3.230}$$

[16] A minor difference is that in the analysis of §3.1c, the angular momentum L could have either sign. Here $L = |\mathbf{L}|$ is always non-negative, while L_z can have either sign.

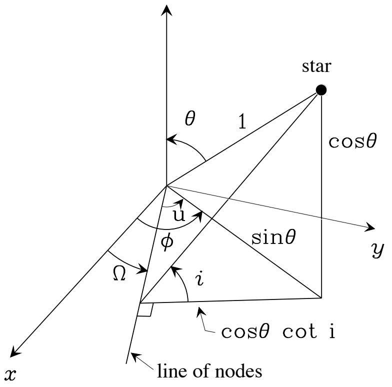


Figure 3.26 Angles defined by an orbit. The orbit is confined to a plane whose normal makes an angle i, the **inclination**, with the z axis. The orbital plane intersects the xy plane along the **line of nodes**. The ascending node is the node at which $\dot z > 0$, and the angle Ω is the longitude of the ascending node. Elementary trigonometry shows that $u = \sin^{-1}(\cot i \cot \vartheta) = \phi - \Omega$ and that $\cos\vartheta = \sin i \sin\psi$, were ψ is the angle between the line of nodes and the radius vector to the star.

We now obtain explicit expressions for the angle variables of any spherical potential by evaluating $\partial S/\partial J_i = \theta_i$, where S is derived from equation (3.220) by replacing E with $H(J_2, J_3)$, L with J_2, and L_z with J_1. We have

$$S = \phi J_1 + \int_{\pi/2}^{\vartheta} \mathrm{d}\vartheta\, \epsilon_\vartheta \sqrt{J_2^2 - \frac{J_1^2}{\sin^2\vartheta}} + \int_{r_{\min}}^{r} \mathrm{d}r\, \epsilon_r \sqrt{2H(J_2, J_3) - 2\Phi(r) - \frac{J_2^2}{r^2}}. \tag{3.231}$$

Figure 3.26 helps us to interpret our final result. It depicts the star after it has passed the line of nodes, moving upward. At this instant, $\dot\vartheta < 0$, and we must choose $\epsilon_\vartheta = -1$ to make the first integral of equation (3.231) increasing. We therefore specialize to this case, and using (3.230) find

$$\begin{aligned}\theta_1 = \frac{\partial S}{\partial J_1} &= \phi + \mathrm{sgn}(J_1)\int_{\pi/2}^{\vartheta} \frac{\mathrm{d}\vartheta}{\sin\vartheta\sqrt{\sin^2\vartheta\sec^2 i - 1}} \\ &= \phi - u,\end{aligned} \tag{3.232a}$$

where[17]

$$\sin u \equiv \cot i \cot\vartheta. \tag{3.232b}$$

Figure 3.26 demonstrates that the new variable u is actually $\phi - \Omega$ and thus that $\theta_1 = \Omega$, the **longitude of the ascending node**.[18] Thus θ_1 is constant because the line of nodes is fixed. If the potential were not spherical, but

[17] This follows because

$$\begin{aligned}\mathrm{d}[\sin^{-1}(\cot i \cot\vartheta)] &= -\csc^2\vartheta \cot i\, \mathrm{d}\vartheta/\sqrt{1 - \cot^2 i \cot^2\vartheta} \\ &= \mathrm{sgn}(\cos i)/(\sin\vartheta\sqrt{\sin^2\vartheta\sec^2 i - 1}).\end{aligned} \tag{3.233}$$

[18] Equation (3.232b) has two solutions in $(0, 2\pi)$ and care must be taken to choose the correct solution.

Table 3.1 Angle-action variables in a spherical potential

actions	$J_\phi = L_z$	$J_\vartheta = L - \lvert L_z\rvert$	J_r
angles	$\theta_\phi = \Omega + \mathrm{sgn}(L_z)\theta_\vartheta$	θ_ϑ	θ_r
Hamiltonian	$H(J_\vartheta + \lvert J_\phi\rvert, J_r)$		
frequencies	$\Omega_\phi = \mathrm{sgn}(L_z)\Omega_\vartheta$	Ω_ϑ	Ω_r
actions	$J_1 = L_z$	$J_2 = L$	$J_3 = J_r$
angles	$\theta_1 = \Omega$	$\theta_2 = \theta_\vartheta$	$\theta_3 = \theta_r$
Hamiltonian	$H(J_2, J_3)$		
frequencies	$\Omega_1 = 0$	$\Omega_2 = \Omega_\vartheta$	$\Omega_3 = \Omega_r$
actions	$J_a = L_z$	$J_b = L$	$J_c = J_r + L$
angles	$\theta_a = \Omega$	$\theta_b = \theta_\vartheta - \theta_r$	$\theta_c = \theta_r$
Hamiltonian	$H(J_b, J_c - J_b)$		
frequencies	$\Omega_a = 0$	$\Omega_b = \Omega_\vartheta - \Omega_r$	$\Omega_c = \Omega_r$

NOTES: The Delaunay variables (J_a, J_b, J_c) are defined in Appendix E. When possible, actions and angles are expressed in terms of the total angular momentum L, the z-component of angular momentum L_z, the radial action J_r, and the longitude of the ascending node Ω (Figure 3.26). Unfortunately, Ω is also used for the frequency corresponding to a given action, but in this case it is always accompanied by a subscript. The Hamiltonian is $H(L, J_r)$.

merely axisymmetric, θ_1 would not be constant and the orbital plane would precess.

Next we differentiate equation (3.231) to obtain θ_3. Only the third term, which is equal to S_r, depends on J_3. Thus we have

$$\theta_3 = \left(\frac{\partial S_r}{\partial J_3}\right)_{J_2} = \left(\frac{\partial S_r}{\partial H}\right)_{J_2}\left(\frac{\partial H}{\partial J_3}\right)_{J_2} = \left(\frac{\partial S_r}{\partial H}\right)_{J_2}\Omega_3, \tag{3.234}$$

where the last step follows from equation (3.190). Similarly,

$$\theta_2 = \left(\frac{\partial S}{\partial J_2}\right)_{J_3} = \left(\frac{\partial S_\vartheta}{\partial J_2}\right)_{J_3} + \left(\frac{\partial S_r}{\partial H}\right)_{J_2}\left(\frac{\partial H}{\partial J_2}\right)_{J_3} + \left(\frac{\partial S_r}{\partial J_2}\right)_{H}. \tag{3.235}$$

We eliminate $\partial S_r/\partial H$ using equation (3.234),

$$\theta_2 = \left(\frac{\partial S_\vartheta}{\partial J_2}\right)_{J_3} + \frac{\Omega_2}{\Omega_3}\theta_3 + \left(\frac{\partial S_r}{\partial J_2}\right)_{H}. \tag{3.236}$$

From equation (3.231) with $\epsilon_\vartheta = -1$, it is straightforward to show that

$$\left(\frac{\partial S_\vartheta}{\partial J_2}\right)_{J_1} = \sin^{-1}\left(\frac{\cos\vartheta}{\sin i}\right). \tag{3.237}$$

Now let ψ be the angle measured in the orbital plane from the line of nodes to the current position of the star. From Figure 3.26 it is easy to see that $\cos\vartheta = \sin i \sin\psi$; thus

$$\left(\frac{\partial S_\vartheta}{\partial J_2}\right)_{J_1} = \psi. \tag{3.238}$$

The other two partial derivatives in equations (3.234) and (3.235) can only be evaluated once $\Phi(r)$ has been chosen. In the case of the isochrone potential (2.47), we have

$$\left(\frac{\partial S_r}{\partial H}\right)_{J_2} = \int_{r_{\min}}^{r} \mathrm{d}I \quad ; \quad \left(\frac{\partial S_r}{\partial J_2}\right)_H = -J_2 \int_{r_{\min}}^{r} \frac{\mathrm{d}I}{r^2} \tag{3.239}$$

where $\mathrm{d}I$ is defined by (3.36). Hence the integrals to be performed are just indefinite versions of the definite integrals that yielded T_r and $\Delta\psi$ in §3.1c. The final answers are most conveniently expressed in terms of an auxiliary variable η that is defined by (cf. eqs. 3.28, 3.32 and 3.34)

$$s = 2 + \frac{c}{b}(1 - e\cos\eta) \quad \text{where} \quad \begin{cases} c \equiv \dfrac{GM}{-2H} - b, \\ e^2 \equiv 1 - \dfrac{J_2^2}{GMc}\Big(1 + \dfrac{b}{c}\Big), \\ s \equiv 1 + \sqrt{1 + r^2/b^2}. \end{cases} \tag{3.240}$$

Then one has[19]

$$\begin{aligned} \theta_3 &= \eta - \frac{ec}{c+b}\sin\eta \\ \theta_2 &= \psi + \frac{\Omega_2}{\Omega_3}\theta_3 - \tan^{-1}\left(\sqrt{\frac{1+e}{1-e}}\tan(\tfrac{1}{2}\eta)\right) \\ &\qquad - \frac{1}{\sqrt{1 + 4GMb/J_2^2}}\tan^{-1}\left(\sqrt{\frac{1+e+2b/c}{1-e+2b/c}}\tan(\tfrac{1}{2}\eta)\right). \end{aligned} \tag{3.241}$$

Thus in the case of the isochrone potential we can analytically evaluate all three angle variables from ordinary phase-space coordinates $(\mathbf{x}, \mathbf{v})$.

To summarize, in an arbitrary spherical potential two of the actions can be taken to be the total angular momentum L and its z-component L_z, and one angle can be taken to be the longitude of the ascending node Ω. The remaining action and angles can easily be determined by numerical evaluation of the integral (3.224) and integrals analogous to those of equation (3.239). In the isochrone potential, all angle-action variables can be obtained analytically from the ordinary phase-space coordinates $(\mathbf{x}, \mathbf{v})$. The analytic relations among angle-action variables in spherical potentials are summarized in Table 3.1.

[19] In numerical work, care must be taken to ensure that the branch of the inverse trigonometric functions is chosen so that the angle variables increase continuously.

3.5.3 Angle-action variables for flattened axisymmetric potentials

In §3.2 we used numerical integrations to show that most orbits in flattened axisymmetric potentials admit three isolating integrals, only two of which were identified analytically. Now we take up the challenge of identifying the missing "third integral" analytically, and deriving angle-action variables from it and the classical integrals. It proves possible to do this only for special potentials, and we start by examining the potentials for which we *have* obtained action integrals for clues as to what a promising potential might be.

(a) Stäckel potentials In §3.3 we remarked that box orbits in a planar non-rotating bar potential resemble Lissajous figures generated by two-dimensional harmonic motion, while loop orbits have many features in common with orbits in axisymmetric potentials. Let us examine these parallels more closely. The orbits of a two-dimensional harmonic oscillator admit two independent isolating integrals, $H_x = \frac{1}{2}(p_x^2 + \omega_x^2 x^2)$ and $H_y = \frac{1}{2}(p_y^2 + \omega_y^2 y^2)$ (eq. 3.207). At each point in the portion of the (x, y) plane visited by the orbit, the particle can have one of four momentum vectors. These momenta arise from the ambiguity in the signs of p_x and p_y when we are given only E_x and E_y, the values of H_x and H_y: $p_x(x) = \pm\sqrt{2E_x - \omega_x^2 x^2}$; $p_y(y) = \pm\sqrt{2E_y - \omega_y^2 y^2}$. The boundaries of the orbit are the lines on which $p_x = 0$ or $p_y = 0$.

Consider now planar orbits in a axisymmetric potential $\Phi(r)$. These orbits fill annuli. At each point in the allowed annulus two momenta are possible: $p_r(r) = \pm\sqrt{2(E - \Phi) - L_z^2/r^2}$, $p_\phi = L_z$. The boundaries of the orbit are the curves on which $p_r = 0$.

These examples have a number of important points in common:

(i) The boundaries of orbits are found by equating to zero one canonical momentum in a coordinate system that reflects the symmetry of the potential.

(ii) The momenta in this privileged coordinate system can be written as functions of only one variable: $p_x(x)$ and $p_y(y)$ in the case of the harmonic oscillator; and $p_r(r)$ and $p_\phi = L_z$ (which depends on neither coordinate) in the case of motion in a axisymmetric potential.

(iii) These expressions for the momenta are found by splitting the Hamilton–Jacobi equation $H - E = 0$ (eq. 3.205) into two parts, each of which is a function of only one coordinate and its conjugate momentum. In the case of the harmonic oscillator, $0 = H - E = \frac{1}{2}|\mathbf{p}|^2 + \frac{1}{2}(\omega_x^2 x^2 + \omega_y^2 y^2) - E = H_x(p_x, x) + H_y(p_y, y) - E$. In the case of motion in an axisymmetric potential, $0 = r^2(H - E) = r^2\left[\frac{1}{2}p_r^2 + \Phi(r)\right] - r^2 E + \frac{1}{2}p_\phi^2$.

The first of these observations suggests that we look for a curvilinear coordinate system whose coordinate curves run parallel to the edges of numerically integrated orbits, such as those plotted in Figure 3.4. Figure 3.27 shows that

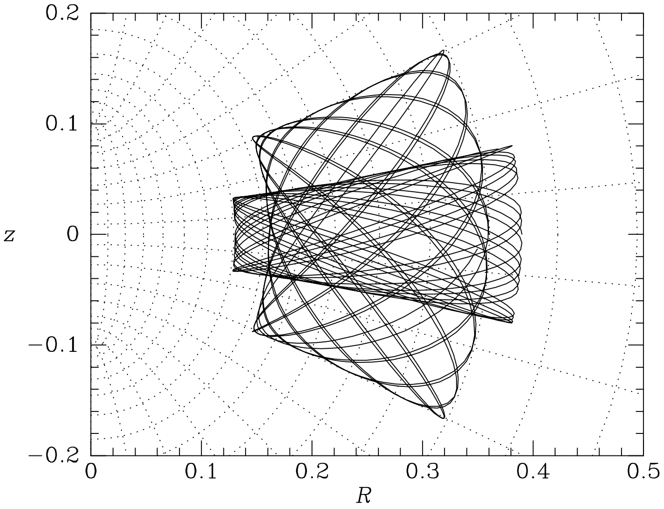


Figure 3.27 The boundaries of orbits in the meridional plane approximately coincide with the coordinate curves of a system of spheroidal coordinates. The dotted lines are the coordinate curves of the system defined by (3.242) and the full curves show the same orbits as Figure 3.4.

the (u, v) coordinate system defined by

$$R = \Delta \sinh u \sin v \quad ; \quad z = \Delta \cosh u \cos v \tag{3.242}$$

achieves this goal to high accuracy: the orbits of Figure 3.4 can be approximately bounded top and bottom by curves of constant v and right and left by curves of constant u.[20]

Now that we have chosen a coordinate system, item (iii) above suggests that we next write the Hamiltonian function in terms of u, v, and their conjugate momenta. The first step is to write the Lagrangian as a function of the new coordinates and their time derivatives. By an analysis that closely parallels the derivation of equations (2.99) we may show that

$$|\dot{\mathbf{x}}|^2 = \Delta^2 \left(\sinh^2 u + \sin^2 v\right) \left(\dot{u}^2 + \dot{v}^2\right) + \Delta^2 \sinh^2 u \sin^2 v \, \dot{\phi}^2, \tag{3.243}$$

and the Lagrangian is

$$\mathcal{L} = \tfrac{1}{2}\Delta^2 \left[\left(\sinh^2 u + \sin^2 v\right) \left(\dot{u}^2 + \dot{v}^2\right) + \sinh^2 u \sin^2 v \, \dot{\phi}^2\right] - \Phi(u, v). \tag{3.244}$$

The momenta are (eq. D.49)

$$\begin{gathered} p_u = \Delta^2 \left(\sinh^2 u + \sin^2 v\right) \dot{u} \quad ; \quad p_v = \Delta^2 \left(\sinh^2 u + \sin^2 v\right) \dot{v} \\ p_\phi = \Delta^2 \sinh^2 u \sin^2 v \, \dot{\phi}, \end{gathered} \tag{3.245}$$

[20] Note that *prolate* spheroidal coordinates are used to fit the boundaries of orbits in *oblate* potentials.

so the Hamiltonian is

$$H(u,v,p_u,p_v,p_\phi) = \frac{p_u^2+p_v^2}{2\Delta^2(\sinh^2 u+\sin^2 v)} + \frac{p_\phi^2}{2\Delta^2 \sinh^2 u \sin^2 v} + \Phi(u,v). \tag{3.246}$$

Since H has no explicit dependence on time, it is equal to some constant E. Likewise, since H is independent of ϕ, the azimuthal momentum p_ϕ is constant at some value L_z.

The examples of motion in harmonic and circular potentials suggest that we seek a form of $\Phi(u,v)$ that will enable us to split a multiple of the equation $H(u,v,p_u,p_v,L_z)=E$ into a part involving only u and p_u and a part that involves only v and p_v. Evidently we require that $(\sinh^2 u+\sin^2 v)\Phi$ be of the form $U(u)-V(v)$, i.e., that[21]

$$\Phi(u,v) = \frac{U(u)-V(v)}{\sinh^2 u+\sin^2 v}, \tag{3.247}$$

for then we may rewrite $H=E$ as

$$2\Delta^2\left[E\sinh^2 u - U(u)\right] - p_u^2 - \frac{L_z^2}{\sinh^2 u} = \frac{L_z^2}{\sin^2 v} + p_v^2 - 2\Delta^2\left[E\sin^2 v + V(v)\right]. \tag{3.248}$$

It can be shown that potentials of the form (3.247) are generated by bodies resembling real galaxies (see Problems 2.6 and 2.14), so there are interesting physical systems for which (3.248) is approximately valid. Potentials of this form are called **Stäckel potentials** after the German mathematician P. Stäckel.[22] Our treatment of these potentials will be restricted; much more detail, including the generalization to triaxial potentials, can be found in de Zeeuw (1985).

If the analogy with the harmonic oscillator holds, p_u will be a function only of u, and similarly for p_v. Under these circumstances, the left side of equation (3.248) does not depend on v, and the right side does not depend on u, so both sides must equal some constant, say $2\Delta^2 I_3$. Hence we would then have

$$p_u = \pm\sqrt{2\Delta^2\left[E\sinh^2 u - I_3 - U(u)\right] - \frac{L_z^2}{\sinh^2 u}}, \tag{3.249a}$$

$$p_v = \pm\sqrt{2\Delta^2\left[E\sin^2 v + I_3 + V(v)\right] - \frac{L_z^2}{\sin^2 v}}. \tag{3.249b}$$

[21] The denominator of equation (3.247) vanishes when $u=0$, $v=0$. However, we may avoid an unphysical singularity in Φ at this point by choosing U and V such that $U(0)=V(0)$.

[22] Stäckel showed that the *only* coordinate system in which the Hamilton–Jacobi equation for $H=\frac{1}{2}p^2+\Phi(\mathbf{x})$ separates is confocal ellipsoidal coordinates. The usual Cartesian, spherical and cylindrical coordinate systems are limiting cases of these coordinates, as is the (u,v,ϕ) system.

It is a straightforward exercise to show that the analogy with the harmonic oscillator *does* hold, by direct time differentiation of both sides of equations (3.249), followed by elimination of $\dot{u}$ and $\dot{p}_u$ with Hamilton's equations (Problem 3.37). Thus the quantity I_3 defined by equations (3.249) *is* an integral. Moreover, we can display I_3 as an explicit function of the phase-space coordinates by eliminating E between equations (3.249) (Problem 3.39).

Equations (3.249) enable us to obtain expressions for the actions J_u and J_v in terms of the integrals E, I_3 and L_z, the last of which is equal to J_ϕ as in the spherical case. Specifically

$$\begin{aligned}
J_u &= \frac{1}{\pi}\int_{u_{\min}}^{u_{\max}} \mathrm{d}u \sqrt{2\Delta^2\left[E\sinh^2 u - I_3 - U(u)\right] - \frac{L_z^2}{\sinh^2 u}},\\
J_v &= \frac{1}{\pi}\int_{v_{\min}}^{v_{\max}} \mathrm{d}v \sqrt{2\Delta^2\left[E\sin^2 v + I_3 + V(v)\right] - \frac{L_z^2}{\sin^2 v}},\\
J_\phi &= L_z,
\end{aligned} \tag{3.250}$$

where $u_{\min}$ and $u_{\max}$ are the smallest and largest values of u at which the integrand vanishes, and similarly for $v_{\min}$ and $v_{\max}$.

As in the spherical case, we obtain expressions for the angle variables by differentiating the generating function $S(u, v, \phi, J_u, J_v, J_\phi)$ of the canonical transformation between angle-action variables and the (u, v, ϕ) system. We take S to be the sum of three parts S_u, S_v and S_ϕ, each of which depends on only one of the three coordinate variables. The gradient of S_u with respect to u is just p_u, so S_u is just the indefinite integral with respect to u of (3.249a). After evaluating S_v and S_ϕ analogously, we use the chain rule to differentiate $S = \sum_i S_i$ with respect to the actions (cf. the derivation of eq. 3.234):

$$\begin{aligned}
\theta_u &= \frac{\partial S}{\partial J_u} = \sum_{i=u,v}\left(\frac{\partial S_i}{\partial H}\Omega_u + \frac{\partial S_i}{\partial I_3}\frac{\partial I_3}{\partial J_u}\right),\\
\theta_v &= \sum_{i=u,v}\left(\frac{\partial S_i}{\partial H}\Omega_v + \frac{\partial S_i}{\partial I_3}\frac{\partial I_3}{\partial J_v}\right),\\
\theta_\phi &= \sum_{i=u,v}\left(\frac{\partial S_i}{\partial H}\Omega_\phi + \frac{\partial S_i}{\partial I_3}\frac{\partial I_3}{\partial L_z}\right) + \phi.
\end{aligned} \tag{3.251}$$

The partial derivatives in these expressions are all one-dimensional integrals that must in general be done numerically.

The condition (3.247) that must be satisfied by an axisymmetric Stäckel potential is very restrictive because it requires that a function of two variables can be written in terms of two functions of one variable. Most potentials that admit a third integral do not satisfy this condition. In particular the logarithmic potential Φ_L (2.71a) that motivated our discussion is not of Stäckel form: we can find a system of spheroidal coordinates that approximately bounds

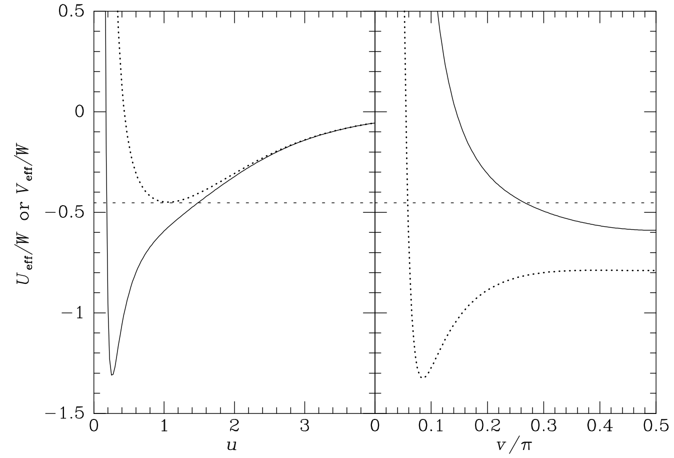


Figure 3.28 Plots of the effective potentials $U_{\rm eff}$ (left) and $V_{\rm eff}$ (right) that are defined by equations (3.252) and (3.253) for $\Delta = 0.6a_3$ and $L_z = 0.05a_3\sqrt{W}$. Curves are shown for $I_3 = -0.1W$ (full) and $I_3 = 0.1W$ (dotted).

any given orbit, but in general different orbits require different coordinate systems.

As an example of the use of equations (3.249) we investigate the shapes they predict for orbits in the potential obtained by choosing in (3.247)

$$\begin{aligned} U(u) &= -W \sinh u \tan^{-1}\left(\frac{\Delta \sinh u}{a_3}\right) \\ V(v) &= W \sin v \tanh^{-1}\left(\frac{\Delta \sin v}{a_3}\right), \end{aligned} \tag{3.252}$$

where W, Δ, and a_3 are constants.[23] An orbit of specified E and I_3 can explore all values of u and v for which equations (3.249) predict positive p_u^2 and p_v^2. This they will do providing E is larger than the largest of the "effective potentials"

$$U_{\rm eff}(u) \equiv \frac{L_z^2}{2\Delta^2 \sinh^4 u} + \frac{I_3 + U(u)}{\sinh^2 u}, \tag{3.253a}$$

$$V_{\rm eff}(v) \equiv \frac{L_z^2}{2\Delta^2 \sin^4 v} - \frac{I_3 + V(v)}{\sin^2 v}. \tag{3.253b}$$

Figure 3.28 shows these potentials for two values of I_3 and all other parameters fixed. Consider the case in which the energy takes the value $-0.453W$

[23] With these choices for U and V, the potential (3.247) becomes the potential of the perfect prolate spheroid introduced in Problem 2.14.

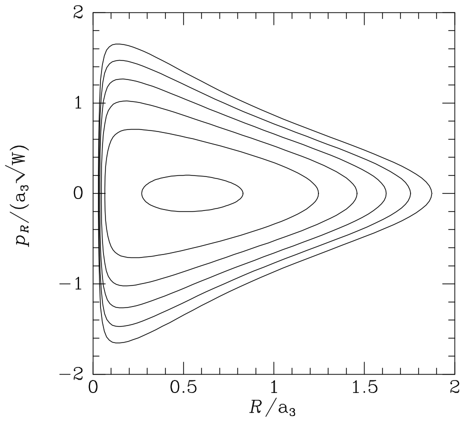


Figure 3.29 Surface of section at $E = -0.5W$ and $L_z = 0.05a_3\sqrt{W}$ constructed from equations (3.249) and (3.252) with $\Delta = 0.6a_3$.

(dashed horizontal line). Then for $I_3 = 0.1W$ (dotted curves), only a single value of u ($u = 1$) is permitted, so the orbit is confined to a segment of an ellipse in the meridional plane—this is a shell orbit. By contrast all values of $|v|$ larger than the intersection of the dashed and dotted curves in the right panel are permitted: these start at $|v| = 0.059\pi$. Consequently, the orbit covers much of the ellipse $u = 1$ (which in three dimensions is a spheroid).

Consider now the case in which $I_3 = -0.1W$ (full curves in Figure 3.28). Now a wide range is permitted in u ($0.17 < u < 1.48$) and a smaller range in v ($|v| > 0.27\pi$). Physically, lowering I_3 transfers some of the available energy from motion perpendicular to the potential's equatorial plane into the star's radial oscillation.

In §3.2.2 we detected the existence of non-classical integrals by plotting surfaces of section. It is interesting to see how I_3 structures surfaces of section. If we were to plot the (u, p_u) surface of section, the consequents of a given orbit (definite values of E, L_z, I_3) would lie on the curve in the (u, p_u) plane whose equation is (3.249a). This equation is manifestly independent of v, so the surface of section would look the same regardless of whether it was for $v = 0$, $v = 0.1$, or whatever. To get the structure of the (R, p_R) surfaces of section that we plotted in §3.2.2, for each allowed value of u we get $p(u)$ from (3.249a) and $p(v)$ from (3.249b) with $v = \pi/2$, and then obtain (R, p_R) from the (u, v, p_u, p_v) coordinates by inverting the transformations (2.96) and (3.245). Figure 3.29 shows a surface of section generated in this way.

In §3.2.1 we saw that motion in the meridional plane is governed by a Hamiltonian $H(R, z, p_R, p_z)$ in which L_z occurs as a parameter and the phase space is four-dimensional. In this space the orbital tori are ordinary two-dimensional doughnuts, and a surface of section is simply a cross-section through a nested sequence of such tori: each invariant curve marks the intersection of a two-dimensional doughnut with the two-dimensional surface of section.

(b) Epicycle approximation In §3.2.3 we used the epicycle approximation to obtain solutions to the equations of motion that are approximately valid for nearly circular orbits in an axisymmetric potential. Here we obtain the corresponding approximate angle-action variables. In cylindrical coordinates the Hamilton–Jacobi equation (3.205) is

$$\tfrac{1}{2}\Big(\frac{\partial S}{\partial R}\Big)^2 + \frac{1}{2R^2}\Big(\frac{\partial S}{\partial \phi}\Big)^2 + \tfrac{1}{2}\Big(\frac{\partial S}{\partial z}\Big)^2 + \Phi(R,z) = E. \tag{3.254}$$

As in equation (2.75a) we assume that Φ is of the form $\Phi_R(R) + \Phi_z(z)$; the radial dependence of $\Phi_z(z)$ is suppressed because the radial motion is small in the epicycle approximation. We further assume that S is of the form $S(\mathbf{J}, R, \phi, z) = S_R(\mathbf{J}, R) + S_\phi(\mathbf{J}, \phi) + S_z(\mathbf{J}, z)$. Now we use the method of separation of variables to split equation (3.254) up into three parts:

$$\begin{gathered} E_z = \tfrac{1}{2}\Big(\frac{\partial S_z}{\partial z}\Big)^2 + \Phi_z(z) \quad ; \quad L_z^2 = \Big(\frac{\partial S_\phi}{\partial \phi}\Big)^2 \\ E - E_z = \tfrac{1}{2}\Big(\frac{\partial S_R}{\partial R}\Big)^2 + \Phi_R(R) + \frac{L_z^2}{2R^2}, \end{gathered} \tag{3.255}$$

where E_z and L_z are the two constants of separation. The first equation of this set leads immediately to an integral for $S_z(z)$

$$S_z(z) = \int_0^z \mathrm{d}z' \, \epsilon_z \sqrt{2[E_z - \Phi_z(z')]}, \tag{3.256}$$

where ϵ_z is chosen to be ± 1 such that the integral increases monotonically along the path. If, as in §3.2.3, we assume that $\Phi_z = \tfrac{1}{2}\nu^2 z^2$, where ν is a constant, then our equation for S_z becomes essentially the same as the first of equations (3.211), and by analogy with equations (3.213) and (3.216), we have

$$J_z = \frac{E_z}{\nu} \quad ; \quad z = -\sqrt{\frac{2J_z}{\nu}}\cos\theta_z. \tag{3.257}$$

The second of equations (3.255) trivially yields

$$S_\phi(\mathbf{J}, \phi) = L_z \phi, \tag{3.258}$$

and it immediately follows that $J_\phi = L_z$. The last of equations (3.255) yields

$$2(E - E_z) = \Big(\frac{\partial S_R}{\partial R}\Big)^2 + 2\Phi_{\rm eff}(R), \tag{3.259a}$$

where (cf. eq. 3.68b)

$$\Phi_{\rm eff}(R) \equiv \Phi_R(R) + \frac{J_\phi^2}{2R^2}. \tag{3.259b}$$

The epicycle approximation involves expanding $\Phi_{\rm eff}$ about its minimum, which occurs at the radius $R_{\rm g}(J_\phi)$ of the circular orbit of angular momentum J_ϕ; with x defined by $R = R_{\rm g} + x$, the expansion is $\Phi(R) = E_{\rm c}(J_\phi) + \frac{1}{2}\kappa^2 x^2$, where $E_{\rm c}(J_\phi)$ is the energy of the circular orbit of angular momentum J_ϕ and κ is the epicycle frequency defined in equation (3.77). Inserting this expansion into (3.259a) and defining $E_R \equiv E - E_z - E_{\rm c}$, we have

$$2E_R = \left(\frac{\partial S_R}{\partial R}\right)^2 + \kappa^2 x^2, \tag{3.260}$$

which is the same as equation (3.210) with K^2 replaced by $2E_R$, x by R, and ω_x by κ. It follows from equations (3.213), (3.216) and (3.217) that

$$J_R = \frac{E_R}{\kappa} \quad ; \quad S_R(\mathbf{J}, R) = J_R(\theta_R - \tfrac{1}{2}\sin 2\theta_R) \quad ; \quad R = R_{\rm g} - \sqrt{\frac{2J_R}{\kappa}}\cos\theta_R. \tag{3.261}$$

The last of these equations is equivalent to equation (3.91) if we set $\theta_R = \kappa t + \alpha$ and $X = -(2J_R/\kappa)^{1/2}$.

Finally, we find an expression for θ_ϕ. With equations (3.258) and (3.261) we have

$$\begin{aligned}\theta_\phi = \frac{\partial S}{\partial J_\phi} = \frac{\partial S_\phi}{\partial J_\phi} + \frac{\partial S_R}{\partial J_\phi} &= \phi + J_R(1 - \cos 2\theta_R)\frac{\partial \theta_R}{\partial J_\phi} \\ &= \phi + 2J_R \sin^2\theta_R \frac{\partial \theta_R}{\partial J_\phi}.\end{aligned} \tag{3.262}$$

The derivative of θ_R has to be taken at constant J_R, J_z, R, ϕ, and z. We differentiate the last of equations (3.261) bearing in mind that both $R_{\rm g}$ and κ are functions of J_ϕ:

$$0 = \frac{{\rm d}R_{\rm g}}{{\rm d}J_\phi} + \frac{1}{2\kappa}\frac{{\rm d}\kappa}{{\rm d}J_\phi}\sqrt{\frac{2J_R}{\kappa}}\cos\theta_R + \sqrt{\frac{2J_R}{\kappa}}\sin\theta_R\frac{\partial\theta_R}{\partial J_\phi}. \tag{3.263}$$

By differentiating $R_{\rm g}^2\Omega_{\rm g} = J_\phi$ with respect to $R_{\rm g}$ we may show with equation (3.80) that

$$\frac{{\rm d}R_{\rm g}}{{\rm d}J_\phi} = \frac{\gamma}{\kappa R_{\rm g}}, \tag{3.264}$$

where $\gamma = 2\Omega_{\rm g}/\kappa$ is defined by equation (3.93b). Inserting this relation into (3.263) and using the result to eliminate $\partial\theta_R/\partial J_\phi$ from (3.262), we have finally

$$\theta_\phi = \phi - \frac{\gamma}{R_{\rm g}}\sqrt{\frac{2J_R}{\kappa}}\sin\theta_R - \frac{J_R}{2}\frac{{\rm d}\ln\kappa}{{\rm d}J_\phi}\sin 2\theta_R. \tag{3.265}$$

This expression should be compared with equation (3.93a). If we set $\theta_\phi = \Omega_{\rm g}t + \phi_0$, $\theta_R = \kappa t + \alpha + \pi$, and $X = (2J_R/\kappa)^{1/2}$ as before, the only difference

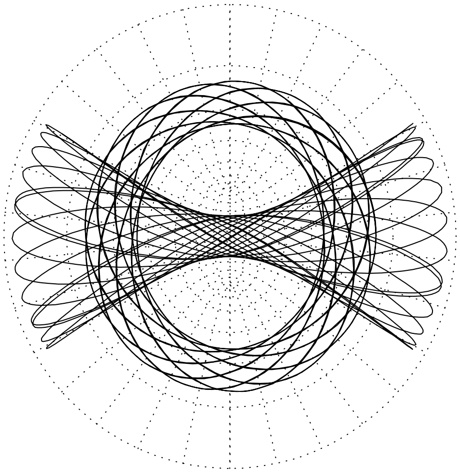


Figure 3.30 The boundaries of loop and box orbits in barred potentials approximately coincide with the curves of a system of spheroidal coordinates. The figure shows two orbits in the potential Φ_L of equation (3.103), and a number of curves on which the coordinates u and v defined by equations (3.267) are constant.

between the two equations is the presence of a term proportional to $\sin 2\theta_R$ in equation (3.265). For nearly circular orbits, this term is smaller than the term proportional to $\sin\theta_R$ by $\sqrt{J_R/J_\phi}$ and represents a correction to equation (3.92) that makes the $(\theta_R, \theta_\phi, J_R, J_\phi)$ coordinates canonical (Dehnen 1999a).

It is worth noting that when $J_R \neq 0$, the frequency associated with ϕ is not the circular frequency, $\Omega_{\rm g}$. To see this, recall that the Hamiltonian $H = E_R + E_{\rm c} + E_z$, and $E_R = \kappa J_R$, while $\mathrm{d}E_{\rm c}/\mathrm{d}J_\phi = \Omega_{\rm g}$, so

$$\Omega_\phi = \frac{\partial H}{\partial J_\phi} = \frac{\mathrm{d}\kappa}{\mathrm{d}J_\phi} J_R + \Omega_{\rm g}. \tag{3.266}$$

3.5.4 Angle-action variables for a non-rotating bar

The (u, v) coordinate system that allowed us to recover angle-action variables for flattened axisymmetric potentials enables us to do the same for a planar, non-rotating bar. This fact is remarkable, because we saw in §3.3 that these systems support two completely different types of orbit, loops and boxes. Figure 3.30 makes it plausible that the (u, v) system can provide analytic solutions for both loops and boxes, by showing that the orbits plotted in Figure 3.8 have boundaries that may be approximated by curves of constant u and v (cf. the discussion on page 226). We can explore this idea quantitatively by defining

$$x = \Delta \sinh u \sin v \quad ; \quad y = \Delta \cosh u \cos v \tag{3.267}$$

and then replacing R by x and z by y in the formulae of the previous subsection. Further setting $\dot\phi = L_z = 0$ we find by analogy with equations (3.249) that

$$p_u = \pm\Delta \sinh u \sqrt{2[E - U_{\rm eff}(u)]} \quad ; \quad p_v = \pm\Delta \sin v \sqrt{2[E - V_{\rm eff}(v)]} \tag{3.268a}$$

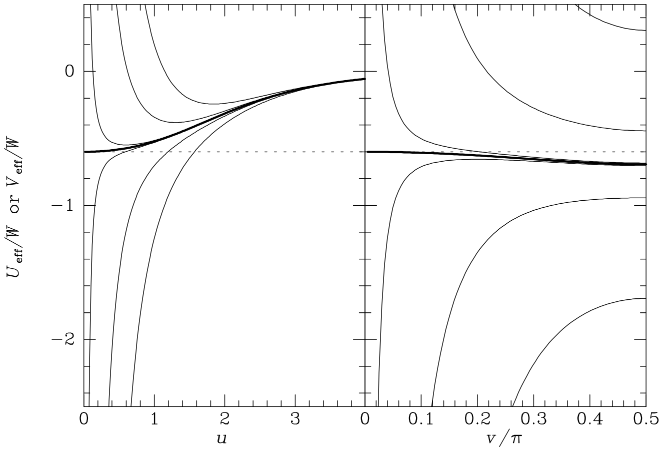


Figure 3.31 The effective potentials defined by equations (3.268b) when U and V are given by equations (3.252). The curves are for $I_2 = 1, 0.25, 0.01, 0, -0.01, -0.25$ and -1, with the largest values coming on top in the left panel and on the bottom in the right panel. The thick curves are for $I_2 = 0$.

where

$$U_{\rm eff}(u) = \frac{I_2 + U(u)}{\sinh^2 u} \quad ; \quad V_{\rm eff}(v) = -\frac{I_2 + V(v)}{\sin^2 v}. \tag{3.268b}$$

Here U and V are connected to the gravitational potential by equation (3.247) as before and I_2 is the constant of separation analogous to I_3.

An orbit of specified E and I_2 is confined to values of u and v at which both $E \geq U_{\rm eff}$ and $E \geq V_{\rm eff}$. Figure 3.31 shows the effective potentials as functions of their coordinates for several values of I_2 when U and V are chosen to be the functions specified by equations (3.252). In each panel the thick curve is for $I_2 = 0$, with curves for $I_2 > 0$ lying above this in the left panel, and below it on the right. Since the curves of $U_{\rm eff}$ have minima only when $I_2 > 0$, there is a lower limit on the star's u coordinate only in this case. Consequently, stars with $I_2 \leq 0$ can reach the center, while stars with $I_2 > 0$ cannot reach the center. This suggests that when $I_2 \leq 0$ the orbit is a box orbit, while when $I_2 > 0$ it is a loop orbit. Comparison of the right and left panels confirms this conjecture by showing that when $I_2 > 0$ (upper curves on left and lower curves on right), the minimum value of $U_{\rm eff}$ is greater than the maximum of $V_{\rm eff}$. Hence when $I_2 > 0$ the condition $E > V_{\rm eff}(v)$ imposes no constraint on v and the boundaries of the orbit are the ellipses $u = u_{\rm min}$ and $u = u_{\rm max}$ on which $E = U_{\rm eff}$. When $I_2 \leq 0$, by contrast, the curves on the right tend to ∞ as $v \to 0$, so sufficiently small values of v are excluded and the boundaries of the orbit are the ellipse $u = u_{\rm max}$ on which $E = U_{\rm eff}(u)$ and the hyperbola $|v| = v_{\rm min}$ on which $E = V_{\rm eff}(v)$.

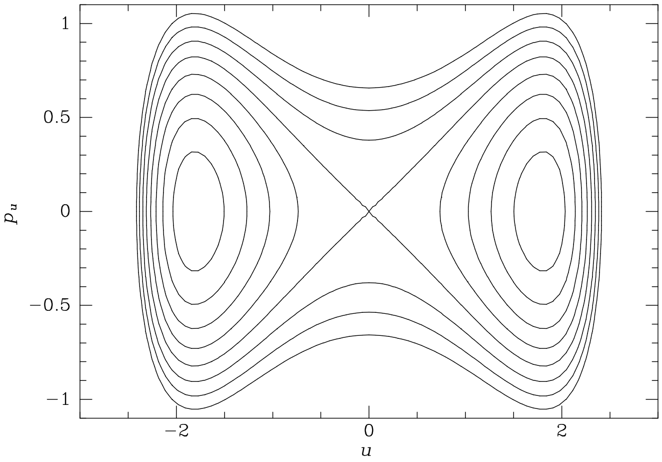


Figure 3.32 The (u, p_u), $v = 0$ surface of section for motion at $E = -0.25$ in the Stäckel potential defined by equations (3.247) and (3.252) with $\Delta = 0.6$ and $a_3 = 1$. Each curve is a contour of constant I_2 (eqs. 3.268). The invariant curves of box orbits ($I_2 = -0.6, -0.4, \ldots$) run round the outside of the figure, while the bull's-eyes at right are the invariant curves of anti-clockwise loop orbits. Temporarily suspending the convention that loops always have $u > 0$, we show the invariant curves of clockwise loops as the bull's-eyes at left.

Figure 3.32 shows the (u, p_u) surface of section, which is in practice nothing more than a contour plot of the integral $I_2(E, u, p_u)$ with E fixed (eq. 3.268a). Each contour shows the curve in which an orbital torus is sliced by the surface of section. As in Figure 3.9, for example, there are two different types of contour, namely those generated by the tori of loop orbits (which come in pairs, because there are both clockwise and anti-clockwise circulating loops), and those generated by the tori of box orbits, which envelop all the tori of the loop orbits.

3.5.5 Summary

We have made a considerable investment in the theory of angle-action variables, which is repaid by the power of these variables in investigations of a wide variety of dynamical problems. This power arises from the following features:

(i) Angle-action variables are canonical. In particular, the phase-space volume $\mathrm{d}^3\boldsymbol{\theta}\mathrm{d}^3\mathbf{J}$ is the same as the phase-space volume $\mathrm{d}^3\mathbf{q}\mathrm{d}^3\mathbf{p}$ for any other set of canonical variables $(\mathbf{q}, \mathbf{p})$, including the usual Cartesian coordinates $(\mathbf{x}, \mathbf{v})$.

(ii) Every set of angle-action variables $(\boldsymbol{\theta}, \mathbf{J})$ is associated with a Hamiltonian[24] $H(\mathbf{J})$, and orbits in this Hamiltonian have the simple form

[24] If a given set of angle-action variables is associated with $H(\mathbf{J})$, then it is also as-

$\mathbf{J} = constant$, $\boldsymbol{\theta} = \boldsymbol{\Omega}t + constant$. A Hamiltonian that admits angle-action variables is said to be **integrable**. The simplicity of angle-action variables makes them indispensable for investigating motion in non-integrable Hamiltonians by using perturbation theory. This technique will be used to explore chaotic orbits in §3.7, and the stability of stellar systems in Chapter 5.

(iii) In the next section we shall see that actions are usually invariant during slow changes in the Hamiltonian.

3.6 Slowly varying potentials

So far we have been concerned with motion in potentials that are time-independent in either an inertial or a rotating frame. It is sometimes necessary to consider how stars move in potentials that are time-dependent. The nature of the problem posed by a time-varying potential depends on the speed with which the potential evolves. In this section we shall confine ourselves to potentials that evolve slowly, in which case angle-action variables enable us to predict how a stellar system will respond to changes in the gravitational field that confines it. Such changes occur when:

(i) Encounters between the individual stars at the core of a dense stellar system (such as a globular cluster or galaxy center) cause the core to evolve on a timescale of order the relaxation time (1.38), which is much longer than the orbital times of individual stars (§7.5).

(ii) Stars of galaxies and globular clusters lose substantial quantities of mass as they gradually evolve and shed their envelopes into interstellar or intergalactic space (Box 7.2).

(iii) Gas settles into the equatorial plane of a pre-existing dark halo to form a spiral galaxy. In this case the orbits of the halo's dark-matter particles will undergo a slow evolution as the gravitational potential of the disk gains in strength.

Potential variations that are slow compared to a typical orbital frequency are called **adiabatic**. We now show that the actions of stars are constant during such adiabatic changes of the potential. For this reason actions are often called **adiabatic invariants**.

3.6.1 Adiabatic invariance of actions

Suppose we have a sequence of potentials $\Phi_\lambda(\mathbf{x})$ that depend continuously on the parameter λ. For each fixed λ we assume that angle-action variables could be constructed for Φ_λ. That is, we assume that at all times phase space is filled by arrays of nested tori on which the phase points of individual stars

sociated with $\tilde{H}(\mathbf{J}) \equiv f[H(\mathbf{J})]$, where f is any differentiable function. Thus, a set of angle-action variables is associated with *infinitely many* Hamiltonians.

travel. We consider what happens when λ is changed from its initial value, say $\lambda = \lambda_0$, to a new value λ_1. After this change has occurred, each star's phase point will start to move on a torus of the set that belongs to Φ_{λ_1}. In general, two stellar phase points that started out on the same torus of Φ_{λ_0} will end up on two different tori of Φ_{λ_1}. But if λ is changed very slowly compared to all the characteristic times $2\pi/\Omega_k$ associated with motion on each torus, all phase points that are initially on a given torus of Φ_{λ_0} will be equally affected by the variation of λ. This statement follows from the time averages theorem of §3.5.1a, which shows that all stars spend the same fraction of their time in each portion of the torus; hence, all stars are affected by slow changes in Φ_λ in the same way. Thus all phase points that start on the same torus of Φ_{λ_0} will end on a single torus of Φ_{λ_1}. Said in other language, any two stars that are initially on a common orbit (but at different phases) will still be on a common orbit after the slow variation of λ is complete.

Suppose the variation of λ starts at time $t = 0$ and is complete by time τ, and let $\mathbf{H}_t$ be the time-evolution operator defined in equation (D.55). Then we have just seen that $\mathbf{H}_\tau$, which is a canonical map (see Appendix D.4.4), maps tori of Φ_{λ_0} onto tori of Φ_{λ_1}. These facts guarantee that actions are adiabatically invariant, for the following reason. Choose three closed curves γ_i, on any torus M of Φ_{λ_0} that through the integrals (3.195) generate the actions J_i of this torus. Then, since $\mathbf{H}_\tau$ is the endpoint of a continuous deformation of phase space into itself, the images $\mathbf{H}_\tau(\gamma_i)$ of these curves are suitable curves along which to evaluate the actions J_i' of $\mathbf{H}_\tau(M)$, the torus to which M is mapped by $\mathbf{H}_\tau$. But by a corollary to the Poincaré invariant theorem (Appendix D.4.2), we have that if γ is any closed curve and $\mathbf{H}_\tau(\gamma)$ is its image under the canonical map $\mathbf{H}_\tau$, then

$$\oint_{\mathbf{H}_\tau(\gamma)} \mathbf{p}\cdot \mathrm{d}\mathbf{q} = \oint_\gamma \mathbf{p}\cdot \mathrm{d}\mathbf{q}. \tag{3.269}$$

Hence $J_i' = J_i$, and the actions of stars do not change if the potential evolves sufficiently slowly.

It should be stressed that any action J_i with fundamental frequency $\Omega_i = 0$ is not an adiabatic invariant. For example, in a spherical potential, J_2 and J_3 are normally adiabatic invariants, but J_1 is not (Table 3.1).

3.6.2 Applications

We illustrate these ideas with a number of simple examples. Other applications of adiabatic invariants will be found in Binney & May (1986), Lichtenberg & Lieberman (1992), and §4.6.1.

(a) Harmonic oscillator We first consider the one-dimensional harmonic oscillator whose potential is

$$\Phi = \tfrac{1}{2}\omega^2 x^2. \tag{3.270}$$

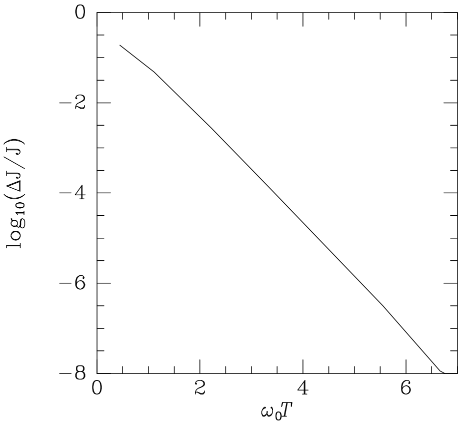


Figure 3.33 Checking the invariance of the action (3.271) when the natural frequency of a harmonic oscillator is varied according to equation (3.277). ΔJ is the RMS change in the action on integrating the oscillator's equation of motion from $t = -20T$ to $t = 20T$, using eight equally spaced phases. The RMS change in J declines approximately as $\Delta J \propto \exp(-2.8\omega_0 T)$.

By equation (3.213) the action is

$$J = \frac{1}{2\omega}\left[p^2 + (\omega x)^2\right] = \frac{H}{\omega}, \tag{3.271}$$

where $H(x,p) = \frac{1}{2}p^2 + \frac{1}{2}\omega^2 x^2$. The general solution of the equations of motion is $x(t) = X\cos(\omega t + \phi)$. In terms of the amplitude of oscillation X we have

$$J = \tfrac{1}{2}\omega X^2. \tag{3.272}$$

Now suppose that the oscillator's spring is slowly stiffened by a factor $s^2 > 1$, so the natural frequency increases to

$$\omega' = s\omega. \tag{3.273}$$

By the adiabatic invariance of J, the new amplitude X' satisfies

$$\tfrac{1}{2}\omega' X'^2 = J = \tfrac{1}{2}\omega X^2. \tag{3.274}$$

Thus the amplitude is diminished to

$$X' = \frac{X}{\sqrt{s}}, \tag{3.275}$$

while the energy, $E = \omega J$, has increased to[25]

$$E' = \omega' J = s\omega J = sE. \tag{3.276}$$

[25] The simplest proof of this result uses quantum mechanics. The energy of a harmonic oscillator is $E = (n + \frac{1}{2})\hbar\omega$ where n is an integer. When ω is slowly varied, n cannot change discontinuously and hence must remain constant. Therefore $E/\omega = E'/\omega'$. Of course, for galaxies n is rather large.

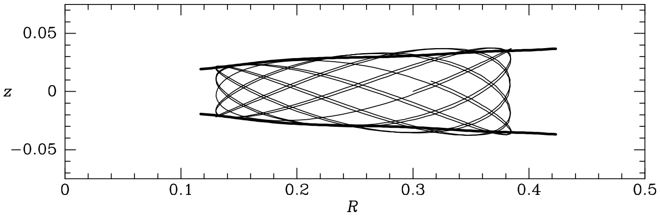


Figure 3.34 The envelope of an orbit in the effective potential (3.70) with $q = 0.5$ (light curve) is well modeled by equation (3.279) (heavy curves).

We now ask how rapidly we can change the frequency ω without destroying the invariance of J. Let ω vary with time according to

$$\omega(t) = \pi\sqrt{3 + \operatorname{erf}(t/T)}. \tag{3.277}$$

Thus the frequency changes from $\omega = \omega_0 \equiv \sqrt{2}\pi$ at $t \ll -T$ to $\omega = 2\pi = \sqrt{2}\omega_0$ at $t \gg T$. In Figure 3.33 we show the results of numerically integrating the oscillator's equation of motion with $\omega(t)$ given by equation (3.277). We plot the RMS difference ΔJ between the initial and final values of J for eight different phases of the oscillator at $t = -20T$. For $\omega_0 T \gtrsim 2$, J changes by less than half a percent, and for $\omega_0 T \gtrsim 4$, J changes by less than 3×10^{-5}. We conclude that the potential does not have to change very slowly for J to be well conserved. In fact, one can show that the fractional change in J is in general less than $\exp(-\omega T)$ for $\omega T \gg 1$ (Lichtenberg & Lieberman 1992).

(b) Eccentric orbits in a disk Consider the shapes shown in Figure 3.4 of the orbits in the meridional plane of an axisymmetric galaxy. On page 167 we remarked that disk stars in the solar neighborhood oscillate perpendicular to the galactic plane considerably more rapidly than they oscillate in the radial direction. Therefore, if we take the radial coordinate $R(t)$ of a disk star to be a known function of time, we may consider the equation of motion (3.67c) of the z-coordinate to describe motion in a slowly varying potential. If the amplitude of the z-oscillations is small, we may expand $\partial\Phi/\partial z$ about $z = 0$ to find

$$\ddot{z} \simeq -\omega^2 z \quad \text{where} \quad \omega(t) \equiv \left(\frac{\partial^2\Phi}{\partial z^2}\right)^{1/2}_{[R(t),0]} \equiv \sqrt{\Phi_{zz}[R(t),0]}. \tag{3.278}$$

If the action integral of this harmonic oscillator is conserved, we expect the amplitude $Z(R)$ to satisfy (see eqs. 3.273 and 3.275)

$$Z(R) = Z(R_0)\left(\frac{\Phi_{zz}(R_0,0)}{\Phi_{zz}(R,0)}\right)^{1/4}. \tag{3.279}$$

Figure 3.34 compares the prediction of (3.279) with the true shape of an orbit in the effective potential (3.70). Evidently the behavior of such orbits can be accurately understood in terms of adiabatic invariants.

(c) Transient perturbations Consider the motion of a star on a loop orbit in a slowly varying planar potential $\Phi(R,\phi)$. The relevant action is

$$J_\phi = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}\phi\, p_\phi. \tag{3.280}$$

We now conduct the following thought experiment. Initially the potential Φ is axisymmetric. Then $p_\phi = L_z$ is an integral, and we can trivially evaluate the integral in (3.280) to obtain $J_\phi = L_z$. We now slowly distort the potential in some arbitrary fashion into a new axisymmetric configuration. At the end of this operation, the azimuthal action, being adiabatically invariant, still has value J_ϕ and is again equal to the angular momentum L_z. Thus the star finishes the experiment with the same angular momentum with which it started,[26] even though its instantaneous angular momentum, p_ϕ, was changing during most of the experiment. Of course, if the potential remains axisymmetric throughout, p_ϕ remains an integral at all times and is exactly conserved no matter how rapidly the potential is varied.

A closely related example is a slowly varying external perturbation of a stellar system, perhaps from the gravitational field of an object passing at a low angular velocity. If the passage is slow enough, the actions are adiabatically invariant, so the distribution of actions in the perturbed system will be unchanged by the encounter. In other words, adiabatic encounters, even strong ones, have no lasting effect on a stellar system (§8.2c).

(d) Slow growth of a central black hole As our final application of the adiabatic invariance of actions, we consider the evolution of the orbit of a star near the center of a spherical galaxy, as a massive black hole grows by slowly accreting matter (Goodman & Binney 1984). A more complete treatment of the problem is given in §4.6.1d. We assume that prior to the formation of the hole, the density of material interior to the orbit can be taken to be a constant, so the potential is that of the spherical harmonic oscillator. It is then easy to show that the star's Hamiltonian can be written (Problem 3.36)

$$H = \Omega_r J_r + \Omega_\phi J_\phi = 2\Omega J_r + \Omega J_\phi, \tag{3.281}$$

where $\Omega = \Omega_\phi = \frac{1}{2}\Omega_r$ is the circular frequency, and $J_\phi = L$ is the magnitude of the angular-momentum vector. The radii $r_{\min}$ and $r_{\max}$ of peri- and apocenter are the roots of

$$0 = \frac{J_\phi^2}{2r^2} + \tfrac{1}{2}\Omega^2 r^2 - H \quad \Rightarrow \quad 0 = r^4 - \frac{2H}{\Omega^2}r^2 + \frac{J_\phi^2}{\Omega^2}. \tag{3.282}$$

[26] This statement does not apply for stars that switch from loop to box orbits and back again as the potential is varied (Binney & Spergel 1983; Evans & Collett 1994). These stars will generally be on highly eccentric orbits initially.

Hence, the axis ratio of the orbit is

$$\begin{aligned} q_{\rm H} = \frac{r_{\rm min}}{r_{\rm max}} &= \left(\frac{H/\Omega^2 - \left[(H/\Omega^2)^2 - (J_\phi/\Omega)^2\right]^{1/2}}{H/\Omega^2 + \left[(H/\Omega^2)^2 - (J_\phi/\Omega)^2\right]^{1/2}}\right)^{1/2} \\ &= \left(\frac{2J_r + J_\phi - 2[J_r(J_r+J_\phi)]^{1/2}}{2J_r + J_\phi + 2[J_r(J_r+J_\phi)]^{1/2}}\right)^{1/2}. \end{aligned} \qquad (3.283)$$

Multiplying top and bottom of the fraction by the top, this last expression reduces to

$$q_{\rm H} = \frac{1}{J_\phi}\left[2J_r + J_\phi - 2\sqrt{J_r(J_r+J_\phi)}\,\right]. \qquad (3.284)$$

When the hole has become sufficiently massive, the Hamiltonian may be taken to be that for Kepler motion (eq. E.6) and the orbit becomes an ellipse with the black hole at the focus rather than the center of the ellipse. A similar calculation yields for the axis ratio of this ellipse

$$q_{\rm K} = \left[1 - \left(\frac{r_{\rm max} - r_{\rm min}}{r_{\rm max} + r_{\rm min}}\right)^2\right]^{1/2} = \frac{J_\phi}{J_r + J_\phi}. \qquad (3.285)$$

When J_r/J_ϕ is eliminated between equations (3.284) and (3.285), we find

$$q_{\rm K} = \frac{4q_{\rm H}}{(1+q_{\rm H})^2}. \qquad (3.286)$$

For example, if $q_{\rm H} = 0.5$ is the original axis ratio, the final one is $q_{\rm K} = 0.889$, and if initially $q_{\rm H} = 0.75$, then finally $q_{\rm K} = 0.980$. Physically, an elongated ellipse that is centered on the black hole distorts into a much rounder orbit with the black hole at one focus.

For any orbit in a spherical potential the mean-square radial speed is

$$\overline{v_r^2} = \frac{\Omega_r}{\pi}\int_0^{\pi/\Omega_r} {\rm d}t\, v_r^2 = \frac{\Omega_r}{\pi}\int_{r_{\rm min}}^{r_{\rm max}} {\rm d}r\, v_r = \Omega_r J_r. \qquad (3.287a)$$

Similarly, the mean-square tangential speed is

$$\overline{v_{\rm t}^2} = \frac{\Omega_\phi}{2\pi}\int_0^{2\pi/\Omega_\phi} {\rm d}t\,(R\dot\phi)^2 = \frac{\Omega_\phi}{2\pi}\int_0^{2\pi} {\rm d}\phi\, p_\phi = \Omega_\phi J_\phi. \qquad (3.287b)$$

Since the actions do not change as the hole grows, the change in the ratio of the mean-square speeds is given by

$$\frac{\left(\overline{v_r^2}/\overline{v_{\rm t}^2}\right)_{\rm K}}{\left(\overline{v_r^2}/\overline{v_{\rm t}^2}\right)_{\rm H}} = \frac{(\Omega_r/\Omega_\phi)_{\rm K}}{(\Omega_r/\Omega_\phi)_{\rm H}} = \frac{1}{2}. \qquad (3.288)$$

Consequently, the growth of the black hole increases the star's tangential velocity much more than it does the radial velocity, irrespective of the original eccentricity of the orbit. In §4.6.1a we shall investigate the implications of this result for measurements of the stellar velocity dispersion near a black hole, and show how the growth of the black hole enhances the density of stars in its vicinity.

3.7 Perturbations and chaos

Analytic solutions to a star's equations of motion exist for only a few simple potentials $\Phi(\mathbf{x})$. If we want to know how stars will move in a more complex potential, for example one estimated from observational data, two strategies are open to us: either solve the equations of motion numerically, or obtain an approximate analytic solution by invoking perturbation theory, which involves expressing the given potential as a sum of a potential for which we can solve the equations of motion analytically and a (one hopes) small additional term.

Even in the age of fast, cheap and convenient numerical computation, perturbative solutions to the equations of motion are useful in two ways. First, they can be used to investigate the stability of stellar systems (§5.3). Second, they give physical insight into the dynamics of orbits. We start this section by developing perturbation theory and sketching some of its astronomical applications; then we describe the phenomenon of orbital chaos, and show that Hamiltonian perturbation theory helps us to understand the physics of this phenomenon.

3.7.1 Hamiltonian perturbation theory

In §3.3.3 we derived approximate orbits in the potential of a weak bar, by treating the potential as a superposition of a small non-axisymmetric potential and a much larger axisymmetric one. Our approach involved writing the orbit $\mathbf{x}(t)$ as a sum of two parts, one of which described the circular orbit of a guiding center, while the other described epicyclic motion. We worked directly with the equations of motion. Angle-action variables enable us to develop a more powerful perturbative scheme, in which we work with scalar functions rather than coordinates, and think of the orbit as a torus in phase space rather than a time-ordered series of points along a trajectory. For more detail see Lichtenberg & Lieberman (1992).

Let H^0 be an integrable Hamiltonian, and consider the one-parameter family of Hamiltonians

$$H^\beta \equiv H^0 + \beta h, \tag{3.289}$$

where $\beta \ll 1$ and h is a Hamiltonian with gradients that are comparable in magnitude to those of H^0. Let $(\boldsymbol{\theta}^\beta, \mathbf{J}^\beta)$ be angle-action variables for H^β. These coordinates are related to the angle-action variables of H^0 by a canonical transformation. As $\beta \to 0$ the generating function S (Appendix D.4.6) of this transformation will tend to the generating function of the identity transformation, so we may write

$$S(\boldsymbol{\theta}^\beta, \mathbf{J}^0) = \boldsymbol{\theta}^\beta \cdot \mathbf{J}^0 + s^\beta(\boldsymbol{\theta}^\beta, \mathbf{J}^0), \tag{3.290}$$

where s^β is $\mathrm{O}(\beta)$, and (eq. D.94)

$$\mathbf{J}^\beta = \frac{\partial S}{\partial \boldsymbol{\theta}^\beta} = \mathbf{J}^0 + \frac{\partial s^\beta}{\partial \boldsymbol{\theta}^\beta} \quad ; \quad \boldsymbol{\theta}^0 = \boldsymbol{\theta}^\beta + \frac{\partial s^\beta}{\partial \mathbf{J}^0}. \tag{3.291}$$

Substituting these equations into (3.289), we have

$$\begin{aligned} H^\beta(\mathbf{J}^\beta) &= H^0(\mathbf{J}^0) + \beta h(\boldsymbol{\theta}^0, \mathbf{J}^0) \\ &= H^0\Big(\mathbf{J}^\beta - \frac{\partial s^\beta}{\partial \boldsymbol{\theta}^\beta}\Big) + \beta h\Big(\boldsymbol{\theta}^\beta + \frac{\partial s^\beta}{\partial \mathbf{J}^0}, \mathbf{J}^\beta - \frac{\partial s^\beta}{\partial \boldsymbol{\theta}^\beta}\Big) \\ &= H^0(\mathbf{J}^\beta) - \boldsymbol{\Omega}^0(\mathbf{J}^\beta) \cdot \frac{\partial s^\beta}{\partial \boldsymbol{\theta}^\beta} + \beta h(\boldsymbol{\theta}^\beta, \mathbf{J}^\beta) + \mathrm{O}(\beta^2), \end{aligned} \tag{3.292}$$

where $\boldsymbol{\Omega}^0$ is the derivative of H^0 with respect to its argument. We next expand h and s^β as Fourier series in the periodic angle variables (Appendix B.4):

$$h(\boldsymbol{\theta}^\beta, \mathbf{J}^\beta) = \sum_{\mathbf{n}} h_{\mathbf{n}}(\mathbf{J}^\beta)\, \mathrm{e}^{\mathrm{i}\mathbf{n}\cdot\boldsymbol{\theta}^\beta} \;\; ; \;\; s^\beta(\boldsymbol{\theta}^\beta, \mathbf{J}^0) = \mathrm{i} \sum_{\mathbf{n}} s^\beta_{\mathbf{n}}(\mathbf{J}^0)\, \mathrm{e}^{\mathrm{i}\mathbf{n}\cdot\boldsymbol{\theta}^\beta}, \tag{3.293}$$

where $\mathbf{n} = (n_1, n_2, n_3)$ is a triple of integers. Substituting these expressions into (3.292) we find

$$H^\beta(\mathbf{J}^\beta) = H^0(\mathbf{J}^\beta) + \beta h_0 + \sum_{\mathbf{n}\neq\mathbf{0}} \Big(\beta h_{\mathbf{n}} + \mathbf{n}\cdot\boldsymbol{\Omega}^0 s^\beta_{\mathbf{n}}\Big)\, \mathrm{e}^{\mathrm{i}\mathbf{n}\cdot\boldsymbol{\theta}^\beta} + \mathrm{O}(\beta^2). \tag{3.294}$$

In this equation $\boldsymbol{\Omega}^0$ and $h_{\mathbf{n}}$ are functions of $\mathbf{J}^\beta$, while $s_{\mathbf{n}}$ is a function of $\mathbf{J}^0$, but to the required order in β, $\mathbf{J}^0$ can be replaced by $\mathbf{J}^\beta$.

Since the left side of equation (3.294) does not depend on $\boldsymbol{\theta}^\beta$, on the right the coefficient of $\exp(\mathrm{i}\mathbf{n}\cdot\boldsymbol{\theta}^\beta)$ must vanish for all $\mathbf{n} \neq 0$. Hence the Fourier coefficients of S are given by

$$s^\beta_{\mathbf{n}}(\mathbf{J}) = -\frac{\beta h_{\mathbf{n}}(\mathbf{J})}{\mathbf{n}\cdot\boldsymbol{\Omega}^0(\mathbf{J})} + \mathrm{O}(\beta^2) \quad (\mathbf{n} \neq 0). \tag{3.295}$$

The $\mathrm{O}(\beta)$ part of equation (3.295) defines the generating function of a canonical transformation. Let $(\boldsymbol{\theta}', \mathbf{J}')$ be the images of $(\boldsymbol{\theta}^0, \mathbf{J}^0)$ under this transformation. Then we have shown that

$$H^\beta(\mathbf{J}^\beta) = H'(\mathbf{J}') + \beta^2 h'(\boldsymbol{\theta}', \mathbf{J}'), \tag{3.296a}$$

where

$$H'(\mathbf{J}') \equiv H^0(\mathbf{J}') + \beta h_0(\mathbf{J}') \tag{3.296b}$$

and h' is a function involving second derivatives of H^0 and first derivatives of h.

The analysis we have developed can be used to approximate orbits in a given potential. As we saw in §3.2.2, if we know an integral other than the Hamiltonian of a system with two degrees of freedom, we can calculate the curve in a surface of section on which the consequents of a numerically

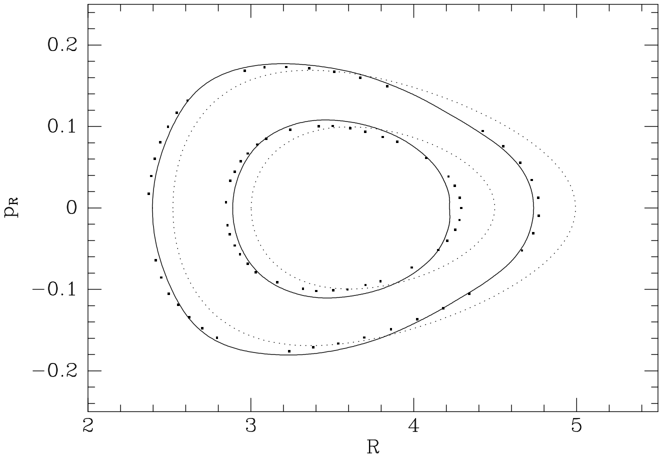


Figure 3.35 A surface of section for orbits in a flatted isochrone potential. The density distribution generating the potential has axis ratio $q = 0.7$. The points are the consequents of numerically calculated orbits. The dotted curves show the orbital tori for the spherical isochrone potential that have the same actions as the numerically integrated orbits. The full curves show the result of using first-order Hamiltonian perturbation theory to deform these tori.

integrated orbit should lie. Since $\mathbf{J}'$ differs from the true action by only $\mathrm{O}(\beta^2)$ it should provide an approximate integral of motion, and it is interesting to compare the invariant curve that it yields with the consequents of a numerically integrated orbit. Figure 3.35 is a surface of section for orbits in a flattened isochrone potential. The density distribution that generates this potential is obtained by replacing r by $\sqrt{R^2 + z^2/q^2}$ and M by M/q in equations (2.48b) and (2.49). The axis ratio q has been set equal to 0.7. The dots show the consequents of numerically integrated orbits. The dotted curves show the corresponding invariant curves for the spherical isochrone. The full curves show the results of applying first-order perturbation theory to the spherical isochrone to obtain better approximations to invariant curves.

The full curves in Figure 3.35 fit the numerical consequents much better than the dotted curves, but the fit is not perfect. An obvious strategy for systematically improving our approximation to the true angle-action variables is to use our existing machinery to derive from (3.296a) a second canonical transformation that would enable us to write H as a sum of a Hamiltonian $H''(\mathbf{J}'')$ that is a function of new actions $\mathbf{J}''$ and a yet smaller perturbation $\beta^4 h''$. After we have performed k transformations, the angle-dependent part of H^β will be of order β^{2^k}. In practice this procedure is unlikely to work because after each application the "unperturbed" frequencies of the orbit change from $\mathbf{\Omega}' = \partial H'/\partial \mathbf{J}'$ to $\mathbf{\Omega}'' = \partial H''/\partial \mathbf{J}''$, and sooner or later we will find that $\mathbf{n} \cdot \mathbf{\Omega}''$ is very close to zero for some $\mathbf{n}$, with the consequence that the corresponding term in the generating function (3.295) becomes large.

This is the problem of **small divisors**. Fortunately, in many applications the coefficients $h_{\mathbf{n}}$ in the numerators of (3.295) decline sufficiently quickly as $|\mathbf{n}|$ increases that for most orbits $|\beta^{2^k} h_{\mathbf{n}}/\mathbf{n}\cdot\mathbf{\Omega}|$ is small for all $\mathbf{n}$.

Box 3.5 outlines how the so-called **KAM theory** enables one to overcome the problem of small divisors for most tori, and for them construct a convergent series of canonical transformations that yield the angle-action variables of H^β to arbitrarily high accuracy for sufficiently small β.

3.7.2 Trapping by resonances

Figure 3.36, like Figure 3.35, is a surface of section for motion in a flattened isochrone potential, but the axis ratio of the mass distribution that generates the potential is now $q = 0.4$ rather than $q = 0.7$. The consequents of two orbits are shown together with the approximations to the invariant curves of these orbits that one obtains from the angle-action variables of the spherical isochrone potential with (full curves) and without (dotted curves) first-order perturbation theory. The inner full invariant curve is not very far removed from the inner loop of orbital consequents, but the outer full invariant curve does not even have the same shape as the crescent of consequents that is generated by the second orbit. The deviation between the outer full invariant curve and the consequents is an example of **resonant trapping**, a phenomenon intimately connected with the problem of small divisors that was described above.

To understand this connection, consider how the frequencies of orbits in the flattened isochrone potential are changed by first-order perturbation theory. We obtain the new frequencies by differentiating equation (3.296b) with respect to the actions. Figure 3.37 shows the resulting ratio $\Omega_r/\Omega_\vartheta$ as a function of J_ϑ at the energy of Figure 3.36. Whereas $\Omega_r > \Omega_\vartheta$ for all unperturbed orbits, for some perturbed orbit the resonant condition $\Omega_r - \Omega_\vartheta = 0$ is satisfied. Consequently, if we attempt to use equation (3.295) to refine the tori that generate the full curves in Figure 3.36, small divisors will lead to large distortions in the neighborhood of the resonant torus. These distortions will be unphysical, but they are symptomatic of a real physical effect, namely a complete change in the way in which orbital tori are embedded in phase space. The numerical consequents in Figure 3.36, which mark cross-sections through two tori, one before and one after the change in the embedding, make the change apparent: one torus encloses the shell orbit whose single consequent lies along $p_R = 0$, while the other torus encloses the resonant orbit whose single consequent lies near $(R, p_R) = (2.2, 0.38)$.

Small divisors are important physically because they indicate that a perturbation is acting with one sign for a long time. If the effects of a perturbation can accumulate for long enough, they can become important, even if the perturbation is weak. So if $\mathbf{N}\cdot\mathbf{\Omega}$ is small for some $\mathbf{N}$, then the term $h_{\mathbf{N}}$ in the Hamiltonian can have big effects even if it is very small.

Box 3.5: KAM theory

Over the period 1954–1967 Kolmogorov, Arnold and Moser demonstrated that, notwithstanding the problem of small divisors, convergent perturbation series *can* be constructed for Hamiltonians of the form (3.289). The key ideas are (i) to focus on a single invariant torus rather than a complete foliation of phase space by invariant tori, and (ii) to determine at the outset the frequencies $\mathbf{\Omega}$ of the torus to be constructed (Lichtenberg & Lieberman 1992). In particular, we specify that the frequency ratios are far from resonances in the sense that $|\mathbf{n}\cdot\mathbf{\Omega}| > \alpha|\mathbf{n}|^{-\gamma}$ for all $\mathbf{n}$ and some fixed, non-negative numbers α and γ. We map an invariant torus of H^0 with frequencies $\mathbf{\Omega}$ into an invariant torus of H^β by means of the generating function

$$S(\boldsymbol{\theta}^\beta,\mathbf{J}^0) = \boldsymbol{\theta}^\beta\cdot(\mathbf{J}^0+\mathbf{j}) + s^\beta(\boldsymbol{\theta}^\beta,\mathbf{J}^0). \tag{1}$$

This differs from (3.290) by the addition of a term $\boldsymbol{\theta}^\beta\cdot\mathbf{j}$, where $\mathbf{j}$ is a constant of order β. Proceeding in strict analogy with the derivation of equations (3.295) and (3.296b), we find that if the Fourier coefficients of s^β are chosen to be

$$s^\beta_{\mathbf{n}} = -\frac{\beta h_{\mathbf{n}}}{\mathbf{n}\cdot\mathbf{\Omega}} \quad (\mathbf{n}\neq 0), \tag{2}$$

then we obtain a canonical transformation to new coordinates $(\boldsymbol{\theta}',\mathbf{J}')$ in terms of which H^β takes the form (3.296a) with

$$H'(\mathbf{J}') = H^0(\mathbf{J}') + \beta h_0(\mathbf{J}') - \mathbf{j}\cdot\mathbf{\Omega}. \tag{3}$$

We now choose the parameter $\mathbf{j}$ such that the frequencies of H' are still the old frequencies $\mathbf{\Omega}$, which were far from any resonance. That is, we choose $\mathbf{j}$ to be the solution of

$$\beta\frac{\partial h_0}{\partial J_j} = \mathbf{j}\cdot\frac{\partial\mathbf{\Omega}}{\partial J_j} = \sum_i j_i\cdot\frac{\partial^2 H^0}{\partial J_i\partial J_j}. \tag{4}$$

This linear algebraic equation will be soluble provided the matrix of second derivatives of H^0 is non-degenerate. With $\mathbf{j}$ the solution to this equation, the problem posed by H^β in the $(\boldsymbol{\theta}',\mathbf{J}')$ coordinates differs from our original problem only in that the perturbation is now $\mathrm{O}(\beta^2)$. Consequently, a further canonical transformation will reduce the perturbation to $\mathrm{O}(\beta^4)$ and so on indefinitely. From the condition $|\mathbf{n}\cdot\mathbf{\Omega}| > \alpha|\mathbf{n}|^{-\gamma}$ one may show that the series of transformations converges.

We now use this idea to obtain an analytic model of orbits near resonances. Our working will be a generalization of the discussion of orbital trapping at Lindblad resonances in §3.3.3b. For definiteness we shall assume

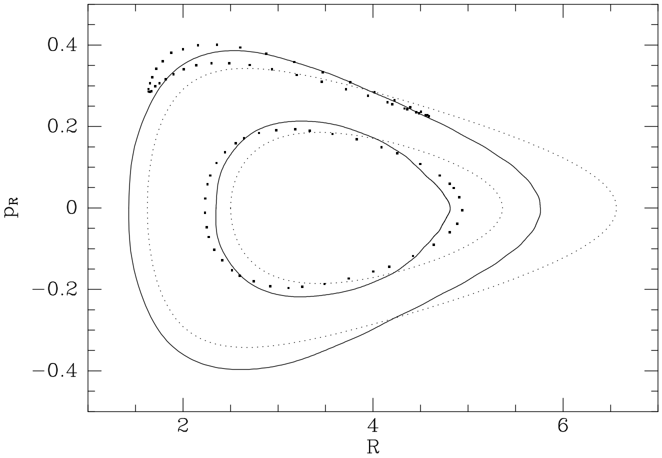


Figure 3.36 The same as Figure 3.35 except that the density distribution generating the potential now has axis ratio $q = 0.4$.

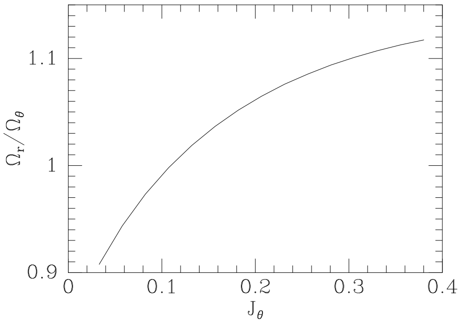


Figure 3.37 The ratio of the frequencies in first-order perturbation theory for a star that moves in a flattened isochrone potential.

that there are three actions and three angles. The resonance of H^0 is characterized by the equation $\mathbf{N}\cdot\mathbf{\Omega} = 0$, and $(\boldsymbol{\theta}, \mathbf{J})$ are angle-action variables for the unperturbed Hamiltonian. Then in the neighborhood of the resonant orbit the linear combination of angle variables $\phi_\mathrm{s} \equiv \mathbf{N}\cdot\boldsymbol{\theta}$ will evolve slowly, and we start by transforming to a set of angle-action variables that includes the **slow angle** ϕ_s. To do so, we introduce new action variables I_s, I_f1, and I_f2 through the generating function

$$S = (\mathbf{N}\cdot\boldsymbol{\theta})I_\mathrm{s} + \theta_1 I_\mathrm{f1} + \theta_2 I_\mathrm{f2}. \tag{3.297}$$

Then (eq. D.93)

$$\begin{aligned} \phi_\mathrm{s} &= \frac{\partial S}{\partial I_\mathrm{s}} = \mathbf{N}\cdot\boldsymbol{\theta} \qquad & J_1 &= \frac{\partial S}{\partial \theta_1} = N_1 I_\mathrm{s} + I_\mathrm{f1} \\ \phi_\mathrm{f1} &= \theta_1 & J_2 &= N_2 I_\mathrm{s} + I_\mathrm{f2} \\ \phi_\mathrm{f2} &= \theta_2 & J_3 &= N_3 I_\mathrm{s}. \end{aligned} \tag{3.298}$$

Since the old actions are functions only of the new ones, H^0 does not acquire any angle dependence when we make the canonical transformation, and the Hamiltonian is of the form

$$H(\boldsymbol{\phi},\mathbf{I}) = H^0(\mathbf{I}) + \beta \sum_{\mathbf{n}} h_{\mathbf{n}}(\mathbf{I})\mathrm{e}^{\mathrm{i}\mathbf{n}\cdot\boldsymbol{\phi}}, \tag{3.299}$$

where it is to be understood that H^0 is a different function of $\mathbf{I}$ than it was of $\mathbf{J}$ and similarly for the dependence on $\mathbf{I}$ of $h_{\mathbf{n}}$. We now argue that any term in the sum that contains either of the **fast angles** $\phi_{\mathrm{f}1}$ and $\phi_{\mathrm{f}2}$ has a negligible effect on the dynamics—these terms give rise to forces that rapidly average to zero. We therefore drop all terms except those with indices that are multiples of $\mathbf{n} = \pm(1,0,0)$, including $\mathbf{n} = \mathbf{0}$. Then our approximate Hamiltonian reduces to

$$H(\boldsymbol{\phi},\mathbf{I}) = H^0(\mathbf{I}) + \beta \sum_k h_k(\mathbf{I})\,\mathrm{e}^{\mathrm{i}k\phi_{\mathrm{s}}}. \tag{3.300}$$

Hamilton's equations now read

$$\begin{gathered} \dot{I}_{\mathrm{s}} = -\mathrm{i}\beta \sum_k k h_k(\mathbf{I})\,\mathrm{e}^{\mathrm{i}k\phi_{\mathrm{s}}} \quad ; \quad \dot{\phi}_{\mathrm{s}} = \Omega_{\mathrm{s}} + \beta \sum_k \frac{\partial h_k}{\partial I_{\mathrm{s}}}\,\mathrm{e}^{\mathrm{i}k\phi_{\mathrm{s}}} \\ \dot{I}_{\mathrm{f}1} = 0 \quad ; \quad \dot{I}_{\mathrm{f}2} = 0, \end{gathered} \tag{3.301}$$

where $\Omega_{\mathrm{s}} \equiv \partial H^0/\partial I_{\mathrm{s}}$. So $I_{\mathrm{f}1}$ and $I_{\mathrm{f}2}$ are two constants of motion and we have reduced our problem to one of motion in the $(\phi_{\mathrm{s}}, I_{\mathrm{s}})$ plane. Eliminating $\mathbf{I}$ between equations (3.298) and (3.301) we find that although all the old actions vary, two linear combinations of them are constant:

$$N_2 J_1 - N_1 J_2 = \text{constant} \quad ; \quad N_3 J_2 - N_2 J_3 = \text{constant}. \tag{3.302}$$

We next take the time derivative of the equation of motion (3.301) for ϕ_{s}. We note that Ω_{s}, but not its derivative with respect to I_{s}, is small because it vanishes on the resonant torus. Dropping all terms smaller than $\mathrm{O}(\beta)$,

$$\ddot{\phi}_{\mathrm{s}} \simeq \frac{\partial \Omega_{\mathrm{s}}}{\partial I_{\mathrm{s}}} \dot{I}_{\mathrm{s}} = -\mathrm{i}\beta \frac{\partial \Omega_{\mathrm{s}}}{\partial I_{\mathrm{s}}} \sum_k k h_k \mathrm{e}^{\mathrm{i}k\phi_{\mathrm{s}}}. \tag{3.303}$$

If we define

$$V(\phi_{\mathrm{s}}) \equiv \beta \frac{\partial \Omega_{\mathrm{s}}}{\partial I_{\mathrm{s}}} \sum_k h_k(\mathbf{I})\,\mathrm{e}^{\mathrm{i}k\phi_{\mathrm{s}}}, \tag{3.304}$$

where $\mathbf{I}$ is evaluated on the resonant torus, then we can rewrite (3.303) as

$$\ddot{\phi}_{\mathrm{s}} = -\frac{\mathrm{d}V}{\mathrm{d}\phi_{\mathrm{s}}}. \tag{3.305}$$

This is the equation of motion of an oscillator. If V were proportional to $\phi_{\rm s}^2$, the oscillator would be harmonic. In general it is an anharmonic oscillator, such as a pendulum, for which $V \propto \cos\phi_{\rm s}$. The oscillator's energy invariant is

$$E_{\rm p} \equiv \tfrac{1}{2}\dot\phi_{\rm s}^2 + V(\phi_{\rm s}). \tag{3.306}$$

V is a periodic function of $\phi_{\rm s}$, so it will have some maximum value $V_{\rm max}$, and if $E_{\rm p} > V_{\rm max}$, $\phi_{\rm s}$ circulates because equation (3.306) does not permit $\dot\phi_{\rm s}$ to vanish. In this case the orbit is not resonantly trapped and the torus is like the ones shown in Figure 3.36 from first-order perturbation theory. If $E_{\rm p} < V_{\rm max}$, the angle variable is confined to the range in which $V \le E_{\rm p}$; the orbit has been trapped by the resonance. On trapped orbits $\phi_{\rm s}$ librates with an amplitude that can be of order unity, and at a frequency of order $\sqrt{\beta}$, while $I_{\rm s}$ oscillates with an amplitude that cannot be bigger than order $\sqrt{\beta}$. Such orbits generate the kind of torus that is delineated by the crescent of numerical consequents in Figure 3.36. We obtain an explicit expression for the resonantly induced change $\Delta I_{\rm s}$ by integrating the equation of motion (3.301) for $I_{\rm s}$:

$$\begin{aligned}\Delta I_{\rm s} &= -\left(\frac{\partial\Omega_{\rm s}}{\partial I_{\rm s}}\right)^{-1}\int {\rm d}\phi_{\rm s}\frac{\partial V/\partial\phi_{\rm s}}{\dot\phi_{\rm s}}\\ &= \pm\left(\frac{\partial\Omega_{\rm s}}{\partial I_{\rm s}}\right)^{-1}\sqrt{2[E_{\rm p}-V(\phi_{\rm s})]}\,,\end{aligned} \tag{3.307}$$

where (3.306) has been used to eliminate $\dot\phi_{\rm s}$.

The full curve in Figure 3.38 shows the result of applying this model of a resonantly trapped orbit to the data depicted in Figure 3.36. Since the model successfully reproduces the gross form of the invariant curve on which the consequents of the trapped orbit lie, we infer that the model has captured the essential physics of resonant trapping. The discrepancies between the full curve and the numerical consequents are attributable to the approximations inherent in the model.

Levitation We now describe one example of an astronomical phenomena that may be caused by resonant trapping of stellar orbits. Other examples are discussed by Tremaine & Yu (2000). In our discussion we shall employ J_r, J_ϑ and J_ϕ to denote the actions of a mildly non-spherical potential that are the natural extensions of the corresponding actions for spherical systems that were introduced in §3.5.2.

The disk of the Milky Way seems to be a composite of two chemically distinct disks, namely the thin disk, to which the Sun belongs, and a thicker, more metal-poor disk (page 13). Sridhar & Touma (1996) have suggested that resonant trapping of the orbits of disk stars may have converted the Galaxy's original thin disk into the thick disk. The theory of hierarchical galaxy formation described in Chapter 9 predicts that the Galaxy was originally dominated by collisionless dark matter, which is not highly concentrated towards the plane. Consequently, the frequency Ω_ϑ at which a

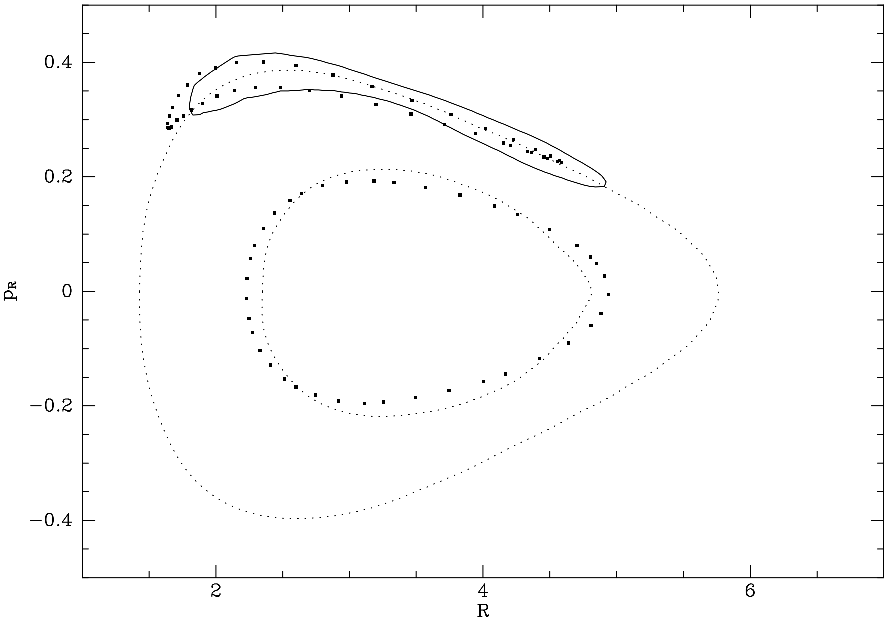


Figure 3.38 Perturbation theory applied to resonant trapping in the flattened isochrone potential. The points are the consequents shown in Figure 3.36, while the full curves in that figure are shown dotted here. The full curve shows the result of using (3.307) to model the resonantly trapped orbit.

star oscillated perpendicular to the plane was originally smaller than the frequency Ω_r of radial oscillations—see equation (3.82). As more and more baryonic material accumulated near the Galaxy's equatorial plane, the ratio $\Omega_\vartheta/\Omega_r$ rose slowly from a value less than unity to its present value. For stars such as the Sun that are on nearly circular orbits within the plane, Ω_r and Ω_ϑ are equal to the current epicycle and vertical frequencies κ and ν, respectively, so now $\Omega_\vartheta/\Omega_r \simeq 2$ (page 167). It follows that the resonant condition $\Omega_r = \Omega_\vartheta$ has at some stage been satisfied for many stars that formed when the inner Galaxy was dark-matter dominated.

Let us ask what happens to a star in the disk as the disk slowly grows and $\Omega_\vartheta/\Omega_r$ slowly increases. At any energy, the first stars to satisfy the resonant condition $\Omega_r = \Omega_\vartheta$ will have been those with the largest values of Ω_ϑ, that is, stars that orbit close to the plane, and have $J_\vartheta \simeq 0$. In an (R, p_R) surface of section, such orbits lie near the zero-velocity curve that bounds the figure (§3.2.2) because J_ϑ increases as one moves in towards the central fixed point on $p_R = 0$. Hence, the resonant condition will first have been satisfied on the zero-velocity curve, and it is here that the resonant island seen in Figure 3.38 first emerged as the potential flattened. As mass accumulated in the disk, the island moved inwards, and, depending on the values of E and L_z, finally disappeared near the central fixed point.

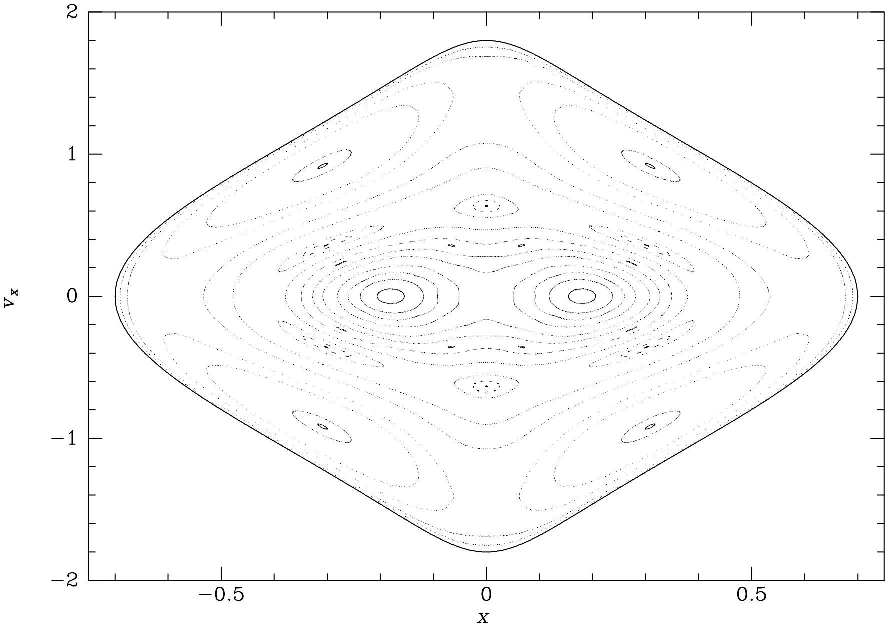


Figure 3.39 A surface of section for motion in Φ_L (eq. 3.103) with $q = 0.6$.

When the advancing edge of a resonant island reaches the star's phase-space location, there are two possibilities: either (a) the star is trapped by the resonance and its phase-space point subsequently moves within the island, or (b) its phase-space point suddenly jumps to the other side of the island. Which of (a) or (b) occurs in a particular case depends on the precise phase of the star's orbit at which the edge reaches it. In practice it is most useful to discard phase information and to consider that either (a) or (b) occurs with appropriate probabilities P_a and $P_\mathrm{b} = 1 - P_\mathrm{a}$. The value of P_a depends on the speed with which the island is growing relative to the speed with which its center is moving (Problem 3.43); it is zero if the island is shrinking.

We have seen that the resonant island associated with $\Omega_r = \Omega_\vartheta$ first emerged on the zero-velocity curve, which in a thin disk is highly populated by stars. Most of these stars were trapped as the island grew. They then moved with the island as the latter moved in towards the central fixed point. The stars were finally released as the island shrank somewhere near that point. The net effect of the island's transitory existence is to convert radial action to latitudinal action, thereby shifting stars from eccentric, planar orbits to rather circular but highly inclined ones. Hence, a hot thin disk could have been transformed into a thick disk.

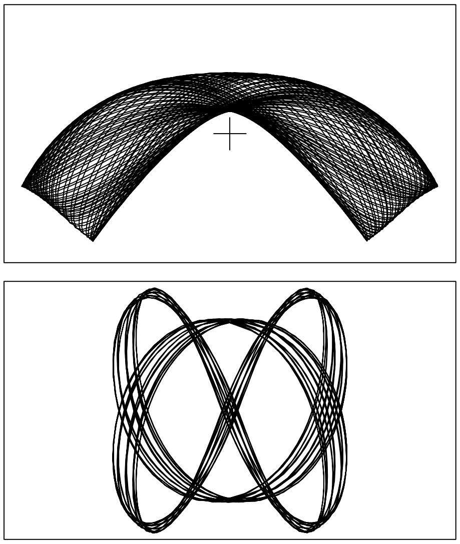


Figure 3.40 The appearance in real space of a banana orbit (top) and a fish orbit (bottom). In the upper panel the cross marks the center of the potential. Resonant box orbits of these types are responsible for the chains of islands in Figure 3.39. The banana orbits generate the outer chain of four islands, and the fish orbits the chain of six islands further in.

3.7.3 From order to chaos

Figure 3.39 is a surface of section for motion in the planar barred potential Φ_L that is defined by equation (3.103) with $q = 0.6$ and $R_{\rm c} = 0.14$. It should be compared with Figures 3.9 and 3.12, which are surfaces of section for motion in Φ_L for more nearly spherical cases, with $q = 0.9$ and 0.8. In Figure 3.39 one sees not only the invariant curves of loop and box orbits that fill the other two figures, but also a number of "islands": a set of four large islands occupies much of the outer region, while a set of six islands of varying sizes is seen further in. In the light of our discussion of resonant trapping, it is natural to refer to the orbits that generate these islands as resonantly trapped box orbits. Figure 3.40 shows what these orbits look like in real space. We see that the outer islands are generated by "banana" orbits in which the x- and y-oscillations are trapped in a $\Omega_x{:}\Omega_y = 1{:}2$ resonance (the star oscillates through one cycle left to right while oscillating through two cycles up and down). Similarly, the inner chain of six islands is associated with a "fish" orbit that satisfies the resonance condition $\Omega_x{:}\Omega_y = 2{:}3$.

The islands in Figure 3.39 can be thought of as orbits in some underlying integrable Hamiltonian H^0 that are trapped by a resonance arising from a perturbation. This concept lacks precision because we do not know what H^0 actually is. In particular, Hamiltonians of the form $H_q(\mathbf{x},\mathbf{v}) = \frac{1}{2}v^2 + \Phi_L(\mathbf{x})$ are probably not integrable for any value of the axis ratio q other than unity. Therefore, we cannot simply assume that $H^0 = H_{0.8}$, say. On the other hand, Figure 3.12, which shows the surface of section for $q = 0.8$, contains no resonant islands—all orbits are either boxes or loops—which we know from our study of Stäckel potentials in §3.5.4 is compatible with an integrable potential. So we can *define* an integrable Hamiltonian H^0 that differs very little from $H_{0.8}$ as follows. On each of the invariant tori that appears in Figure 3.12 we set $H^0 = H_{0.8}$, and at a general phase-space point we obtain the value of H^0 by a suitable interpolation scheme from nearby points at which $H^0 = H_{0.8}$.

The procedure we have just described for defining H^0 (and thus the perturbation $h = H - H^0$) suffers from the defect that it is arbitrary: why start from the invariant tori of $H_{0.8}$ rather than $H_{0.81}$ or some other Hamiltonian? A numerical procedure that might be considered less arbitrary has been described by Kaasalainen & Binney (1994). In any event, it is worth bearing in mind in the discussion that follows that H^0 and h are not uniquely defined, and one really ought to demonstrate that for a given H the islands that are predicted by perturbation theory are reasonably independent of H^0. As far as we know, no such demonstration is available.

If we accept that the island chains in Figure 3.39 arise from box orbits that are resonantly trapped by some perturbation h on a Stäckel-like Hamiltonian H^0, two questions arise. First, "are box orbits trapped around resonances other than the 1:2 and 2:3 resonances that generate the banana and fish orbits of Figure 3.40?" Certainly infinitely many resonances are available to trap orbits because as one moves along the sequence of box orbits from thin ones to fat ones, the period of the y-oscillations is steadily growing in parallel with their amplitude, while the period of the x-oscillations is diminishing for the same reason.[27] In fact, the transition to loop orbits can be associated with resonant trapping by the 1:1 resonance, so between the banana orbits and the loop orbits there is not only the 2:3 resonance that generates the fishes, but also the 4:5, 5:6, ..., resonances. In the potential Φ_L on which our example is based, the width of the region in phase space in which orbits are trapped by the $m{:}n$ resonance diminishes rapidly with $|m + n|$ and the higher-order resonances are hard to trace in the surface of section—but the 4:5 resonance can be seen in Figure 3.39.

The second question is "do resonances occur within resonant islands?" Consider the case of the banana orbits shown in Figure 3.40 as an example. Motion along this orbit is quasiperiodic with two independent frequencies. One independent frequency Ω_b is associated with motion along the bow-shaped closed orbit that runs through the heart of the banana, while the other is the frequency of libration Ω_l about this closed orbit. The libration frequency decreases as one proceeds along the sequence of banana orbits from thin ones to fat ones, so infinitely many resonant conditions $\Omega_\mathrm{b}{:}\Omega_l = m{:}n$ will be satisfied within an island of banana orbits. In the case of Φ_L there is no evidence that any of these resonances traps orbits, but in another case we might expect trapping to occur also within families of resonantly trapped orbits.

This discussion is rather disquieting because it implies that the degree to which resonant trapping causes the regular structure of phase space inherited from the underlying integral potential H^0 to break up into islands depends

[27] The period of a nonlinear oscillator almost always increases with amplitude.

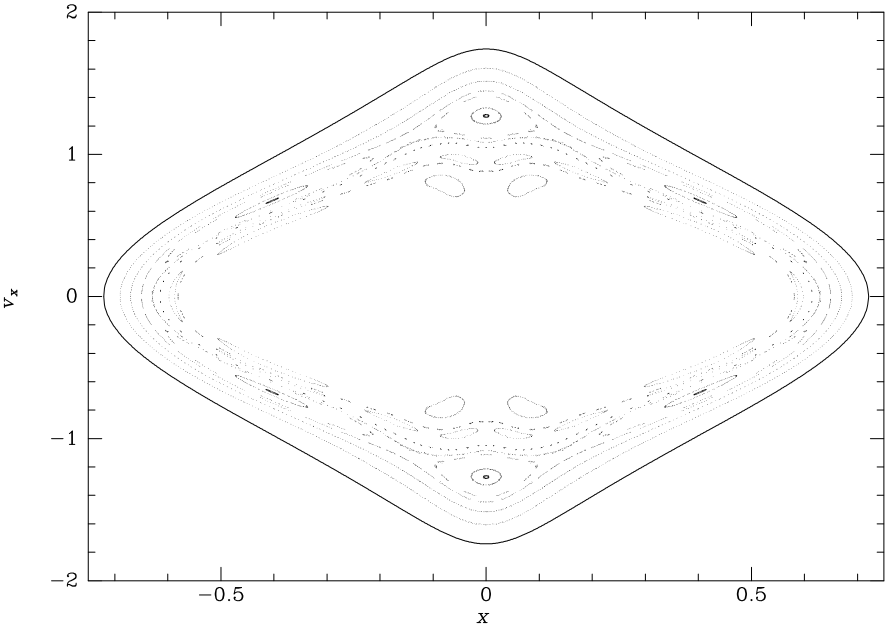


Figure 3.41 Surface of section for motion in the potential $\Phi_{\rm N}$ of equation (3.309) with $R_{\rm e} = 3$. The inner region has been blanked out and is shown in expanded form in Figure 3.42.

on the detailed structure of the perturbation h. Since we have no unique way of defining h we cannot compute its Fourier coefficients and cannot predict how important islands will be.

We illustrate this point by examining motion in a potential that is closely related to Φ_L in which resonant trapping is *much* more important (Binney 1982). In polar coordinates equation (3.103) for Φ_L reads

$$\Phi_L(R,\phi) = \tfrac{1}{2}v_0^2 \ln\left[R_{\rm c}^2 + \tfrac{1}{2}R^2(q^{-2}+1) - \tfrac{1}{2}R^2(q^{-2}-1)\cos 2\phi\right]. \qquad (3.308)$$

The potential

$$\begin{aligned}\Phi_{\rm N}(R,\phi) = \tfrac{1}{2}v_0^2 \ln\Bigg[R_{\rm c}^2 + \tfrac{1}{2}R^2(q^{-2}+1) - \tfrac{1}{2}R^2(q^{-2}-1)\cos 2\phi \\ - \frac{R^3}{R_{\rm e}}\cos 2\phi\Bigg],\end{aligned} \qquad (3.309)$$

where $R_{\rm e}$ is a constant, differs from Φ_L only by the addition of $(R^3/R_{\rm e})\cos 2\phi$ to the logarithm's argument. For $R \ll R_{\rm e}$ this term is unimportant, but as R grows it makes the isopotential curves more elongated. Let us set $R_{\rm e} = 3$, $R_{\rm c} = 0.14$, and $q = 0.9$, and study the surface of section generated by orbits

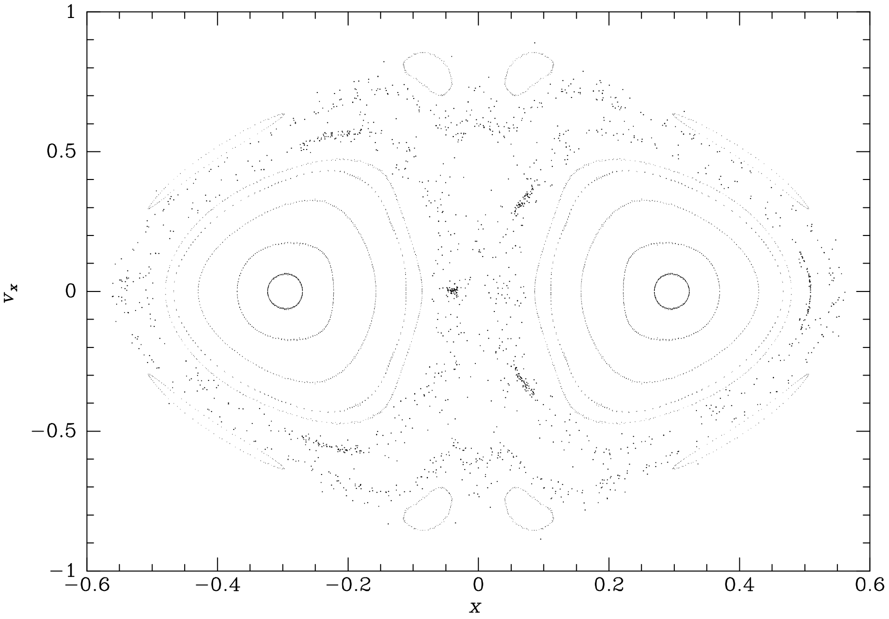


Figure 3.42 The inner part of the surface of section shown in Figure 3.41—the chain of eight islands around the edge is the innermost chain in Figure 3.41. In the gap between this chain and the bull's-eyes are the consequents of two irregular orbits.

in $\Phi_{\rm N}$ that is most nearly equivalent to the surface of section for Φ_L with the same values of $R_{\rm c}$ and q that is shown in Figure 3.9. Figure 3.41 shows the outer part of this surface of section. Unlike Figure 3.9 it shows several chains of islands generated by resonantly trapped box orbits. The individual islands are smaller than those in Figure 3.39, and the regions of untrapped orbits between chains of islands are very thin. Figure 3.42 shows the inner part of the same surface of section. In the gap between the region of regular box orbits that is shown in Figure 3.41 and the two bull's-eyes associated with loop orbits, there is an irregular fuzz of consequents. These consequents belong to just two orbits but they do not lie on smooth curves; they appear to be randomly scattered over a two-dimensional region. Since the gap within which these consequents fall lies just on the boundary of the loop-dominated region, we know that it contains infinitely many resonant box orbits. Hence, it is natural to conclude that the breakdown of orbital regularity, which the random scattering of consequents betrays, is somehow caused by more than one resonance simultaneously trying to trap an individual orbit. One says that the orbits have been made irregular by **resonance overlap**.

(a) Irregular orbits We now consider in more detail orbits whose consequents in a surface of section do not lie on a smooth curve, but appear to be irregularly sprinkled through a two-dimensional region. If we take the

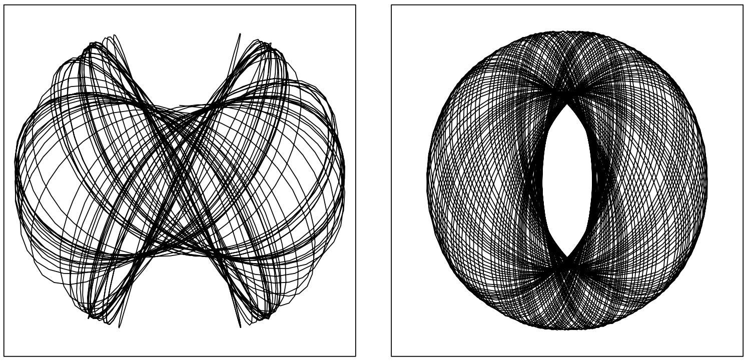


Figure 3.43 Two orbits from the surface of section of Figure 3.42. The left orbit is not quasiperiodic, while the right one is.

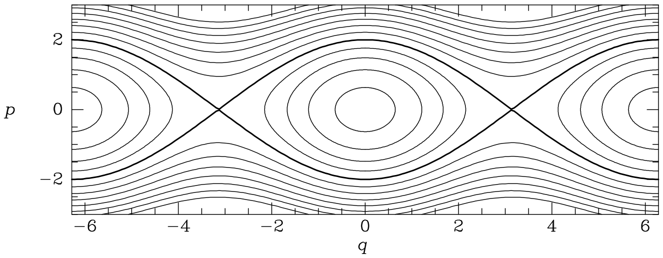


Figure 3.44 Trapped and circulating orbits in a phase plane. The homoclinic orbit, shown by the heavy curve, divides the trapped orbits, which form a chain of islands, from the circulating orbits, whose consequents lie on the wavy lines at top and bottom.

Fourier transform of the time dependence of some coordinate, for example $x(t)$, along such an orbit, we will find that the orbit is not quasiperiodic; the Fourier transform $X(\omega)$ (eq. B.69) has contributions from frequencies that are not integer linear combinations of two or three fundamental frequencies. Figure 3.43 shows the appearance in real space of an orbit that is not quasiperiodic (left) and one that is (right). The lack of quasiperiodicity gives the orbit a scruffy, irregular appearance, so orbits that are not quasiperiodic are called **chaotic** or **irregular orbits**.

There are generally some irregular orbits at the edge of a family of resonantly trapped orbits. Figure 3.44 is a sketch of a surface of section through such a region of phase-space when all orbits are quasiperiodic. The islands formed by the trapped orbits touch at their pointed ends and there are invariant curves of orbits that circulate rather than librate coming right up to these points. The points at which the islands touch are called **hyperbolic fixed points** and the invariant curves that pass through these points are generated by **homoclinic orbits**. In the presence of irregular orbits, the islands of trapped orbits do not quite touch and the invariant curves of the circulating orbits do not reach right into the hyperbolic fixed point. Consequently there is space between the resonant islands and the region of the circulating orbits. Irregular orbits fill this space.

A typical irregular orbit alternates periods when it is resonantly trapped with periods of circulation. Consequently, if one Fourier transforms $x(t)$ over an appropriate time interval, the orbit may appear quasiperiodic, but the fundamental frequencies that would be obtained from the transform by the method of Box 3.6 would depend on the time interval chosen for Fourier transformation.

If the islands in a chain are individually small, it can be very hard to decide whether an orbit is librating or circulating, or doing both on an irregular pattern.

When it is available, a surface of section is the most effective way of diagnosing the presence of resonantly trapped and irregular orbits. Unfortunately, surfaces of section can be used to study three-dimensional orbits only when an analytic integral other than the Hamiltonian is known, as in the case of orbits in an axisymmetric potential (§3.2). Two other methods are available to detect irregular orbits when a surface of section cannot be used.

(b) Frequency analysis By numerically integrating the equations of motion from some initial conditions, we obtain time series $x(t)$, $y(t)$, etc., for each of the phase-space coordinates. If the orbit is regular, these time series are equivalent to those obtained by substituting $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \boldsymbol{\Omega} t$ in the Fourier expansions (3.191) of the coordinates. Hence, the frequencies Ω_i may be obtained by Fourier transforming the time series and identifying the various linear combinations $\mathbf{n}\cdot\boldsymbol{\Omega}$ of the fundamental frequencies that occur in the Fourier transform (Box 3.6; Binney & Spergel 1982). If a single system of angle-action variables covers the entire phase space (as in the case of Stäckel potentials), the actions J_i of the orbit that one obtains from a given initial condition $\mathbf{w}$ are continuous functions $J(\mathbf{w})$ of $\mathbf{w}$, so the frequencies $\Omega_i = \partial H/\partial J_i$ are also continuous functions of $\mathbf{w}$. Consequently, if we choose initial conditions $\mathbf{w}_\alpha$ at the nodes of some regular two-dimensional grid in phase space, the frequencies will vary smoothly from point to point on the grid. If, by contrast, resonant trapping is important, the actions of orbits will sometimes change discontinuously between adjacent grid points, because one orbit will be trapped, while the next is not. Discontinuities in $\mathbf{J}$ give rise to discontinuities in $\boldsymbol{\Omega}$. Moreover, the resonance that is entrapping orbits will be apparent from the ratios $r_a \equiv \Omega_2/\Omega_1$ and $r_b \equiv \Omega_3/\Omega_1$. Hence a valuable way of probing the structure of phase space is to plot a dot at (r_a, r_b) for each orbit obtained by integrating from a regular grid of initial conditions $\mathbf{w}_\alpha$ (Laskar 1990; Dumas & Laskar 1993).

Figure 3.45 shows an example of such a plot of frequency ratios. The orbits plotted were integrated in the potential

$$\Phi(\mathbf{x}) = \tfrac{1}{2}\ln[x^2 + (y/0.9)^2 + (z/0.7)^2 + 0.1]. \qquad (3.310)$$

Box 3.6: Numerical determination of orbital frequencies

The determination of orbital frequencies Ω_i from a numerically integrated orbit is not entirely straightforward because (i) the orbit is integrated for only a finite time interval $(0, T)$, and (ii) the function $x(t)$ is sampled only at discrete times $t_0 = 0, \dots, t_{K-1} = T$, which we shall assume to be equally spaced. Let $\Delta = t_{i+1} - t_i$. Then a "line" $X\mathrm{e}^{\mathrm{i}\omega t}$ in $x(t)$ contributes to the discrete Fourier transform (Appendix G) an amount

$$\begin{aligned}\hat{x}_p &= X\sum_{k=0}^{K-1}\mathrm{e}^{\mathrm{i}k\Delta(\omega-\omega_p)}\\ &= X\mathrm{e}^{\mathrm{i}\alpha u}\frac{\sin\pi u}{\sin(\pi u/K)},\end{aligned}\qquad\text{where}\quad\begin{cases}\omega_p \equiv \dfrac{2\pi p}{K\Delta},\\ u \equiv K\Delta(\omega-\omega_p)/(2\pi),\\ \alpha \equiv \pi(K-1)/K.\end{cases}\tag{1}$$

$|\hat{x}_p|$ is large whenever the sine in the denominator vanishes, which occurs when $\omega_p \simeq \omega + 2\pi m/\Delta$, where m is any integer. Thus peaks can arise at frequencies far from ω; a peak in $|\hat{x}_p|$ that is due to a spectral line far removed from ω is called an **alias** of the line. Near to a peak we can make the approximation $\sin(\pi u/K) \simeq \pi u/K$, so $|\hat{x}_p|$ declines with distance u from the peak only as u^{-1}.

Orbital frequencies can be estimated by fitting equation (1) to the data and thus determining ω. The main difficulty with this procedure is confusion between spectral lines—this confusion can arise either because two lines are nearby, or because a line has a nearby alias. One way to reduce this confusion is to ensure a steeper falloff than u^{-1} by multiplying the original time sequence by a "window" function $w(t)$ that goes smoothly to zero at the beginning and end of the integration period (Press et al. 1986; Laskar 1990). Alternatively, one can identify peaks in the second difference of the spectrum, defined by $\hat{x}''_p = \hat{x}_{p+1} + \hat{x}_{p-1} - 2\hat{x}_p$. One can show that for $u/K \ll 1$ the contribution to $\hat{x}''_p$ of a line is

$$\hat{x}''_p = \frac{2XK}{\pi}\frac{\mathrm{e}^{\mathrm{i}\alpha u}\sin\pi u}{u(u^2-1)},\tag{2}$$

which falls off as u^{-3}. The frequency, etc., of the line can be estimated from the ratio of the $\hat{x}''_p$ on either side of the line's frequency.

Ω_i was defined to be the non-zero frequency with the largest amplitude in the spectrum of the ith coordinate, and 10 000 orbits were obtained by dropping particles from a grid of points on the surface $\Phi(\mathbf{x}) = 0.5$. Above and to the right of the center of the figure, the points are organized into regular ranks that reproduce the grid of initial conditions in slightly distorted form.

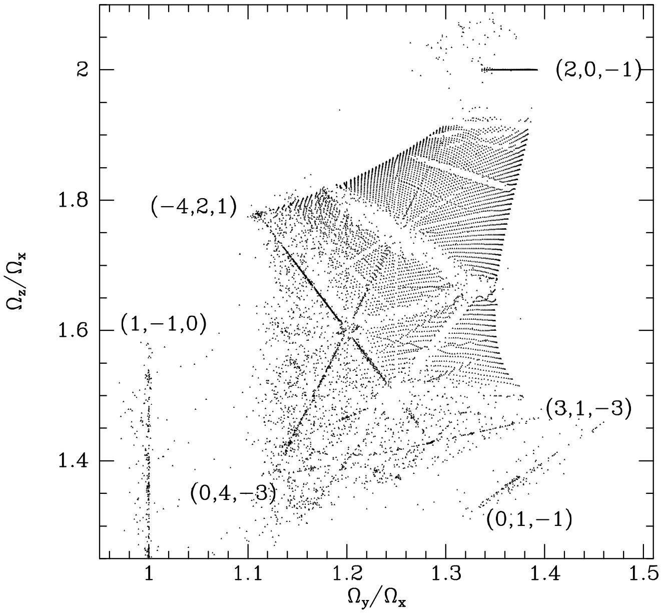


Figure 3.45 The ratios of orbital frequencies for orbits integrated in a three-dimensional non-rotating bar potential.

We infer that resonant trapping is unimportant in the phase-space region sampled by these initial conditions. Running through the ranks we see several depopulated lines, while both within the ranks and beyond other lines are conspicuously heavily populated: orbits that have been resonantly trapped produce points that lie along these lines. The integers n_i in the relevant resonant condition $\mathbf{n}\cdot\mathbf{\Omega}=0$ are indicated for some of the lines.

In some parts of Figure 3.45, for example the lower left region, the grid of initial conditions has become essentially untraceable. The disappearance of the grid indicates that irregular motion is important. In fact, the frequencies Ω_i are not well defined for an irregular orbit, because its time series, $x(t)$, $y(t)$, etc., are not quasiperiodic. When software designed to extract the frequencies of regular orbits is used on a time series that is not quasiperiodic, the frequencies returned vary erratically from one initial condition to the next and the resulting points in the plane of frequency ratios scatter irregularly.

(c) Liapunov exponents If we integrate Hamilton's equations for some time t, we obtain a mapping $\mathbf{H}_t$ of phase space onto itself. Let $\mathbf{H}_t$ map the phase space point $\mathbf{w}_0$ into the point $\mathbf{w}_t$. Points near $\mathbf{w}_0$ will be mapped to points that lie near $\mathbf{w}_t$, and if we confine our attention to a sufficiently small region around $\mathbf{w}_0$, we may approximate $\mathbf{H}_t$ by a linear map of the neighborhood of $\mathbf{w}_0$ into a neighborhood of $\mathbf{w}_t$. We now determine this map. Let $\mathbf{w}'_0$ be a point near $\mathbf{w}_0$, and $\delta\mathbf{w}(t)=\mathbf{H}_t\mathbf{w}'_0-\mathbf{H}_t\mathbf{w}_0$ be the difference between the phase-space coordinates of the points reached by integrating Hamilton's equations for time t from the initial conditions $\mathbf{w}'_0$ and $\mathbf{w}_0$. Then the equations of motion of the components of $\delta\mathbf{w}$ are

$$\begin{aligned}\dot{\delta\mathbf{x}} &= \left(\frac{\partial H}{\partial \mathbf{v}}\right)_{\mathbf{w}'_t} - \left(\frac{\partial H}{\partial \mathbf{v}}\right)_{\mathbf{w}_t} \simeq \left(\frac{\partial^2 H}{\partial \mathbf{w}\partial \mathbf{v}}\right)_{\mathbf{w}_t} \cdot \delta\mathbf{w} \\ \dot{\delta\mathbf{v}} &= -\left(\frac{\partial H}{\partial \mathbf{x}}\right)_{\mathbf{w}'_t} + \left(\frac{\partial H}{\partial \mathbf{x}}\right)_{\mathbf{w}_t} \simeq -\left(\frac{\partial^2 H}{\partial \mathbf{w}\partial \mathbf{x}}\right)_{\mathbf{w}_t} \cdot \delta\mathbf{w},\end{aligned} \tag{3.311}$$

where the approximate equality in each line involves approximating the first derivatives of H by the leading terms in their Taylor series expansions. Equations (3.311) are of the form

$$\frac{\mathrm{d}\delta\mathbf{w}}{\mathrm{d}t} = \mathbf{M}_t \cdot \delta\mathbf{w} \quad \text{where} \quad \mathbf{M}_t \equiv \begin{pmatrix} \dfrac{\partial^2 H}{\partial \mathbf{x}\partial \mathbf{v}} & \dfrac{\partial^2 H}{\partial \mathbf{v}\partial \mathbf{v}} \\ -\dfrac{\partial^2 H}{\partial \mathbf{x}\partial \mathbf{x}} & -\dfrac{\partial^2 H}{\partial \mathbf{v}\partial \mathbf{x}} \end{pmatrix}. \tag{3.312}$$

For any initial vector $\delta\mathbf{w}_0$ these equations are solved by $\delta\mathbf{w}_t = \mathbf{U}_t \cdot \delta\mathbf{w}_0$, where $\mathbf{U}_t$ is the matrix that solves

$$\frac{\mathrm{d}\mathbf{U}_t}{\mathrm{d}t} = \mathbf{M}_t \cdot \mathbf{U}_t. \tag{3.313}$$

We integrate this set of ordinary coupled linear differential equations from $\mathbf{U}_0 = \mathbf{I}$ in parallel with Hamilton's equations of motion for the orbit. Then we are in possession of the matrix $\mathbf{U}_t$ that describes the desired linear map of a neighborhood of $\mathbf{w}_0$ into a neighborhood of $\mathbf{w}_t$. We perform a "singular-value decomposition" of $\mathbf{U}_t$ (Press et al. 1986), that is we write it as a product $\mathbf{U}_t = \mathbf{R}_2 \cdot \mathbf{S} \cdot \mathbf{R}_1$ of two orthogonal matrices $\mathbf{R}_i$ and a diagonal matrix $\mathbf{S}$.[28] $\mathbf{U}_t$ conserves phase-space volume (page 803), so it never maps any vector to zero and the diagonal elements of $\mathbf{S}$ are all non-zero. In fact they are all positive because $\mathbf{U}_t$ evolves continuously from the identity, and their product is unity. A useful measure of the amount by which $\mathbf{U}_t$ shears phase space is the magnitude s of the largest element of $\mathbf{S}$. The **Liapunov exponent** of the orbit along which (3.313) has been integrated is defined to be

$$\lambda = \lim_{t\to\infty} \frac{\ln s}{t}. \tag{3.314}$$

[28] Any linear transformation of an N-dimensional vector space can be decomposed into a rotation, a rescaling in N perpendicular directions, and another rotation. $\mathbf{R}_1$ rotates axes to the frame in which the coordinate directions coincide with the scaling directions. $\mathbf{S}$ effects the rescaling. $\mathbf{R}_2$ first rotates the coordinate directions back to their old values and then effects whatever overall rotation is required.

Since the scaling s is dimensionless, the Liapunov exponent λ has dimensions of a frequency. In practice one avoids integrating (3.313) for long times because numerical difficulties would be encountered once the ratio of the largest and smallest numbers on the diagonal of $\mathbf{S}$ became large. Instead one integrates along the orbit for some time t_1 to obtain a value s_1, and then sets $\mathbf{U}_t$ back to the identity and continues integrating for a further time t_2 to obtain s_2, after which $\mathbf{U}_t$ is again set to the identity before the integration is continued. After N such steps one estimates λ from

$$\lambda \simeq \frac{\sum_i^N \ln s_i}{\sum_i^N t_i}. \tag{3.315}$$

Using this procedure one finds that along a regular orbit $\lambda \to 0$, while along an irregular orbit λ is non-zero.

Angle-action variables enable us to understand why λ is zero for a regular orbit. A point near $\mathbf{w}_0$ will have angles and actions that differ from those of $\mathbf{w}_0$ by small amounts $\delta\theta_i$, δJ_i. The action differences are invariant as we move along the orbit, while the angle differences increase linearly in time due to differences in the frequencies Ω_i of the orbits on which our initial point and $\mathbf{w}_0$ lie. Consequently, the scalings s_i associated with angle differences increase linearly in time, and, by (3.314), the Liapunov exponent is $\lambda = \lim_{t\to\infty} t^{-1}\ln t = 0$.

If the Liapunov exponent of an orbit is non-zero, the largest scaling factor s must increase exponentially in time. Thus in this case initially neighboring orbits diverge exponentially in time. It should be noted, however, that this exponential divergence holds only so long as the orbits remain close in phase space: the definition of the Liapunov exponent is in terms of the linearized equations for orbital perturbations. The approximations involved in deriving these equations will soon be violated if the solutions to the equations are exponentially growing. Hence, we cannot conclude from the fact that an orbit's Liapunov exponent is non-zero that an initially neighboring orbit will necessarily stray far from the original orbit.

3.8 Orbits in elliptical galaxies

Elliptical galaxies nearly always have cusps in their central density profiles in which $\rho \sim r^{-\alpha}$ with $0.3 \lesssim \alpha \lesssim 2$ (BM §4.3.1). Black holes with masses $\sim 0.2\%$ of the mass of the visible galaxy are believed to reside at the centers of these cusps (§1.1.6 and BM §11.2.2). Further out the mass distributions of many elliptical galaxies are thought to be triaxial (BM §4.3.3). These features make the orbital dynamics of elliptical dynamics especially rich, and illustrate aspects of galaxy dynamics that we have already discussed in this chapter (Merritt & Fridman 1996; Merritt & Valluri 1999).

3.8.1 The perfect ellipsoid

A useful basic model of the orbital dynamics of a triaxial elliptical galaxy is provided by extensions to three dimensions of the two-dimensional Stäckel potentials of §3.5.4 (de Zeeuw 1985). The simplest three-dimensional system that generates a Stäckel potential through Poisson's equation is the **perfect ellipsoid**, in which the density is given by

$$\rho(\mathbf{x}) = \frac{\rho_0}{(1+m^2)^2} \quad \text{where} \quad m^2 \equiv \frac{x^2 + (y/q_1)^2 + (z/q_2)^2}{a_0^2}. \tag{3.316}$$

In this formula q_1 and q_2 are the axis ratios of the ellipsoidal surfaces of constant density, and a_0 is a scale length. At radii significantly smaller than a_0, the density is approximately constant, while at $r \gg a_0$ the density falls off $\propto r^{-4}$. Since these asymptotic forms differ from those characteristic of elliptical galaxies, we have to expect the orbital structures of real galaxies to differ in detail from that of the perfect ellipsoid, but nevertheless the model exhibits much of the orbital structure seen in real elliptical galaxies.

By an analysis similar to that used in §3.5.4 to explore the potential of a planar bar, one can show that the perfect ellipsoid supports four types of orbit. Figure 3.46 depicts an orbit of each type. At top left we have a box orbit. The key feature of a box orbit is that it touches the isopotential surface for its energy at its eight corners. Consequently, the star comes to rest for an instant at these points; a box orbit is conveniently generated numerically by releasing a star from rest on the equipotential surface. The potential's longest axis emerges from the orbit's convex face. The other three orbits are all **tube orbits**: stars on these orbits circulate in a fixed sense around the hole through the orbit's center, and are never at rest. The most important tube orbits are the short-axis loops shown at top right, which circulate around the potential's shortest axis. These orbits are mildly distorted versions of the orbits that dominate the phase space of a flattened axisymmetric potential. The tube orbits at the bottom of Figure 3.46 are called outer (left) and inner long-axis tube orbits, and circulate around the longest axis of the potential. Tube orbits around the intermediate axis are unstable. All these orbits can be quantified by a single system of angle-action coordinates $(J_\lambda, J_\mu, J_\nu)$ that are generalizations of the angle-action coordinates for spherical potentials $(J_r, J_\vartheta, J_\phi)$ of Table 3.1 (de Zeeuw 1985).

3.8.2 Dynamical effects of cusps

The most important differences between a real galactic potential and the best-fitting Stäckel potential are at small radii. Box orbits, which alone penetrate to arbitrarily small radii, are be most affected by these differences. The box orbits of a given energy form a two-parameter family: the parameters can be taken to be an orbit's axis ratios. Resonant relations $\mathbf{n} \cdot \mathbf{\Omega} = 0$ between the fundamental frequencies of an orbit are satisfied at various points

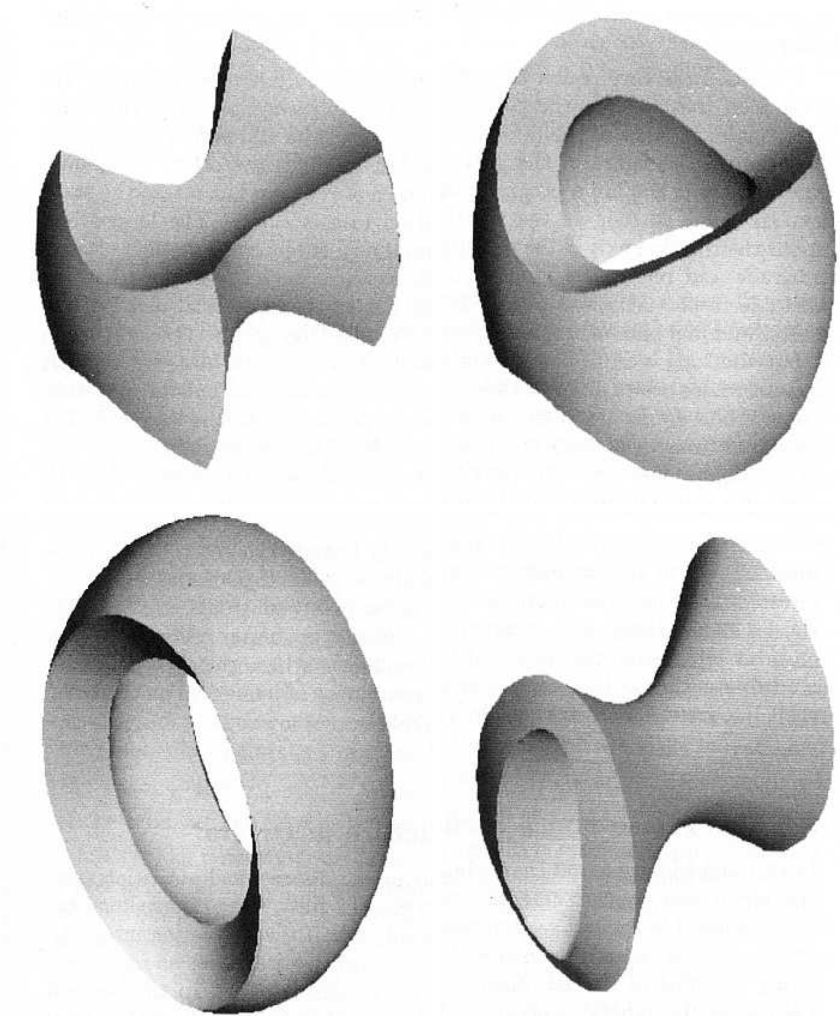


Figure 3.46 Orbits in a non-rotating triaxial potential. Clockwise from top left: (a) box orbit; (b) short-axis tube orbit; (c) inner long-axis tube orbit; (d) outer long-axis tube orbit. From Statler (1987), by permission of the AAS.

in parameter space, but in a Stäckel potential none of these resonances traps other orbits. We expect perturbations to cause some resonances to become trapping. Hence it is no surprise to find that in potentials generated by slightly cusped mass distributions, significant numbers of orbits are trapped by resonances. (In Figure 3.45 we have already encountered extensive resonant trapping of box orbits in a triaxial potential that differs from a Stäckel potential.)

A regular orbit on which the three angle variables satisfy the condition $\mathbf{n} \cdot \mathbf{\Omega} = 0$ is a two-dimensional object since its three actions are fixed, and one of its angles is determined by the other two. Consequently, the orbit occupies a surface in real space. A generic resonantly trapped orbit is a three-dimensional structure because it has a finite libration amplitude around the resonant orbit. In practice the amplitude of the libration is usually small, with the result that the orbit forms a sheet of small but finite thickness around the resonant orbit. It is found that stable resonant box orbits are **centrophobic**, that is, they avoid the galactic center (Merritt & Valluri 1999).

Steepening the cusp in the galaxy's central density profile enhances the difference between the galactic potential and the best-fitting Stäckel model and thus the importance of resonances. More and more resonances overlap (§3.7.3) and the fraction of irregular orbits increases.

The existence of large numbers of irregular orbits in elliptical galaxies is likely to have important but imperfectly understood astronomical implications because irregular orbits display a kind of creep or diffusion. To understand this phenomenon, imagine that there is a clean distinction between regular and irregular regions of $2N$-dimensional phase space. The regular region is occupied by regular orbits and is strictly off-limits to any irregular orbit, while the irregular region is off-limits to regular orbits. However, while each regular orbit is strictly confined to its N-dimensional torus and never trespasses on the territory of a different regular orbit, over time an irregular orbit explores at least some of the irregular region of phase space. In fact, the principal barrier to an irregular orbit's ability to wander is walls formed by regular orbits. In the case $N = 2$ of two-dimensional motion, the energetically accessible part of phase space is three-dimensional, while the walls formed by regular orbits are two-dimensional. Hence such a wall can completely bound some portion of irregular phase space, and forever exclude an irregular orbit from part of irregular phase space. In the case $N = 3$ that is relevant for elliptical galaxies, the energetically accessible region of phase space is five-dimensional while the wall formed by a regular orbit is three-dimensional. Since the boundary of a five-dimensional volume is a four-dimensional region, it is clear that no regular orbit can divide the irregular region of phase space into two. Hence, it is believed that given enough time an irregular orbit with $N \geq 3$ degrees of freedom will eventually visit every part of the irregular region of phase space.

The process by which irregular orbits wander through phase space is called **Arnold diffusion** and is inadequately understood. Physically, it probably involves repeated trapping by a multitude of high-order resonances. In elliptical galaxies and the bars of barred disk galaxies, the rate of Arnold diffusion may be comparable to the Hubble time and could be a major factor in determining the rate of galactic evolution.

If the timescale associated with Arnold diffusion were short enough, galaxy models would need to include only one irregular orbit. The phase-space density $f_{\rm irr}$ contributed by this orbit would be the same at all points on the energy hypersurface $H(\mathbf{x},\mathbf{v}) = E$ except in the regular region of phase space, where $f_{\rm irr}$ would vanish.[29] It is not yet clear how galaxy modeling is best done when the timescale for Arnold diffusion is comparable to the Hubble time.

3.8.3 Dynamical effects of black holes

Introducing even a small black hole at the center of a triaxial galaxy that has a largely regular phase space destroys much of that regularity. There is a simple physical explanation of this phenomenon (Gerhard & Binney 1985; Merritt & Quinlan 1998).

Consider a star on the box orbit shown at top left in Figure 3.46. Each crossing time the star passes through the orbit's waist on an approximately rectilinear trajectory, and is deflected through some angle $\theta_{\rm defl}$ by the black hole's gravitational field. If M is the mass of the hole, and v and b are, respectively, the speed and the distance from the galactic center at which the star would have passed the waist had the hole not deflected it, then from equation (3.52) we have that

$$\theta_{\rm defl} = 2\tan^{-1}\left(\frac{GM}{bv^2}\right). \tag{3.317}$$

The speed v will be similar for all passages, but the impact parameter b will span a wide range of values over a series of passages. For any value of M, no matter how small, there is a chance that b will be small enough for the star to be scattered onto a significantly different box orbit.

The tensor virial theorem (§4.8.3) requires that the velocity dispersion be larger parallel to the longest axis of a triaxial system than in the perpendicular directions. Repeated scattering of stars by a nuclear black hole will tend to make the velocity dispersion isotropic, and thus undermine the orbital support for the triaxiality of the potential. If the potential loses its triaxiality, angular momentum will become a conserved quantity, and every star will have a non-zero pericentric distance. Hence stars will no longer be exposed to the risk of coming arbitrarily close to the black hole, and stars will disappear from the black hole's menu.

Let us assume that the distribution of a star's crossing points is uniform within the waist and calculate the expectation value of the smallest value taken by r in N passages. Let the area of the waist be πR^2. Then the probability of there being n crossing points in a circle of radius r is given by the Poisson distribution (Appendix B.8) as

$$P(n|r) = \frac{(Nr^2/R^2)^n}{n!}\mathrm{e}^{-Nr^2/R^2}. \tag{3.318}$$

[29] See Häfner et al. (2000) for a method of exploiting the uniformity of $f_{\rm irr}$ in galaxy modeling.

The probability that the closest passage lies in $(r, r+\mathrm{d}r)$ is the probability that there are zero passages inside r and a non-zero number of passages in the surrounding annulus, has area $2\pi r\mathrm{d}r$. Thus this probability is

$$\mathrm{d}P = \left(1-\mathrm{e}^{-2Nr\mathrm{d}r/R^2}\right)\mathrm{e}^{-Nr^2/R^2} \simeq \frac{2Nr\mathrm{d}r}{R^2}\mathrm{e}^{-Nr^2/R^2}. \tag{3.319}$$

The required expectation value of r_1 is now easily calculated:

$$\langle r_1\rangle = \int \mathrm{d}r\,\frac{2Nr^2}{R^2}\mathrm{e}^{-Nr^2/R^2} = \sqrt{\frac{\pi}{N}}\frac{R}{2}. \tag{3.320}$$

From equation (3.317) the deflection that corresponds to $\langle r_1\rangle$ is

$$\theta_{\mathrm{defl,max}} = 2\tan^{-1}\left(\frac{2\sqrt{N}GM}{\sqrt{\pi}v^2R}\right). \tag{3.321}$$

Two empirical correlations between galactic parameters enable us to estimate $\theta_{\mathrm{defl,max}}$ for a star that reaches maximum radius R_{max} in an elliptical galaxy with measured line-of-sight velocity dispersion $\sigma_\parallel$. First we take the black hole's mass M from the empirical relation (1.27). In the galaxy's lifetime τ we have $N \simeq \sigma_\parallel\tau/2R_{\mathrm{max}}$, and we relate R_{max} to D_n, the diameter within which the mean surface brightness of an elliptical galaxy is $20.75\,\mathrm{mag\,arcsec^{-2}}$ in the B band: D_n is correlated with $\sigma_\parallel$ such that (BM eq. 4.43)

$$D_n = 5.2\Big(\frac{\sigma_\parallel}{200\,\mathrm{km\,s^{-1}}}\Big)^{1.33}\,\mathrm{kpc}. \tag{3.322}$$

With these relations, (3.321) becomes

$$\theta_{\mathrm{defl,max}} \simeq 2\tan^{-1}\left[0.08\frac{D_n^{3/2}}{R_{\mathrm{max}}^{3/2}}\frac{R_{\mathrm{max}}}{R}\frac{\sigma_\parallel^2}{v^2}\Big(\frac{\sigma_\parallel}{200\,\mathrm{km\,s^{-1}}}\Big)^{0.5}\Big(\frac{\tau}{10\,\mathrm{Gyr}}\Big)^{1/2}\right]. \tag{3.323}$$

For the moderately luminous elliptical galaxies that are of interest here, D_n is comparable to, or slightly larger than, the effective radius (Dressler et al. 1987), and thus similar to the half-mass radius $r_{\mathrm{h}} = 1.3R_{\mathrm{e}}$ for the $R^{1/4}$ profile. Thus for the majority of stars $D_n/R_{\mathrm{max}} \simeq 1$. From Figure 3.46 we estimate $R_{\mathrm{max}}/R \simeq 10$. To estimate the ratio $\sigma_\parallel/v$ we deduce from equations (2.66) and (2.67) that for a Hernquist model with scale radius a the potential drop $\Delta\Phi = \Phi(a) - \Phi(0)$ between $r_{\mathrm{h}} = 2.41a$ and the center is $0.71GM_{\mathrm{gal}}/a$, so $v^2 = 2\Delta\Phi = 1.4GM_{\mathrm{gal}}/a$. From Figure 4.4 we see that $\sigma_\parallel \simeq 0.2\sqrt{GM_{\mathrm{gal}}/a}$, so $(\sigma_\parallel/v)^2 \simeq 35$. Inserting these values into equation (3.323) we find $\theta_{\mathrm{defl,max}} \simeq 2.6°$. Scattering by such a small angle will probably not undermine a galaxy's triaxiality, but stars with smaller apocenter distances R_{max} will be deflected through significant angles, so it is likely that the black hole will erode triaxiality in the galaxy's inner parts (Norman, May, & van Albada 1985; Merritt & Quinlan 1998).

Problems

3.1 [1] Show that the radial velocity along a Kepler orbit is

$$\dot{r} = \frac{GMe}{L}\sin(\psi - \psi_0), \tag{3.324}$$

where L is the angular momentum. By considering this expression in the limit $r \to \infty$ show that the eccentricity e of an unbound Kepler orbit is related to its speed at infinity by

$$e^2 = 1 + \left(\frac{Lv_\infty}{GM}\right)^2. \tag{3.325}$$

3.2 [1] Show that for a Kepler orbit the eccentric anomaly η and the true anomaly $\psi - \psi_0$ are related by

$$\cos(\psi - \psi_0) = \frac{\cos\eta - e}{1 - e\cos\eta} \quad ; \quad \sin(\psi - \psi_0) = \sqrt{1 - e^2}\frac{\sin\eta}{1 - e\cos\eta}. \tag{3.326}$$

3.3 [1] Show that the energy of a circular orbit in the isochrone potential (2.47) is $E = -GM/(2a)$, where $a = \sqrt{b^2 + r^2}$. Let the angular momentum of this orbit be $L_\mathrm{c}(E)$. Show that

$$L_\mathrm{c} = \sqrt{GMb}\left(x^{-1/2} - x^{1/2}\right), \qquad \text{where} \qquad x \equiv -\frac{2Eb}{GM}. \tag{3.327}$$

3.4 [1] Prove that if a homogeneous sphere of a pressureless fluid with density ρ is released from rest, it will collapse to a point in time $t_\mathrm{ff} = \frac{1}{4}\sqrt{3\pi/(2G\rho)}$. The time t_ff is called the **free-fall time** of a system of density ρ.

3.5 [3] Generalize the timing argument in Box 3.1 to a universe with non-zero vacuum-energy density. Evaluate the required mass of the Local Group for a universe of age $t_0 = 13.7\,\mathrm{Gyr}$ with (a) $\Omega_{\Lambda 0} = 0$; (b) $\Omega_{\Lambda 0} = 0.76$, $h_7 = 1.05$. Hints: the energy density in radiation can be neglected. The solution requires evaluation of an integral similar to (1.62).

3.6 [1] A star orbiting in a spherical potential suffers an arbitrary instantaneous velocity change while it is at pericenter. Show that the pericenter distance of the ensuing orbit cannot be larger than the initial pericenter distance.

3.7 [2] In a spherically symmetric system, the apocenter and pericenter distances are given by the roots of equation (3.14). Show that if $E < 0$ and the potential $\Phi(r)$ is generated by a non-negative density distribution, this equation has either no root, a repeated root, or two roots (Contopoulos 1954). Thus there is at most one apocenter and pericenter for a given energy and angular momentum. Hint: take the second derivative of $E - \Phi$ with respect to $u = 1/r$ and use Poisson's equation.

3.8 [1] Prove that circular orbits in a given potential are unstable if the angular momentum per unit mass on a circular orbit decreases outward. Hint: evaluate the epicycle frequency.

3.9 [2] Compute the time-averaged moments of the radius, $\langle r^n \rangle$, in a Kepler orbit of semi-major axis a and eccentricity e, for $n = 1, 2$ and $n = -1, -2, -3$.

3.10 [2] $\Delta\psi$ denotes the increment in azimuthal angle during one complete radial cycle of an orbit.

(a) Show that in the potential (3.57)

$$\Delta\psi = \frac{2\pi L}{\sqrt{-2Er_\mathrm{a}r_\mathrm{p}}}, \tag{3.328}$$

where r_a and r_p are the apo- and pericentric radii of an orbit of energy E and angular momentum L. Hint: by contour integration one can show that for $A > 1$, $\int_{-\pi/2}^{\pi/2} \mathrm{d}\theta/(A + \sin\theta) = \pi/\sqrt{A^2 - 1}$.

(b) Prove in the epicycle approximation that along orbits in a potential with circular frequency $\Omega(R)$,

$$\Delta\psi = 2\pi\left(4 + \frac{\mathrm{d}\ln\Omega^2}{\mathrm{d}\ln R}\right)^{-1/2}. \tag{3.329}$$

(c) Show that the exact expression (3.328) reduces for orbits of small eccentricity to (3.329).

3.11 [1] For what spherically symmetric potential is a possible trajectory $r = a\mathrm{e}^{b\psi}$?

3.12 [2] Prove that the mean-square velocity is on a bound orbit in a spherical potential $\Phi(r)$ is

$$\langle v^2\rangle = \left\langle r\frac{\mathrm{d}\Phi}{\mathrm{d}r}\right\rangle, \tag{3.330}$$

where $\langle\cdot\rangle$ denotes a time average.

3.13 [2] Let $\mathbf{r}(s)$ be a plane curve depending on the parameter s. Then the **curvature** is

$$K = \frac{|\mathbf{r}' \times \mathbf{r}''|}{|\mathbf{r}'|^3}, \tag{3.331}$$

where $\mathbf{r}' \equiv \mathrm{d}\mathbf{r}/\mathrm{d}s$. The local radius of curvature is K^{-1}. Prove that the curvature of an orbit with energy E and angular momentum L in the spherical potential $\Phi(r)$ is

$$K = \frac{L\,\mathrm{d}\Phi/\mathrm{d}r}{2^{3/2}r[E - \Phi(r)]^{3/2}}. \tag{3.332}$$

Hence prove that no orbit in any spherical mass distribution can have an inflection point (in contrast to the cover illustration of Goldstein, Safko, & Poole 2002).

3.14 [1] Show that in a spherical potential the vertical and circular frequencies ν and Ω (eqs. 3.79) are equal.

3.15 [1] Prove that at any point in an axisymmetric system at which the local density is negligible, the epicycle, vertical, and circular frequencies κ, ν, and Ω (eqs. 3.79) are related by $\kappa^2 + \nu^2 = 2\Omega^2$.

3.16 [1] Using the epicycle approximation, prove that the azimuthal angle $\Delta\psi$ between successive pericenters lies in the range $\pi \le \Delta\psi \le 2\pi$ in the gravitational field arising from any spherical mass distribution in which the density decreases outwards.

3.17 [3] The goal of this problem is to prove the results of Problem 3.16 without using the epicycle approximation (Contopoulos 1954).

(a) Using the notation of §3.1, show that

$$E - \Phi - \frac{L^2}{2r^2} = (u_1 - u)(u - u_2)\left\{\tfrac{1}{2}L^2 + \Phi[u, u_1, u_2]\right\}, \tag{3.333}$$

where $u_1 = 1/r_1$ and $u_2 = 1/r_2$ are the reciprocals of the pericenter and apocenter distances of the orbit respectively, $u = 1/r$, and

$$\Phi[u, u_1, u_2] = \frac{1}{u_1 - u_2}\left[\frac{\Phi(u_1) - \Phi(u)}{u_1 - u} - \frac{\Phi(u) - \Phi(u_2)}{u - u_2}\right]. \tag{3.334}$$

This expression is a second-order divided difference of the potential Φ regarded as a function of u, and a variant of the mean-value theorem of calculus shows that $\Phi[u, u_1, u_2] = \frac{1}{2}\Phi''(\bar{u})$ where $\bar{u}$ is some value of u in the interval (u_1, u_2). Then use the hint in Problem 3.7 and equation (3.18b) to deduce that $\Delta\psi \le 2\pi$ when the potential Φ is generated by a non-negative, spherically symmetric density distribution.

(b) A lower bound on $\Delta\psi$ can be obtained from working in a similar manner with the function

$$\chi(\omega) = \frac{2\omega\Phi}{L}, \quad \text{where} \quad \omega \equiv \frac{L}{r^2}. \tag{3.335}$$

Show that

$$\frac{2\omega E}{L} - \chi(\omega) - \omega^2 = (\omega_1 - \omega)(\omega - \omega_2)\left\{1 + \chi[\omega, \omega_1, \omega_2]\right\}, \tag{3.336}$$

where $\omega_1 = L/r_1^2$, $\omega_2 = L/r_2^2$ and $\chi[\omega, \omega_1, \omega_2]$ is a second-order divided difference of $\chi(\omega)$. Now deduce that $\Delta\psi \geq \pi$ for any potential in which the circular frequency $\Omega(r)$ decreases outwards.

3.18 [1] Let $\Phi(R, z)$ be the Galactic potential. At the solar location, $(R, z) = (R_0, 0)$, prove that

$$\frac{\partial^2\Phi}{\partial z^2} = 4\pi G\rho_0 + 2(A^2 - B^2), \tag{3.337}$$

where ρ_0 is the density in the solar neighborhood and A and B are the Oort constants. Hint: use equation (2.73).

3.19 [3] Consider an attractive power-law potential, $\Phi(r) = Cr^\alpha$, where $-1 \leq \alpha \leq 2$ and $C > 0$ for $\alpha > 0$, $C < 0$ for $\alpha < 0$. Prove that the ratio of radial and azimuthal periods is

$$\frac{T_r}{T_\psi} = \begin{cases} 1/\sqrt{2+\alpha} & \text{for nearly circular orbits} \\ \begin{cases} 1/2, & \text{for } \alpha > 0 \\ 1/(2+\alpha), & \text{for } \alpha < 0 \end{cases} & \text{for nearly radial orbits.} \end{cases} \tag{3.338}$$

What do these results imply for harmonic and Kepler potentials?
Hint: depending on the sign of α use a different approximation in the radical for v_r. For $b > 0$, $\int_1^\infty \mathrm{d}x/(x\sqrt{x^b - 1}) = \pi/b$ (see Touma & Tremaine 1997).

3.20 [1] Show that in spherical polar coordinates the Lagrangian for motion in the potential $\Phi(\mathbf{x})$ is

$$\mathcal{L} = \tfrac{1}{2}[\dot{r}^2 + (r\dot{\theta})^2 + (r\sin\theta\dot{\phi})^2] - \Phi(\mathbf{x}). \tag{3.339}$$

Hence show that the momenta p_θ and p_ϕ are related to the the magnitude and z-component of the angular-momentum vector $\mathbf{L}$ by

$$p_\phi = L_z \quad ; \quad p_\theta^2 = L^2 - \frac{L_z^2}{\sin^2\theta}. \tag{3.340}$$

3.21 [3] Plot a $(y, \dot{y})$, $(x = 0, \dot{x} > 0)$ surface of section for motion in the potential Φ_L of equation (3.103) when $q = 0.9$ and $E = -0.337$. Qualitatively relate the structure of this surface of section to the structure of the $(x, \dot{x})$ surface of section shown in Figure 3.9.

3.22 [3] Sketch the structure of the $(x, \dot{x})$, $(y = 0, \dot{y} > 0)$ surface of section for motion at energy E in a Kepler potential when (a) the (x, y) coordinates are inertial, and (b) the coordinates rotate at 0.75 times the circular frequency Ω at the energy E. Hint: see Binney, Gerhard, & Hut (1985).

3.23 [3] The Earth is flattened at the poles by its spin. Consequently orbits in its potential do not conserve total angular momentum. Many satellites are launched in inclined, nearly circular orbits only a few hundred kilometers above the Earth's surface, and their orbits must remain nearly circular, or they will enter the atmosphere and be destroyed. Why do the orbits remain nearly circular?

3.24 [2] Let $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$ be unit vectors in an inertial coordinate system centered on the Sun, with $\hat{\mathbf{e}}_1$ pointing away from the Galactic center (towards $\ell = 180°$, $b = 0$) and $\hat{\mathbf{e}}_2$ pointing towards $\ell = 270°$, $b = 90°$. The mean velocity field $\mathbf{v}(\mathbf{x})$ relative to the Local Standard of Rest can be expanded in a Taylor series,

$$v_i = \sum_{j=1}^{2} H_{ij}x_j + \mathrm{O}(x^2). \tag{3.341}$$

(a) Assuming that the Galaxy is stationary and axisymmetric, evaluate the matrix $\mathbf{H}$ in terms of the Oort constants A and B.

(b) What is the matrix $\mathbf{H}$ in a rotating frame, that is, if $\hat{\mathbf{e}}_1$ continues to point to the center of the Galaxy as the Sun orbits around it?

(c) In a homogeneous, isotropic universe, there is an analogous 3×3 matrix $\mathbf{H}$ that describes the relative velocity $\mathbf{v}$ between two fundamental observers separated by $\mathbf{x}$. Evaluate this matrix in terms of the Hubble constant.

3.25 [3] Consider two point masses m_1 and $m_2 > m_1$ that travel in a circular orbit about their center of mass under their mutual attraction. (a) Show that the Lagrange point L_4 of this system forms an equilateral triangle with the two masses. (b) Show that motion near L_4 is stable if $m_1/(m_1 + m_2) < 0.03852$. (c) Are the Lagrange points L_1, L_2, L_3 stable? See Valtonen & Karttunen (2006).

3.26 [2] Show that the leapfrog integrator (3.166a) is second-order accurate, in the sense that the errors in $\mathbf{q}$ and $\mathbf{p}$ after a timestep h are $\mathrm{O}(h^3)$.

3.27 [2] Forest & Ruth (1990) have devised a symplectic, time-reversible, fourth-order integrator of timestep h by taking three successive drift-kick-drift leapfrog steps of length ah, bh, and ah where $2a + b = 1$. Find a and b. Hint: a and b need not both be positive.

3.28 [2] Confirm the formulae for the Adams–Bashforth, Adams–Moulton, and Hermite integrators in equations (3.169), (3.170), and (3.171), and derive the next higher order integrator of each type. You may find it helpful to use computer algebra.

3.29 [1] Prove that the fictitious time τ in Burdet–Heggie regularization is related to the eccentric anomaly η by $\tau = (T_r/2\pi a)\eta + \mathit{constant}$, if the motion is bound ($E_2 < 0$) and the external field $\mathbf{g} = 0$.

3.30 [1] We wish to integrate numerically the motions of N particles with positions $\mathbf{x}_i$, velocities $\mathbf{v}_i$, and masses m_i. The particles interact only by gravitational forces (the gravitational N-body problem). We are considering using several possible integrators: modified-Euler, leapfrog, or fourth-order Runge–Kutta. Which of these will conserve the total momentum $\sum_{i=1}^N m_i \mathbf{v}_i$? Which will conserve the total angular momentum $\sum_{i=1}^N m_i \mathbf{x}_i \times \mathbf{v}_i$? Assume that all particles are advanced with the same timestep, and that forces are calculated exactly. You may solve the problem either analytically or numerically.

3.31 [2] Show that the generating function of the canonical transformation from angle-action variables (θ_i, J_i) to the variables (q_i, p_i) discussed in Box 3.4 is

$$S(q, J) = \mp\tfrac{1}{2} q\sqrt{2J - q^2} \pm J \cos^{-1}\left(\frac{q}{\sqrt{2J}}\right). \tag{3.342}$$

3.32 [1] Let $\epsilon(R)$ and $\ell(R)$ be the specific energy and angular momentum of a circular orbit of radius R in the equatorial plane of an axisymmetric potential.

(a) Prove that

$$\frac{\mathrm{d}\ell}{\mathrm{d}R} = \frac{R\kappa^2}{2\Omega} \quad ; \quad \frac{\mathrm{d}\epsilon}{\mathrm{d}R} = \tfrac{1}{2} R\kappa^2, \tag{3.343}$$

where Ω and κ are the circular and epicycle frequencies.

(b) The energy of a circular orbit as a function of angular momentum is $\epsilon(\ell)$. Show that $\mathrm{d}\epsilon/\mathrm{d}\ell = \Omega$ in two ways, first from the results of part (a) and then using angle-action variables.

3.33 [2] The angle variables θ_i conjugate to the actions J_i can be implicitly defined by the coupled differential equations $\mathrm{d}w_\alpha/\mathrm{d}\theta_i = [w_\alpha, J_i]$, where w_α is any ordinary phase-space coordinate. Using this result, show that the angle variable for the harmonic oscillator, $H = \frac{1}{2}(p^2 + \omega^2 q^2)$, may be written

$$\theta(x, p) = -\tan^{-1}\left(\frac{p}{\omega q}\right). \tag{3.344}$$

Hint: the action is $J = H/\omega$.

3.34 [2] Consider motion for $L_z = 0$ in the Stäckel potential (3.247).

(a) Express I_3 as a function of u, v, p_u, and p_v.

(b) Show that $H \cos^2 v + I_3 = \frac{1}{2}(p_v^2/\Delta^2) - V$.

(c) Show that $[H, I_3] = 0$.

(d) Hence show that J_u and J_v are in involution, that is $[J_u, J_v] = 0$. Hint: if $f(a, b)$ is any differentiable function of two variables, and A is any differentiable function of the phase-space variables, then $[A, f] = [A, a](\partial f/\partial a) + [A, b](\partial f/\partial b)$.

3.35 [2] A particle moves in a one-dimensional potential well $\Phi(x)$. In angle-action variables, the Hamiltonian has the form $H(J) = cJ^{4/3}$ where c is a constant. Find $\Phi(x)$.

3.36 [2] Obtain the Hamiltonian and fundamental frequencies as functions of the actions for the three-dimensional harmonic oscillator by examining the limit $b \to \infty$ of equations (3.226).

3.37 [2] For motion in a potential of the form (3.247), obtain

$$\dot{p}_u = \frac{2E \sinh u \cosh u - \mathrm{d}U/\mathrm{d}u}{\sinh^2 u + \sin^2 v} + \frac{L_z^2 \cosh u}{\Delta^2 \sinh^3 u(\sinh^2 u + \sin^2 v)}, \tag{3.345}$$

where (u, v) are the prolate spheroidal coordinates defined by equations (3.242), by (a) differentiating equation (3.249a) with respect to t and then using $\dot{u} = \partial H/\partial p_u$, and (b) from $\dot{p}_u = -\partial H/\partial u$.

3.38 [2] For the coordinates defined by equation (3.267), show that the integral defined by equations (3.268) can be written

$$I_2 = \frac{\sinh^2 u[\frac{1}{2}(p_v^2/\Delta^2) - V] - \sin^2 v[\frac{1}{2}(p_u^2/\Delta^2) + U]}{\sinh^2 u + \sin^2 v}. \tag{3.346}$$

Show that in the limit $\Delta \to 0$, $u \to \infty$ we have $\Delta \sinh u \to \Delta \cosh u \to R$ and $v \to \pi/2 - \phi$, where R and ϕ are the usual polar coordinates. Hence show that in this limit $2\Delta^2 I_2 \to L_z^2$.

3.39 [2] Show that the third integral of an axisymmetric Stäckel potential can be taken to be

$$\begin{aligned} I_3(u, v, p_u, p_v, p_\phi) = {} & \frac{1}{\sinh^2 u + \sin^2 v} \times \\ & \left[\sinh^2 u\left(\frac{p_v^2}{2\Delta^2} - V\right) - \sin^2 v\left(\frac{p_u^2}{2\Delta^2} + U\right)\right] + \frac{p_\phi^2}{2\Delta^2}\left(\frac{1}{\sin^2 v} - \frac{1}{\sinh^2 u}\right). \end{aligned} \tag{3.347}$$

Hint: generalize the work of Problem 3.38.

3.40 [1] Show that when orbital frequencies are incommensurable, adiabatic invariance of actions implies that closed orbits remain closed when the potential is adiabatically deformed. An initially circular orbit in a spherical potential Φ does not remain closed when Φ is squashed along any line that is not parallel to the orbit's original angular-momentum vector. Why does this statement remain true no matter how slowly Φ is squashed?

3.41 [2] From equations (3.39b) and (3.190), show that the radial action J_r of an orbit in the isochrone potential (2.47) is related to the energy E and angular momentum L of this orbit by

$$J_r = \sqrt{GMb}\left[x^{-\frac{1}{2}} - f(L)\right], \tag{3.348}$$

where $x \equiv -2Eb/(GM)$ and f is some function. Use equation (3.327) to show that $f(L) = (\sqrt{l^2+1} - l)^{-1} = \sqrt{l^2+1} + l$, where $l \equiv |L|/(2\sqrt{GMb})$, and hence show that the isochrone Hamiltonian can be written in the form (3.226a).

3.42 [2] Angle-action variables are also useful in general relativity. For example, the relativistic analog to the Hamilton–Jacobi equation (3.218) for motion in the point-mass potential $\Phi(r) = -GM/r$ is

$$E^2\left(\frac{1+\frac{1}{4}r_{\rm S}/r}{1-\frac{1}{4}r_{\rm S}/r}\right)^2 = c^4 + \frac{c^2}{(1+\frac{1}{4}r_{\rm S}/r)^4}\left[\Big(\frac{\partial S}{\partial r}\Big)^2 + \Big(\frac{1}{r}\frac{\partial S}{\partial \vartheta}\Big)^2 + \Big(\frac{1}{r\sin\vartheta}\frac{\partial S}{\partial \phi}\Big)^2\right], \tag{3.349}$$

where $r_{\rm S} \equiv 2GM/c^2$ is the **Schwarzschild radius**, the energy per unit mass E includes the rest-mass energy c^2, and the equations are written in the isotropic metric, i.e., $\mathrm{d}s^2$ at any point is proportional to its Euclidean form (Landau & Lifshitz 1999).

(a) Show that the Hamiltonian can be written in the form

$$H(p_r, p_\vartheta, p_\phi) = \frac{1-\frac{1}{4}r_{\rm S}/r}{1+\frac{1}{4}r_{\rm S}/r}\sqrt{c^4 + \frac{c^2p^2}{(1+\frac{1}{4}r_{\rm S}/r)^4}}, \tag{3.350}$$

where $p^2 = p_r^2 + p_\vartheta^2/r^2 + p_\phi^2/(r\sin\vartheta)^2$.

(b) For systems in which relativistic effects are weak, show that the Hamiltonian can be written in the form

$$H = c^2 + H_{\rm Kep} + H_{\rm gr} + \mathrm{O}(c^{-4}), \tag{3.351}$$

where $H_{\rm Kep} = \frac{1}{2}p^2 - GM/r$ is the usual Kepler Hamiltonian and

$$H_{\rm gr} = \frac{1}{c^2}\left(\frac{G^2M^2}{2r^2} - \frac{p^4}{8} - \frac{3GMp^2}{2r}\right). \tag{3.352}$$

(c) To investigate the long-term effects of relativistic corrections on a Kepler orbit, we may average $H_{\rm gr}$ over an unperturbed Kepler orbit. Show that this average may be written

$$\langle H_{\rm gr}\rangle = \frac{G^2M^2}{c^2a^2}\left(\tfrac{15}{8} - \frac{3}{\sqrt{1-e^2}}\right), \tag{3.353}$$

where a and e are the semi-major axis and eccentricity. Hint: use the results of Problem 3.9.

(d) Show that relativistic corrections cause the argument of pericenter ω to precess by an amount

$$\Delta\omega = \frac{6\pi GM}{c^2a(1-e^2)} \tag{3.354}$$

per orbit. Hint: convert $\langle H_{\rm gr}\rangle$ to angle-action variables using Table E.1 and use Hamilton's equations.

3.43 [2] The Hamiltonian $H(\mathbf{x}, \mathbf{p}; \lambda)$, where λ is a parameter, supports a family of resonant orbits. In the (x_1, p_1) surface of section, the family's chain of islands is bounded by orbits with actions $J_1 \equiv (2\pi)^{-1}\oint \mathrm{d}x_1\, p_1 = J_\pm(\lambda)$, where $J_+ > J_-$. Let λ increase sufficiently slowly for the actions of non-resonant orbits to be conserved, and assume that $J_+' > J_-' > 0$, where a prime denotes differentiation with respect to λ. Show that, as λ grows, an orbit of unknown phase and action slightly larger than J_+ will be captured by the resonance with probability $P_{\rm c} = 1 - J_-'/J_+'$. Hint: exploit conservation of phase-space volume as expressed by equation (4.10).