2

Galactic Structure and Dynamics

2.1 Introduction

In this chapter several aspects of galactic structure and evolution are discussed. Some of the ideas introduced in Chap. 1 are used and the contents of this chapter will be needed in Chaps. 7–10.

In the study of stellar structure and evolution in Vol. II, we could begin with a series of physically justifiable assumptions, derive the relevant equations describing the stars, and integrate these equations to understand the structure and evolution of stars. Such an approach is impossible in the case of galaxies for several reasons. To begin with, we do not understand how galaxies have formed. (It is true that there are several uncertainties in the case of star formation as well but they refer to details rather than to the fundamental process itself.) Second, observational data related to the galaxies are by no means statistically as well determined and abundant as data related to the stars. The reason essentially has to do with the fact that galaxies are located farther away from us and thus are more difficult to observe with the same level of accuracy. Third, galaxies, being collisionless systems of stars, are intrinsically more complicated compared with stars – which are made of collisional gas – from the point of view of mathematical description.[1]

Given these difficulties, it is better to divide the study of galactic structure and dynamics into several separate aspects and investigate each of them as though they are disconnected from each other. The latter assumption, of course, is only approximately true and we eventually will need to worry about how the individual aspects fit in the overall scheme of things. We shall address some of these broader questions in Chap. 7 and concentrate on more well-defined and conceptually simpler issues in this chapter.

A natural subdivision of the topics in the study of galaxies is based on different time scales and length scales that are involved in their description. The longest time scale that is operational with regard to galaxies is the time scale of formation, $t_{\rm for} \approx \rho_*/\dot{\rho}_*$, where ρ_* is the density of stars in a galaxy and $\dot{\rho}_*$ is the rate of star

formation. Both these quantities could vary in space and time; but for typical values, this time scale is 1- 10 Gyr and is comparable with the cosmological time scales in the redshift range $z \simeq (0–5)$, depending on the model chosen. At time scales shorter than this, we may think of the galaxy as a well-defined physical system that has already formed, and we can ignore the effects related to the actual scenario for the formation of galaxy. The second time scale is the dynamical one, $t_{\rm dyn} \approx (R/v)$, where R is the typical size of the galaxy and v is the typical velocity of stars; this is $t_{\rm dyn} \simeq 0.1$ Gyr. This is, for example, the time scale that would govern the process of violent relaxation (Vol. I, Chap. 10, Section 10.6) if it was operational in the early stages of galaxy formation. Finally, the stars in the galaxy evolve at a time scale that is typically 0.01–1 Gyr, depending on their masses. Over this time scale, the gaseous and chemical contents of the galaxy will change significantly. (The time scale for gravitational two-body relaxation, $t_{\rm relax} \approx (N/\ln N)t_{\rm dyn}$, is considerably larger than the age of the universe and can be ignored; see Vol. I, Chap. 10, Section 10.7.)

The preceding time scales are relevant for an isolated galaxy. If a galaxy is in close proximity with one or more galaxies, then it is necessary to take into account the effects of galaxy–galaxy encounters. As we shall see in Chap. 7, this effect is significant because virtually every galaxy would have undergone a close encounter with another galaxy some time during its past history.

The simplest (and mathematically most well-defined) question related to galaxies is the modelling of the steady-state distribution of stars in it. In addressing this problem, we treat stars as structureless point particles interacting only through gravity. Our description of galaxies begins with such a crude first approximation in Sections 2.2–2.7. At the next level, we could incorporate some aspects of stellar evolution and its effects on the chemical evolution of galaxies. This is attempted in Section 2.8. Finally, in Section 2.9 we describe some simple effects that could arise because of galaxy–galaxy interaction. Issues related to the formation and early evolution of galaxies as well as the effect of galaxies on the IGM will be discussed in Chaps. 7 and 9.

2.2 Models for Galaxies in Steady State

In mathematical terms, the modelling of the steady-state distribution of stars in a galaxy requires finding self-consistent solutions to the collisionless Boltzmann equation (CBE) in which the smooth distribution of stars generates a gravitational potential and the orbits admissible in this potential are populated by stars so as to ensure self-consistency (see Vol. I, Chap. 10, Section 10.4).

The distribution of stars in a galaxy can be described by a function $f(\mathbf{x}, \mathbf{v}, t)$ that may be interpreted as the (relative) probability of finding a star in the phase space in the interval $(\mathbf{x}, \mathbf{x} + d^3\mathbf{x}; \mathbf{v}, \mathbf{v} + d^3\mathbf{v})$ at time t. The smoothed-out mass

density of the stars at any point $\mathbf{x}$ will be

$$\rho(\mathbf{x}, t) = m \int f(\mathbf{x}, \mathbf{v}, t)\, d^3\mathbf{v}, \tag{2.1}$$

where m is the average mass of the stars. Such a smoothed-out density will produce a smooth gravitational potential $\phi(\mathbf{x}, t)$, where

$$\nabla^2\phi = 4\pi G\rho. \tag{2.2}$$

We may assume that each star moves in this smooth gravitational field along some specific orbit. The conservation of the total number of stars, expressed by $(df/dt) = 0$, will reduce to

$$\begin{aligned} \frac{df}{dt} &= \frac{\partial f}{\partial \mathbf{x}} + \dot{\mathbf{v}} \cdot \frac{\partial f}{\partial \mathbf{v}} + \dot{\mathbf{x}} \cdot \frac{\partial f}{\partial \mathbf{x}} \\ &= \frac{\partial f}{\partial t} + \mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{x}} - \nabla\phi \cdot \frac{\partial f}{\partial \mathbf{v}} = 0. \end{aligned} \tag{2.3}$$

Models for galaxies are based on the solution to coupled equations (2.2) and (2.3).

It should be noted that the *actual* gravitational potential at any point $\mathbf{x}$, because of the stars, is

$$\phi_{\text{act}}(\mathbf{x}, t) = -\sum_i \frac{Gm}{|\mathbf{x} - \mathbf{x}_i(t)|}, \tag{2.4}$$

where $\mathbf{x}_i(t)$ is the position of the ith star at time t. This will be different from the ϕ in Eq. (2.2), which was produced by the smooth density. Because of this difference, the actual trajectories of the stars will differ appreciably from the orbits of the smooth potential after sufficiently long time intervals. However, it can be easily shown (see Vol. 1, Chap. 10, Section 10.7) that this time scale is very large compared with the orbital time scales in galaxies. Hence Eqs. (2.2) and (2.3) provide an excellent description of galaxies over reasonable time scales.

Because these equations allow a wide variety of solutions, different classes of models are possible for the galaxies. Unfortunately, there is no way of choosing any subset of them as appropriate to a galaxy without further input. Theoretically, we would have expected the initial conditions (at the time of formation of galaxies) and the process of violent relaxation (see Vol. I, Chap. 10, Section 10.6) to have led to a subset of all possible solutions to the CBE. However, it has not been possible to provide a general analysis of the time-dependent CBE, thereby determining the nature of final configurations to which the system evolves at late times. The best we can do at present is to provide a series of simplified time-independent models that serve as a backdrop for comparing theory and observations. We shall discuss a few such models in this section.

When the galaxy is in steady state, $f(t, \mathbf{x}, \mathbf{v}) = f(\mathbf{x}, \mathbf{v})$. To produce such a steady-state solution we can proceed as follows: Let $C_i = C_i(\mathbf{x}, \mathbf{v})$, $i = 1$,

2, . . . , be a set of isolating integrals of motion for the stars moving in the potential ϕ (which, right now, is not known). It is obvious that any function $f(C_i)$ of the C_i will satisfy the steady-state Boltzmann equation: $(df/dt) = (\partial f/\partial C_i)\dot{C}_i = 0$ because $\dot{C}_i$ is identically zero. If we can now determine ϕ from f self-consistently and populate the orbits of ϕ with stars, we have solved the problem. Let us consider some specific examples to see how this idea works.

In these calculations, it is convenient to shift the origin of $|\phi|$ by defining a new potential $\psi \equiv -\phi + \phi_0$, where ϕ_0 is a constant. (We will choose the value of ϕ_0 such that ψ vanishes at the "boundary" of the galaxy.) The new potential satisfies

$$\nabla^2\psi = -4\pi G\rho, \tag{2.5}$$

and the boundary condition $\psi \to \phi_0$ as $|\mathbf{x}| \to \infty$. We also define a "shifted" energy for the stars, $\epsilon = -E + \phi_0$; because $\phi_0 = \psi + \phi$, $\epsilon = -E + \psi + \phi = -(1/2)v^2 + \psi$.

The simplest galactic models are the ones in which $f(\mathbf{x}, \mathbf{v})$ depends on $\mathbf{x}$ and $\mathbf{v}$ only through the quantity ϵ so that $f = f(\epsilon) = f(\psi - \frac{1}{2}v^2)$. The density $\rho(\mathbf{x})$ corresponding to this distribution is

$$\rho(\mathbf{x}) = \int_0^{\sqrt{2\psi}} 4\pi v^2\, dv f\left(\psi - \frac{1}{2}v^2\right) = \int_0^{\psi} 4\pi\, d\epsilon f(\epsilon)\sqrt{2(\psi - \epsilon)}. \tag{2.6}$$

The limits of integration are chosen in such a way to pick only the stars bound in the galaxy's potential. The right-hand side is a known function of ψ, once $f(\epsilon)$ is specified. The Poisson equation,

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d\psi}{dr}\right) = -4\pi G\rho = -16\pi^2 G\int_0^{\psi} d\epsilon f(\epsilon)\sqrt{2(\psi - \epsilon)}, \tag{2.7}$$

can now be solved – with some central value $\psi(0)$ and the boundary condition $\psi'(0) = 0$ – determining $\psi(r)$. Once $\psi(r)$ is known, all other variables can be computed.

Three different choices for $f(\epsilon)$ have been extensively used in the literature to describe spherically symmetric systems. We will summarise these models briefly.

2.2.1 Polytropes

The simplest form of $f(\epsilon)$ is a power law with $f(\epsilon) = A\epsilon^{n-3/2}$ for $\epsilon > 0$ and zero otherwise. Using Eq. (2.6), we see that this corresponds to density distributions of the form $\rho = B\psi^n$ (for $\psi > 0$) with $B = (2\pi)^{3/2}A\Gamma(n - 1/2)[\Gamma(n + 1)]^{-1}$. (Clearly, we need $n > 1/2$ to obtain finite density.) The Poisson equation now becomes

$$\frac{1}{l^2}\frac{d}{dl}\left(l^2\frac{d\xi}{dl}\right) = \begin{cases} -\xi^n & \xi \geq 0 \\ 0 & \xi \leq 0 \end{cases}, \tag{2.8}$$

where we have introduced the variables

$$L = [4\pi G\psi(0)^{n-1}B]^{-\frac{1}{2}}, \qquad l = (r/L), \qquad \xi = [\psi/\psi(0)]. \tag{2.9}$$

This equation is the Lane–Emden equation that was discussed extensively in Vol. I, Chap. 10, Section 10.3. The case that is of interest in galactic modelling arises for $n = 5$. In this case, Eq. (2.8) has the simple solution

$$\xi = \left(1 + \frac{1}{3}l^2\right)^{-\frac{1}{2}}, \tag{2.10}$$

which corresponds to a density profile of $\rho \propto [1 + (1/3)l^2]^{-5/2}$ and a total mass of $M = [\sqrt{3}L\psi(0)/G]$. This model (called the *Plummer model*) provides a reasonable description of some elliptical galaxies.

2.2.2 Isothermal Spheres

This model corresponds to the distribution function

$$f(\epsilon) = \frac{\rho_0}{(2\pi\sigma^2)^{3/2}} \exp\left(\frac{\epsilon}{\sigma^2}\right), \tag{2.11}$$

parameterised by two constants, ρ_0 and σ. We can easily verify that the mean-square velocity $\langle v^2 \rangle$ is $3\sigma^2$ and that the density distribution is $\rho(r) = \rho_0 \exp(\psi/\sigma^2)$. The central density is $\rho_c = \rho_0 \exp[\psi(0)/\sigma^2]$. It is conventional to define a core radius and a set of dimensionless variables by

$$r_0 = \left(\frac{9\sigma^2}{4\pi G\rho_c}\right)^{1/2}, \qquad l = \frac{r}{r_0}, \qquad \xi = \frac{\rho}{\rho_c}. \tag{2.12}$$

Then the Poisson equation can be rewritten in the form

$$\frac{1}{l^2}\frac{d}{dl}\left(\frac{l^2}{\xi}\frac{d\xi}{dl}\right) = -9\xi. \tag{2.13}$$

This equation has to be integrated numerically [with the boundary condition $\xi(0) = 1$ and $\xi'(0) = 0$] to determine the density profile. It was shown (in Vol. I, Chap. 10, Section 10.3) that the solution has the asymptotic limit of $\xi \simeq (2/9l^2)$. Hence the exact solution has a mass profile with $M(r) \propto r$ at larger r; the model has to be cut off at some radius to provide a finite mass.

For $l \lesssim 2$, the numerical solution is well approximated by the function $\xi(l) = (1 + l^2)^{-\frac{3}{2}}$ (see Fig. 2.1). The projected, two-dimensional, surface density corresponding to this $\xi(l)$ is $S(R) = 2(1 + R^2)^{-1}$, where R is the projected radial distance. In other words, the (true) central surface density $\Sigma(0) \simeq 2\rho_c r_0$, where ρ_c is the central volume density and r_0 is the core radius. Because the mass density of the isothermal sphere varies as r^{-2} at large distances, the potential ϕ

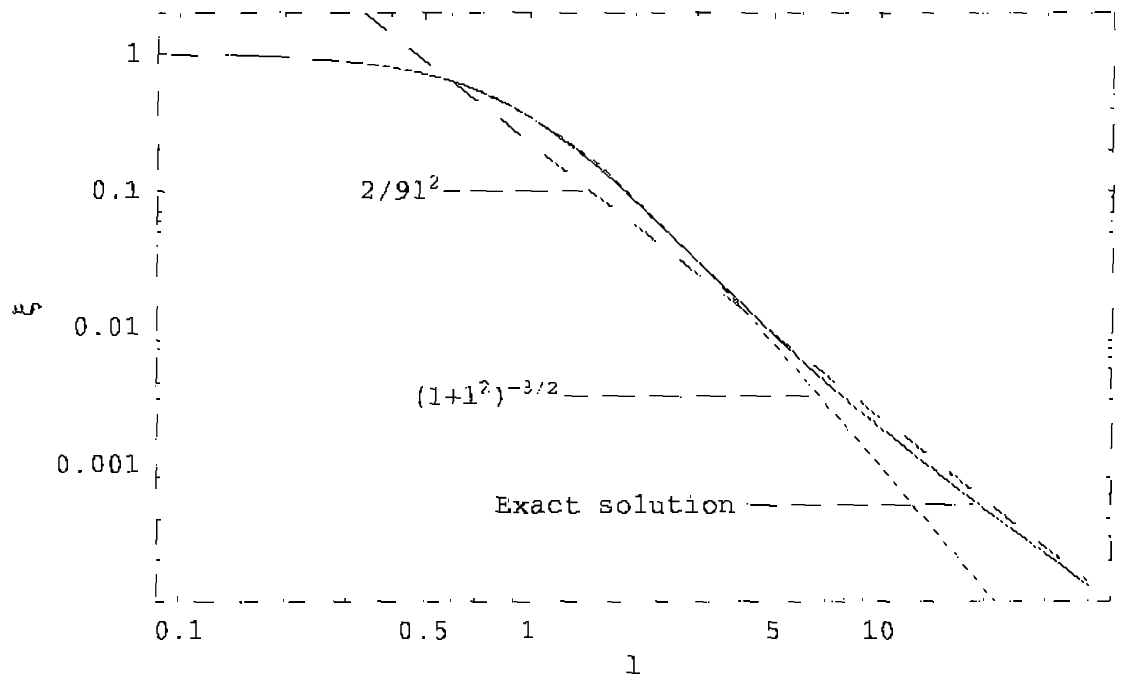


Fig. 2.1. The exact isothermal density profile $\xi(l)$ (solid curve) as well as the two asymptotic forms valid at small and large radii. For large radii, $\xi(l) \approx (2/9)l^{-2}$, which is a long-dashed straight line in the logarithmic plot, and for small radii $\xi(l) \approx (1+l^2)^{-3/2}$, which is shown by a short-dashed curve.

varies as $\ln r$. A disk of test stars embedded in such an isothermal sphere will have constant rotational velocities.

2.2.3 King Model

These are based on the distribution function with

$$f(\epsilon) = \frac{\rho_c}{(2\pi\sigma^2)^{3/2}}\left(e^{\epsilon/\sigma^2} - 1\right), \quad \epsilon \geq 0, \tag{2.14}$$

and $f(\epsilon) = 0$ for $\epsilon < 0$. Because $f(\epsilon)$ vanishes for $\epsilon < 0$, this model may be thought of as a truncated isothermal sphere. The density, when integrated from the origin numerically, will vanish at some radius r_t (called the tidal radius). The quantity $c = \log(r_t/r_0)$ is a measure of how concentrated the system is. Bright elliptical galaxies are well described by such a model with $[\psi(0)/\sigma^2] \simeq 10$ and $c \simeq 2.4$.

The preceding examples show how galactic models of reasonable description can be constructed as solutions to the CBE. Unfortunately, this particular approach has no dynamical content and is merely a tool to fit observations. In fact, the situation is worse than that; for a wide class of acceptable $\rho(r)$, it is always possible to obtain an $f(\epsilon)$ that will self-consistently reproduce the given density profile. To see this, we begin by noting that a given $\rho(r)$ uniquely determines a

$\psi(r)$ and hence the function $\rho(\psi)$. Writing Eq. (2.6) as

$$\frac{1}{\sqrt{8}\pi}\rho(\psi) = 2\int_0^{\psi} f(\epsilon)\sqrt{\psi - \epsilon}\, d\epsilon \tag{2.15}$$

and differentiating both sides with respect to ψ, we get

$$\frac{1}{\sqrt{8}\pi}\frac{d\rho}{d\psi} = \int_0^{\psi} \frac{f(\epsilon)\, d\epsilon}{\sqrt{\psi - \epsilon}}. \tag{2.16}$$

This equation (called Abel's integral equation) has the solution

$$f(\epsilon) = \frac{1}{\sqrt{8}\pi^2}\frac{d}{d\epsilon}\int_0^{\epsilon}\left(\frac{d\rho}{d\psi}\right)\frac{d\psi}{\sqrt{\epsilon - \psi}}, \tag{2.17}$$

which determines $f(\epsilon)$. [Although this procedure gives an $f(\epsilon)$, there is no assurance that it will be positive definite; obviously, the method works only for those $\rho(r)$ for which $f(\epsilon) > 0$ for all ϵ.]

The above procedure leads to a galaxy model that will have an isotropic velocity dispersion. This follows from the fact that, in the above approach, $\rho(r)$ was used to determine an $f(\epsilon)$ that depends on only v^2. Because observations suggest that galaxies have an anisotropic velocity dispersion, it would be interesting to ask whether our procedure can be modified to cover them. If we allow the distribution function to depend on the angular momentum J as well as on ϵ, then a wider class of models can be constructed to fit the same $\rho(r)$. To see this, assume that f depends on ϵ and J^2 only through the combination

$$Q = \epsilon - \frac{J^2}{2R_0^2}, \quad R_0 = \text{constant}. \tag{2.18}$$

We can easily integrate out the angular variables in the velocity space and obtain the density distribution

$$\rho(r) = \frac{2\pi\sqrt{8}}{(1 + r^2/R_0^2)}\int_0^{\psi} f(Q)\sqrt{\psi - Q}\, dQ \equiv \frac{2\pi\sqrt{8}}{(1 + r^2/R_0^2)}\mu(\psi). \tag{2.19}$$

Comparing Eqs. (2.19) and (2.6), we see that

$$f(Q) = \frac{1}{\pi^2\sqrt{8}}\frac{d}{dQ}\int_0^{Q}\left(\frac{d\mu}{d\psi}\right)\frac{d\psi}{\sqrt{Q - \psi}}. \tag{2.20}$$

Thus, given a $\rho(r)$, we determine $\psi(r)$ and another function $\mu(r) \equiv \rho(r)\,(1 + r^2/R_0^2)\,(2\pi\sqrt{8})^{-1}$. Eliminating r between $\mu(r)$ and $\psi(r)$, we obtain $\mu(\psi)$ and consequently $f(Q)$. This distribution function will lead to the density distribution $\rho(r)$ we started with. When $f = f(\epsilon)$, the velocity dispersion is isotropic with $\langle v_r^2\rangle = \langle v_\theta^2\rangle = \langle v_\phi^2\rangle$. For a distribution function of the form $f(\epsilon, J^2)$, we have $v_\theta^2 = v_\phi^2 \neq v_r^2$, thereby providing some amount of anisotropy in the velocity dispersion.

2.2.4 Axisymmetric Systems

Similar modelling also works for a disklike system if the distribution function is taken to be of the form $f = f(\epsilon, J_z)$. We will mention a few examples: If the distribution function is taken to be

$$f(\epsilon, J_z) = \begin{cases} A J_z^n \exp\left(\epsilon/\sigma^2\right), & (\text{for } J_z \geq 0) \\ 0, & (\text{for } J_z \leq 0) \end{cases}, \tag{2.21}$$

then we obtain a surface density for the disk, which is

$$\Sigma(R) = \frac{(n+1)\sigma^2}{2\pi G R} \equiv \frac{\Sigma_0 R_0}{R}. \tag{2.22}$$

The circular velocity of stars, $v_c^2 = -R(\partial\psi/\partial R) = 2\pi G \Sigma_0 R_0$, is a constant for this model, called *Mestel's disk*. These parameters are related to the parameters in the distribution function by

$$n = \left(v_c^2/\sigma^2\right) - 1, \qquad A = \Sigma_0 R_0 \left[2^{n/2}\sqrt{\pi}\Gamma(n+1/2)\sigma^{n+2}\right]^{-1}. \tag{2.23}$$

The velocity dispersion in the radial direction σ^2 is a free parameter characterising the disk. The parameter n is a measure of the "coldness" of the disk.

A more complicated set of disk models (called the *Kalnajs disk*) can be obtained from the distribution function, which has the form

$$f(\epsilon, J_z) = A\left[\left(\Omega_0^2 - \Omega^2\right)a^2 + 2(\epsilon + \Omega J_z)\right]^{-1/2} \tag{2.24}$$

when the term in the square brackets is positive and zero otherwise. This function leads to the following surface density and potential:

$$\Sigma(R) = \Sigma_0\left(1 - \frac{R^2}{a^2}\right)^{1/2}, \qquad \phi(R) = \frac{\pi^2 G \Sigma_0}{4a} R^2 \equiv \frac{1}{2}\Omega_0^2 R^2, \tag{2.25}$$

with Σ_0 and A related by $\Sigma_0 = 2\pi A a\sqrt{\Omega_0^2 - \Omega^2}$. The parameter Ω is free and describes the mean (systematic) rotational velocity of the system: $\langle v_\phi \rangle = \Omega R$ (see Exercise 2.3).

Exercise 2.1

Logarithmic potential: Another galactic model, which we shall have occasion to discuss extensively, has the gravitational potential of the form

$$\phi = \Phi_L(R, z) \equiv \frac{1}{2} v_0^2 \ln\left(R_c^2 + R^2 + \frac{z^2}{q_\Phi^2}\right), \tag{2.26}$$

where R_c, v_0, and q_Φ are constants. Show that the mass density $(4\pi G)^{-1}\nabla^2\phi$ corresponding to the potential in Eq. (2.26) is given by

$$\rho_L(R, z) = \left(\frac{v_0^2}{4\pi G q_\Phi^2}\right)\frac{\left(2q_\Phi^2 + 1\right)R_c^2 + R^2 + 2\left(1 - \frac{1}{2}q_\Phi^{-2}\right)z^2}{\left(R_c^2 + R^2 + z^2 q_\Phi^{-2}\right)^2}. \tag{2.27}$$

As we shall see in Section 2.4, this logarithmic potential helps us to understand several features of orbital structure in real galaxies. Find a distribution function that will lead to this potential self-consistently.

Exercise 2.2

Disks with negligible angular momentum: We intuitively associate highly flattened systems with systems that have a large angular momentum. Consider a disk galaxy in which all stars rotate in the same direction and are supported against gravity by centrifugal force. Is it possible to construct now another *disk* galaxy with the same spatial distribution of stars but with negligible angular momentum?

Exercise 2.3

Properties of disks: Verify the various results quoted in the text as regards the polytropes, isothermal spheres, King's model, Mestel's disk, and the Kalnajs disk. (a) Show that galaxy models with distribution function of the form $f(E)$ have isotropic velocity dispersion, that is, $\langle v_r^2 \rangle = \langle v_\theta^2 \rangle = \langle v_\phi^2 \rangle$. Also show that, if $f = f(E, J^2)$, then $\langle v_\theta^2 \rangle = \langle v_\phi^2 \rangle \neq \langle v_r^2 \rangle$. (b) Show that, for the isothermal sphere, $\langle v^2 \rangle = 3\sigma^2$. (c) Show that, for the Mestel's disk, $\langle v_r^2 \rangle = \sigma^2$ and $\langle v_\phi \rangle = \sqrt{2}\Gamma(n/2+1)\,[\Gamma(n/2+1/2)]^{-1}\sigma$. Hence show that, as n increases to large values, $[\langle v_\phi \rangle/\sigma]$ increases as n. (d) For the Kalnajs disk, show that $\langle v_\phi \rangle / r = \Omega$ is independent of r; thus the disk rotates like a rigid body. Also show that, for $\Omega \ll \Omega_0$, the system is "hot," with the support against gravitational force arising from random velocities, whereas, for $\Omega \approx \Omega_0$, the system is "cold" and centrifugally supported. (e) Consider a distribution function $f = f(E, J_z)$ that leads to a density $\rho(r, z)$. Let $f_\pm(E, J_z) = \frac{1}{2}[f(E, J_z) \pm f(E, -J_z)]$. Show that ρ is determined by f_+ and $\langle v_\phi \rangle$ is determined by f_-.

Exercise 2.4

Vertical density profile of a disk: Find the density profile of a thin self-gravitating disk of stars whose velocity distribution is everywhere Maxwellian and the distribution varies along only the z direction. [Answer. We take the one-dimensional phase-space density to be

$$f = f(E_z) = \rho_0 \left(2\pi\sigma_z^2\right)^{-1/2} \exp\left(-E_z/\sigma_z^2\right), \qquad E_z = \frac{v_z^2}{2} + \Phi(z). \tag{2.28}$$

In terms of the dimensionless variables,

$$\phi \equiv \frac{\Phi}{\sigma_z^2}, \qquad \zeta \equiv \frac{z}{z_0}, \qquad z_0 \equiv \frac{\sigma_z}{\sqrt{8\pi G \rho_0}}, \tag{2.29}$$

the Poisson equation can be reduced to the form

$$2\frac{d^2\phi}{d^2\zeta^2} = e^{-\phi}. \tag{2.30}$$

Solving this with the boundary condition $\phi(0) = 0 = (d\phi/d\zeta)_0$, we get the distribution

$$\rho(z) = \rho_0 \operatorname{sech}^2\left(\frac{1}{2}\frac{z}{z_0}\right), \tag{2.31}$$

which describes the vertical density profile.]

2.3 Aspects of Stellar Orbits

It was pointed out in the beginning of Section 2.2 that it is not possible to derive the structure and gravitational potential of galaxies by starting from first principles and studying their formation. To choose among various solutions of the CBE, like the ones described in the last section, it is necessary to compare their properties with observations. These observations are invariably related to the motion of the stars in the galactic potential, and hence it is worthwhile to discuss the orbits of stars in some general class of potentials that will be appropriate for the description of the galaxies.

When the stars are treated as point particles and the motion (as well as the gravitational field) is treated as nonrelativistic, the equation of motion for any given star is simply $\ddot{\mathbf{x}} = -\nabla\Phi(\mathbf{x})$, which can be integrated once the form of Φ is known. The analytical solutions for the orbits, of course, are available for only a small subset of all potentials; but because they offer valuable insight into the nature of motion, we begin with these cases.

2.3.1 *Spherically Symmetric Potentials*

In the case of spherically symmetric potentials, there are two constants of motion; namely the total angular momentum per unit mass $\mathbf{L} = \mathbf{r} \times \mathbf{v}$ and the energy per unit mass E. The orbit is confined to a plane perpendicular to the direction of $\mathbf{L}$ and the trajectory, in terms of the polar coordinates in the plane, $[r(t), \psi(t)]$, is given by

$$r^2\dot{\psi} = L, \qquad \dot{r} = \frac{dr}{dt} = \pm\sqrt{2[E - \Phi(r)] - \frac{L^2}{r^2}} \tag{2.32}$$

(see Vol. I, Chap. 2, Section 2.3). These equations reduce the problem to quadrature. In particular, the radial period is given by

$$T_r = 2\int_{r_1}^{r_2} \frac{dr}{\sqrt{2[E - \Phi(r)] - (L^2/r^2)}}, \tag{2.33}$$

where r_1 and r_2 are the turning points at which $\dot{r} = 0$. During this period, the azimuthal angle changes by the amount

$$\Delta\psi = 2\int_{r_1}^{r_2} \frac{d\psi}{dr}\,dr = 2\int_{r_1}^{r_2} \frac{L}{r^2}\frac{dt}{dr}\,dr = 2L\int_{r_1}^{r_2} \frac{dr}{r^2\sqrt{2[E - \Phi(r)] - (L^2/r^2)}}. \tag{2.34}$$

The corresponding azimuthal period is

$$T_\psi = \frac{2\pi}{\Delta\psi}\,T_r. \tag{2.35}$$

Because $\Delta\psi/2\pi$ will not – in general – be a rational number, the orbit will not be closed and will fill the annular region between circles of radii r_1 and r_2.

Among all spherically symmetric potentials, there are two that form natural boundaries for more realistic potentials. The first is the Kepler potential $\Phi \propto r^{-1}$ that arises if we assume that the entire mass of the galaxy is concentrated at the centre and the star is moving under its influence. The second is the harmonic potential with $\Phi \propto r^2$ that arises if all the mass in the galaxy is distributed uniformly within a sphere of radius r. Because any realistic mass distribution will be centrally condensed, it is likely to be between these two limiting forms. To study the two limiting cases, it is convenient to have a form of $\Phi(r)$ that interpolates between the two and for which analytic solutions are available. One such special potential is

$$\Phi = -\frac{GM}{b + \sqrt{(r^2 + b^2)}}, \tag{2.36}$$

called the *isochrone potential*. When $b \to 0$, this leads to the Kepler potential whereas for $b \to \infty$ the leading order behavior is (Φ+ constant) $\propto r^2$. This potential is generated by a density distribution of the form

$$\rho(r) = \frac{1}{4\pi G}\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d\Phi}{dr}\right) = M\left[\frac{3(b+a)a^2 - r^2(b+3a)}{4\pi(b+a)^3a^3}\right], \tag{2.37}$$

where $a = \sqrt{b^2 + r^2}$, which has the limiting forms

$$\rho(0) = \frac{3M}{16\pi Gb^3}, \qquad \rho(r) \simeq \frac{bM}{2\pi r^4} \quad (r \gg b). \tag{2.38}$$

To integrate the equations of motion in this potential, it is convenient to define a variable s by

$$s \equiv -\frac{GM}{b\Phi} = 1 + \sqrt{\frac{r^2}{b^2} + 1}, \qquad \frac{r^2}{b^2} = s^2\left(1 - \frac{2}{s}\right), \qquad s \geq 2. \tag{2.39}$$

We can now express the integrals that determine the trajectory of the particle in terms of s. Simple analysis gives the radial period as

$$T_r = \frac{2b}{\sqrt{-2E}}\int_{s_1}^{s_2}\frac{(s-1)\,ds}{\sqrt{(s_2 - s)(s - s_1)}} = \frac{2\pi b}{\sqrt{-2E}}\left[\frac{1}{2}(s_1 + s_2) - 1\right], \tag{2.40}$$

where s_1 and s_2 corresponds to the turning points r_1 and r_2 and we have assumed that $E < 0$. These roots are determined by

$$2Es^2 - 2\left(2E - \frac{GM}{b}\right)s - \frac{4GM}{b} - \frac{L^2}{b^2} = 0, \tag{2.41}$$

from which it follows that the sum of the roots is given by $s_1 + s_2 = 2[1 - (GM/2Eb)]$; hence the radial period is

$$T_r = \frac{2\pi GM}{(-2E)^{3/2}}. \tag{2.42}$$

This result is independent of b and is, of course, valid even in the Keplerian limit. (The fact that the radial period is independent of the angular momentum lends the name *isochrone* to this potential.) The change in the azimuthal angle per cycle is given by

$$\begin{aligned}\Delta\psi &= 2L\int_{s_1}^{s_2}\frac{dt}{r^2} = \frac{2L}{b\sqrt{-2E}}\int_{s_1}^{s_2}\frac{(s-1)\,ds}{s(s-2)\sqrt{(s_2-s)(s-s_1)}} \\ &= \pi\left(1+\frac{L}{\sqrt{L^2+4GMb}}\right).\end{aligned} \tag{2.43}$$

In the limit of $b \to 0$, $\Delta\psi \to 2\pi$, which is the Keplerian result. On the other hand, as $b \to \infty$, $\Delta\psi \to \pi$, which is characteristic of orbits in the harmonic potential. In general, we have $\pi < \Delta\psi < 2\pi$. Because galaxies are more extended than point masses, a typical star in a galaxy completes one radial oscillation in a shorter time than is required for one complete azimuthal cycle around the galactic centre.

2.3.2 Rotation Curves of Disk Galaxies

We next consider issues related to orbits in axisymmetric systems like disk galaxies. There are several interesting features that emerge in this study, and we begin with the question of determining the circular velocity of a particle orbiting in the plane of a disk galaxy. We saw in Chap. 1, Section 1.5, that the surface-density profile of disks is fairly well described by the exponential form $\Sigma(R) = \Sigma_0 \exp(-R/R_d)$. We shall now determine the gravitational potential that is due to such a disk and the properties of stellar orbits.

The easiest way to determine the potential will be to begin by noting that, in axially symmetric cases, the functions

$$\Phi_{\pm}(R, z) = \exp(\pm kz)\, J_0(kR) \tag{2.44}$$

(where J_0 is the Bessel function) are solutions to the Laplace equation. Consider now the function defined by

$$\Phi_k(R, z) = \exp(-k|z|)\, J_0(kR), \tag{2.45}$$

which has the discontinuity in the derivative at $z = 0$ given by

$$\lim_{z\to 0+}\left(\frac{\partial\Phi_k}{\partial z}\right) = -kJ_0(kR), \qquad \lim_{z\to 0-}\left(\frac{\partial\Phi_k}{\partial z}\right) = +kJ_0(kR). \tag{2.46}$$

It follows from Gauss theorem that Φ_k is the gravitational potential that is due to the surface-density distribution,

$$\Sigma_k(R) = -\frac{k}{2\pi G} J_0(kR), \tag{2.47}$$

located at $z = 0$. Superposing the density distribution by using a weightage function $S(k)$, we find that the surface-density profile,

$$\Sigma(R) = \int_0^\infty S(k)\Sigma_k(R)\,dk = -\frac{1}{2\pi G}\int_0^\infty S(k)J_0(kR)\,k\,dk, \tag{2.48}$$

will generate the gravitational potential

$$\Phi(R, z) = \int_0^\infty S(k)\Phi_k(R, z)\,dk = \int_0^\infty S(k)J_0(kR)e^{-k|z|}\,dk. \tag{2.49}$$

This analysis is sufficiently general and could have been done with a wide class of radial functions. The advantage of choosing $J_0(kR)$ lies in the fact that Eq. (2.48) can be inverted to give

$$S(k) = -2\pi G\int_0^\infty J_0(kR)\Sigma(R)R\,dR. \tag{2.50}$$

Using this in Eq. (2.49), we can express the potential directly in terms of the density $\Sigma(R)$, thereby providing the complete solution to the problem. Given the form of the potential, we can also determine the rotation velocity in terms of $S(k)$:

$$v_c^2(R) = R\left(\frac{\partial\Phi}{\partial R}\right)_{z=0} = -R\int_0^\infty S(k)J_1(kR)k\,dk. \tag{2.51}$$

For the case of exponential disks, we can evaluate the integral in Eq. (2.50) to obtain $S(k)$ as

$$S(k) = -\frac{2\pi G\Sigma_0 R_d^2}{[1+(kR_d)^2]^{3/2}}. \tag{2.52}$$

The potential is now given by Eq. (2.49):

$$\Phi(R, z) = -2\pi G\Sigma_0 R_d^2\int_0^\infty \frac{J_0(kR)e^{-k|z|}}{[1+(kR_d)^2]^{3/2}}\,dk, \tag{2.53}$$

which unfortunately cannot be expressed in simple closed form for arbitrary z. In the $z = 0$ plane, however, the result can be expressed in terms of modified Bessel functions by

$$\Phi(R, 0) = -\pi G\Sigma_0 R\,[I_0(y)K_1(y) - I_1(y)K_0(y)], \qquad y \equiv \frac{R}{2R_d}, \tag{2.54}$$

with the corresponding circular speed

$$v_c^2(R) = R\left(\frac{\partial \Phi}{\partial R}\right) = 4\pi G \Sigma_0 R_d y^2 [I_0(y)K_0(y) - I_1(y)K_1(y)]. \quad (2.55)$$

Figure 2.2 shows $v_c^2(R)/4\pi G\Sigma_0 R_d$ as a function of (R/R_d). A reasonable approximation to v_c for $R \lesssim 4R_d$ is given by

$$v_c(R) \simeq 0.876\left(\frac{GM}{R_d}\right)^{1/2}\left(\frac{\bar{y}^{1.3}}{1+\bar{y}^{2.3}}\right)^{1/2}, \quad (2.56)$$

where $\bar{y} = 0.533(R/R_d)$ and M is the total mass of the disk. This reaches a maximum around $(R/R_d)_{\max} \simeq 2.3$ and falls as $v_c \propto R^{-1/2}$ for $(R/R_d) \gtrsim 7$.

Exercise 2.5

Embedding a disk in a halo: Consider a system with the density distribution $\rho(r, \theta) = \rho_0 S(\theta)\,(r_0/r)^2$ that is axisymmetric and falls as r^{-2}. Let $S(\theta)$ be normalised such that $\int_0^\pi S(\theta)\sin\theta\, d\theta = 2$. In general, we would have expected the radial and angular components of the force,

$$F_r = -\frac{\partial \Phi}{\partial r}, \qquad F_\theta = -\frac{1}{r}\frac{\partial \Phi}{\partial \theta}, \quad (2.57)$$

to depend on θ as well as on r. (a) Show that $(\partial F_r/\partial\theta) = 0$, i.e., the radial force is

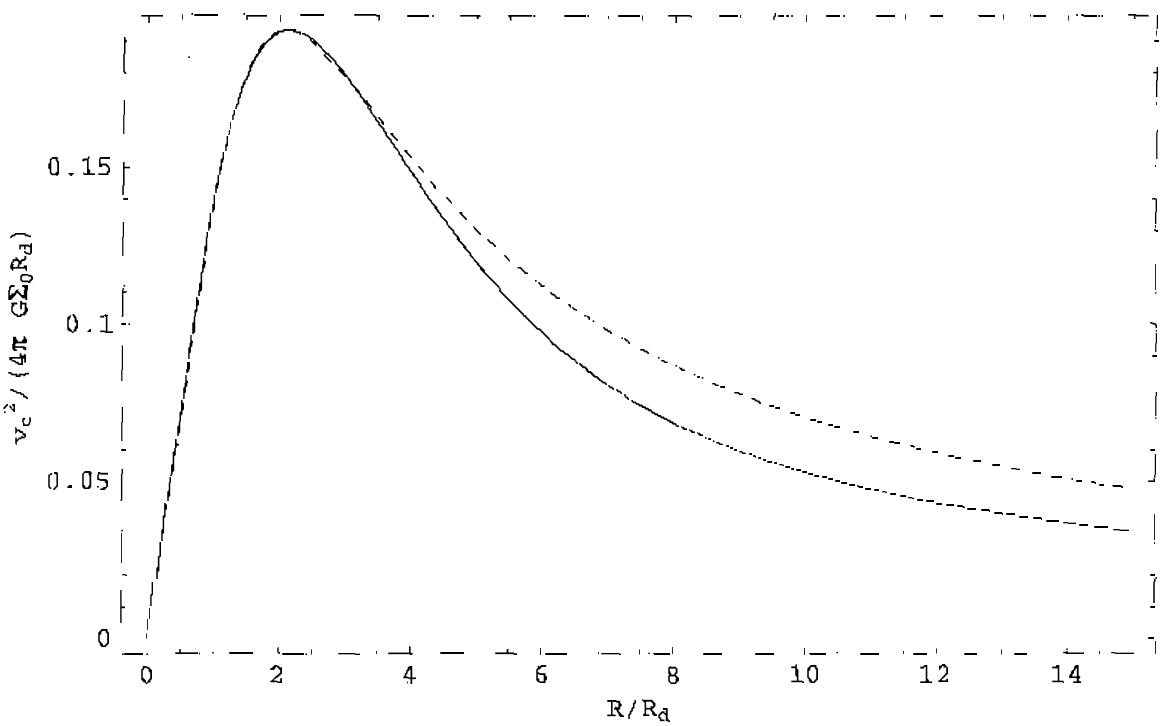


Fig. 2.2. Circular velocity of rotation as a function of radius in an exponential disk based on Eq. (2.55). The rotational velocity starts decreasing within a few factors of R_d. The dashed curve is the analytic fit given by relation (2.56).

independent of the latitude. Hence show that $F_r(r, \theta) = F_r(r) = -4\pi G\rho_0 r_0^2/r \equiv -v_0^2/r$. (b) Show that the potential must have the form

$$\Phi(r, \theta) = v_0^2[\ln(r/r_0) + P(\theta)]. \tag{2.58}$$

Find the equation connecting $P(\theta)$ and $S(\theta)$. (c) What is the potential if $S(\theta) = a\delta_{\text{Dirac}}(\theta - \pi/2) + b$, which corresponds to the disk embedded in a spherical halo?

Exercise 2.6

Density from rotation curves: Obtain an expression for $\Sigma(R)$ in terms of $v_c^2(R)$. [Hint: Starting with the result in Eq. (2.50), use the integral expression for $\Sigma(R)$ in terms of $S_0(k)$.]

Exercise 2.7

Asymptotic form for potentials: Consider an axisymmetric body with density distribution $\rho(R, z)$ that is symmetric about the z axis and vanishes for $r > r_{\max}$. Show that, for distances large compared with $r_{\max}$, the gravitational potential is given by

$$\Phi(R, z) \simeq -\frac{GM}{r} - \frac{G}{4}\frac{(R^2 - 2z^2)}{r^5}\int \rho(R', z')(R'^2 - 2z'^2)\, d^3\mathbf{r}'. \tag{2.59}$$

Apply this formula for the exponential disk with $\Sigma(R) = \Sigma_0 \exp(-R/R_d)$ and show that the potential is given by

$$\Phi(R, z) \simeq -\frac{GM}{r}\left[1 + \frac{3R_d^2(R^2 - 2z^2)}{2r^4}\right], \tag{2.60}$$

where M is the mass of the disk.

The potential for the exponential disk is of some practical significance in modelling the rotation curves of galaxies. The asymptotic behavior of Bessel functions in Eq. (2.55) (see Fig. 2.2) shows that the rotation curve that is due to the exponential disk alone will decrease with radius as $R \to \infty$. Because the observed rotation curve is either flat or even gently rising with radius, it is usual to assume that the density distribution has another component that is possibly due to a spherically distributed halo. The density profile of the halo needs to be obtained by a comparison of the observed rotation curve with the one that is due to the exponential disk. To illustrate the uncertainties involved in this modelling, let us consider a simple halo profile of the form

$$\rho_h(r) = \frac{\rho_h(0)}{1 + (r/a)^\gamma}. \tag{2.61}$$

This halo produces a rotational velocity given by

$$v_h^2(r) = \frac{GM_h(r)}{r} = \frac{4\pi G}{r}\int_0^r dx\, x^2 \rho_h(x). \tag{2.62}$$

The observed rotation curve in the plane of the galaxy is given by the addition

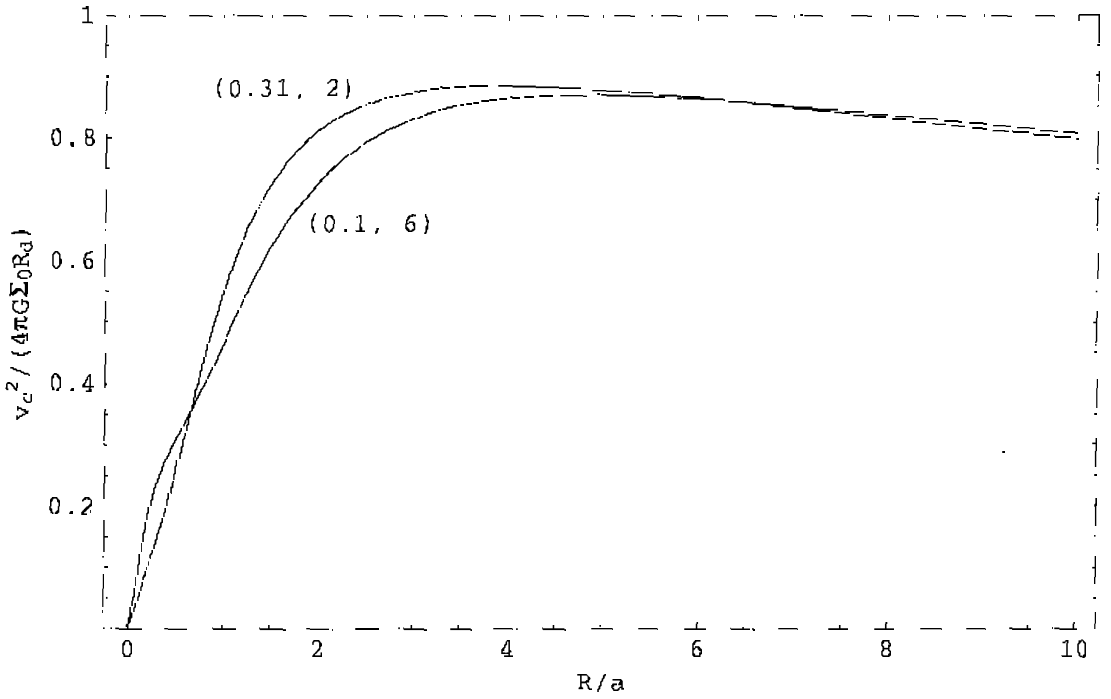


Fig. 2.3. Circular velocity as a function of radius for stars in a disk embedded in a halo. The curves have very different parameter values but still produce similar rotational velocities.

of the contributions in the quadrature: $v_c^2 = v_{\rm disk}^2 + v_h^2$. Figure 2.3 illustrates this degeneracy in a simple situation. The y axis gives v_c^2 in units of $4\pi G\,\Sigma_0 R_d$ as a function of R/a. The curves are parameterised by the two ratios $(q_1, q_2) \equiv [\Sigma_0 R_d^2/\rho_h(0)a^3, a/R_d]$ in which the first parameter q_1 essentially gives the ratio of the disk mass to the halo mass and the second parameter q_2 gives the scale length of halo to the disk. The two curves are for $(q_1, q_2) = (0.31, 2)$ and $(q_1, q_2) = (0.1, 6)$ with $\gamma = 2.4$. One of the curves has a halo mass that is approximately three times the disk mass whereas the other one has a halo that is 10 times as massive as the disk. It is clear from the figure that the rotation curve at large r is essentially the same, indicating that we cannot distinguish between these two models, for example, from the asymptotic form of the rotation curves.

The difficulty described above persists in even more sophisticated modelling. Assuming a constant (M/L) ratio for the disk, and a disk scale length of $\sim$3 kpc, good fits to observed rotation curves can be obtained for a wide class of model parameters. The model with maximum disk mass has $(M/L) \simeq 3.6$, consistent with the result for reasonable stellar population. However, this does not, in general, determine all the halo properties. For example, we can obtain equally good fits by varying in a correlated fashion γ and a of Eq. (2.61) in the range $1.9 < \gamma < 2.9$ and 7 kpc $< a <$ 12 kpc. The core radius of the halo for this profile [usually defined to be $r_{\rm core} = a(2\sqrt{2} - 1)^{1/\gamma}$], however, does not vary

much and is given by $r_{\text{core}} = 12.5 \pm 1.5$ kpc. Thus rotation curves, by themselves, do not fix the disk and halo properties uniquely.

2.3.3 Epicyclic Approximation in Axisymmetric Potentials

A convenient coordinate system in the case of an axisymmetric galaxy is the cylindrical one with (R, ϕ) denoting the polar coordinates of a $z =$ constant plane, with z being measured along the axis of symmetry. The gravitational potential is $\Phi = \Phi(R, z)$ and will be independent of the angular coordinate ϕ.

If a star is confined to move in a $z =$ constant plane, then it cannot distinguish an axially symmetric potential from a spherically symmetric one, and the problem reduces to the one discussed in Subsection 2.3.1. Even when the motion of the star takes it out of the meridinal plane, it is possible to reduce the problem to a two-dimensional one along the following lines. The equation of motion $\ddot{\mathbf{r}} = -\nabla\Phi$ can be written in component form in the cylindrical coordinates as follows:

$$\frac{d}{dt}(R^2\dot{\phi}) = 0, \tag{2.63}$$

$$\ddot{R} - R\dot{\phi}^2 = -\frac{\partial\Phi}{\partial R}, \qquad \ddot{z} = -\frac{\partial\Phi}{\partial z}. \tag{2.64}$$

Equation (2.63) represents the conservation of the z component of angular momentum (per unit mass) $L_z = R^2\dot{\phi}$. Using this, the Eqs. (2.64) can be reduced to that of a two-dimensional problem with an effective potential Φ_{eff} in the form

$$\ddot{R} = -\frac{\partial\Phi_{\text{eff}}}{\partial R}, \qquad \ddot{z} = -\frac{\partial\Phi_{\text{eff}}}{\partial R}, \qquad \Phi_{\text{eff}} \equiv \Phi(R, z) + \frac{L_z^2}{2R^2}. \tag{2.65}$$

The dynamics can be determined once $\Phi(R, z)$ is given. The general behaviour of motion in such potentials has been discussed using surfaces of section in Vol. I, Chap. 2, Section 2.8.

From a practical point of view, it is interesting to study the behaviour of orbits near the minimum of the effective potential without assuming a specific form for $\Phi(R, z)$. The condition for the minimum, $\nabla\Phi_{\text{eff}} = 0$, has a fairly simple interpretation. The vanishing of the gradient with respect to z is satisfied everywhere in the equatorial plane $z = 0$ whereas the condition $(\partial\Phi_{\text{eff}}/\partial R) = 0$ is satisfied at a radius R_g, where

$$\left(\frac{\partial\Phi}{\partial R}\right)_{(R_g, z=0)} = \frac{L_z^2}{R_g^3} = R_g\dot{\phi}^2. \tag{2.66}$$

This is a condition for the existence of a circular orbit with angular speed $\dot{\phi}$. Thus the minimum of Φ_{eff} occurs at the radius of the circular orbit corresponding to the angular momentum L_z. To study the deviation from the circular motion, it is convenient to define a coordinate $x \equiv (R - R_g)$ and expand the potential in a

Taylor series around $(x, z) = (0, 0)$. To the lowest order, this will give a harmonic potential of the form

$$\Phi_{\rm eff} \approx \frac{1}{2}(\kappa^2 x^2 + \nu^2 z^2) + \mathcal{O}(xz^2) + \text{constant}, \tag{2.67}$$

with the two frequencies (κ, ν) defined by the relations

$$\kappa^2 \equiv \left(\frac{\partial^2 \Phi_{\rm eff}}{\partial R^2}\right)_{(R_g,0)} = \left(\frac{\partial^2 \Phi}{\partial R^2}\right)_{(R_g,0)} + \frac{3L_z^2}{R_g^4}, \qquad \nu^2 = \left(\frac{\partial^2 \Phi}{\partial R^2}\right)_{(R_g,0)}. \tag{2.68}$$

Because the circular frequency at a radius R is given by

$$\Omega^2(R) = \frac{1}{R}\left(\frac{\partial \Phi}{\partial R}\right)_{(R,0)} = \frac{L_z^2}{R^4}, \tag{2.69}$$

we can also write

$$\kappa^2 = \left(R\frac{d\Omega^2}{dR} + 4\Omega^2\right)_{R=R_g}. \tag{2.70}$$

More generally, we can use this relation to define the function $\kappa(R)$ at any radius R. Usually Ω is nearly constant near the centre of a galaxy and decreases at larger radius. In this case we have $\Omega \leq \kappa \leq 2\Omega$.

Conventionally, we characterise the angular velocity $\Omega(R)$ in terms of two observationally determined quantities, called *Oort constants*. If the circular speed at the radius R is denoted by $v_c(R)$, then the Oort constants are defined by

$$A \equiv \frac{1}{2}\left(\frac{v_c}{R} - \frac{dv_c}{dR}\right), \qquad B \equiv -\frac{1}{2}\left(\frac{v_c}{R} + \frac{dv_c}{dR}\right). \tag{2.71}$$

From these definitions, it follows that $\Omega(R) = (A - B)$ and

$$A \equiv \frac{1}{2}\left(\frac{v_c}{R} - \frac{dv_c}{dR}\right) = -\frac{1}{2}\left(R\frac{d\Omega}{dR}\right), \qquad B = -\left(\frac{1}{2}R\frac{d\Omega}{dR} + \Omega\right). \tag{2.72}$$

For example, in our galaxy, at the location of the Sun $R = R_0 \simeq 8.5$ kpc, $v_c \approx 220$ km s^{-1}, $A \approx (14.8 \pm 0.8)$ km s^{-1} kpc^{-1}, and $B \approx (-12.4 \pm 0.6)$ km s^{-1} kpc^{-1}.[2] These imply that $(v_c/R) = (A - B) \approx 26.5$ km s^{-1} kpc^{-1}, which is consistent (within errors) with the value obtained from $v_c = 220$ km s^{-1} and $R_0 = 8.5$ kpc. Combining these, we get the value of κ at the location of the Sun as $\kappa_0^2 = -4B(A - B) = -4B\Omega_0$; the numerical value is $\kappa_0 \approx (36 \pm 10)$ km s^{-1} kpc^{-1}. Thus

$$\frac{\kappa_0}{\Omega_0} = 2\sqrt{\frac{-B}{A - B}} = 1.3 \pm 0.2, \tag{2.73}$$

implying that the Sun makes $\sim$1.3 oscillations in the radial direction by the time it completes one orbit around the galactic centre.

Given the harmonic approximation to the potential (2.67), it is easy to solve for the orbits. The x and z motions are oscillatory and are given by

$$z = Z\cos(\nu t + \zeta), \qquad x(t) = X\cos(\kappa t + \psi), \tag{2.74}$$

where Z and X are constants. To determine the azimuthal motion, we use the constancy of L_z and write

$$\dot{\phi} = \frac{L_z}{R^2} = \frac{L_z}{R_g^2}\left(1 + \frac{x}{R_g}\right)^{-2} \simeq \Omega_g\left(1 - \frac{2x}{R_g}\right). \tag{2.75}$$

Substituting for $x(t)$ from Eqs. (2.74) and integrating, we get

$$\phi = \Omega_g t + \phi_0 - \frac{2\Omega_g X}{\kappa R_g}\sin(\kappa t + \psi). \tag{2.76}$$

The first two terms on the right-hand side represent a rotation with angular velocity Ω_g that we can eliminate by introducing a local set of Cartesian axes (x, y, z) in which x and z are – as already defined – along the radial and the symmetry axes and the y coordinate is in the azimuthal direction. From Eq. (2.76), we get

$$y = -\frac{2\Omega_g}{\kappa}X\sin(\kappa t + \psi) \equiv -Y\sin(\kappa t + \psi). \tag{2.77}$$

Clearly the motion of the star in the $(x-y)$ plane is an ellipse with a ratio of the semiaxis of $(X/Y) = (\kappa/2\Omega_g)$. The centre of the ellipse rotates at the angular speed Ω_g around the galactic centre.

As an application of the preceding results (usually called the *guiding centre* or the *epicyclic approximation*) let us consider the question of determining the Oort constants from the observed motion of stars. In principle, the measurements of proper motions in the x and the y directions should yield the maximum values κX and κY, the ratio of which contains the necessary information. Such a direct measurement, of course, is not feasible because of the time scales involved. A more indirect way begins with the measurement of the radial velocity v_R and the differential angular azimuthal velocity $v_\phi(R_0) - v_c(R_0)$ for a group of stars as they pass near the Sun at the radius R_0. This quantity is given by

$$\begin{aligned} v_\phi(R_0) - v_c(R_0) &= R_0(\dot{\phi} - \Omega_0) = R_0(\dot{\phi} - \Omega_g + \Omega_g - \Omega_0) \\ &\simeq R_0\left[(\dot{\phi} - \Omega_g) - \left(\frac{d\Omega}{dR}\right)_{R_0} x\right] \simeq -R_0 x\left[\frac{2\Omega_g}{R_g} + \left(\frac{d\Omega}{dR}\right)_{R_0}\right], \end{aligned} \tag{2.78}$$

where we have used Eqs. (2.75) and (2.77). When (Ω_g/R_g) is replaced by (Ω_0/R_0) – which is permissible in the coefficient of the small quantity x –

and the definition of the Oort constant is used, it follows that

$$v_\phi - v_c(R_0) \simeq 2Bx. \tag{2.79}$$

Similarly, we have

$$\langle v_R^2 \rangle = \langle \kappa^2 x^2 \rangle \simeq -4B(A - B)\langle x^2 \rangle, \tag{2.80}$$

where the average is over many stars. This allows us to write

$$\frac{\langle [v_\phi - v_c(R_0)]^2 \rangle}{\langle v_R^2 \rangle} \simeq \frac{-B}{A - B} = \frac{\kappa_0^2}{4\Omega_0^2}, \tag{2.81}$$

where all the quantities on the left-hand side can be determined from observations.

Exercise 2.8

Determining a product: The line-of-sight velocity along the direction l can be expressed by Eq. (1.94) of Chap. 1. Show how observations of line-of-sight velocity can be used to determine the product AR_0. [Answer. Using the Taylor expansion

$$\Omega(R = R_0 \sin l) - \Omega(R_0) \simeq \left(\frac{d\Omega}{dR}\right)_{R_0} R_0(\sin l - 1), \tag{2.82}$$

which is valid near $l = \pi/2$, and expressing $(d\Omega/dR)$ in terms of A, we get $v(l) \approx 2AR_0(1 - \sin l)$ near $l = \pi/2$. Observations near $l = \pi/2$ will now allow the determination of the product AR_0 fairly accurately, even though neither A nor R_0 is known to the same precision.]

2.3.4 *Planar Nonaxisymmetric Potentials*

When even the assumption of axisymmetry is relaxed, the orbits becomes difficult to tackle analytically. Some progress can be achieved if the motion is confined to the meridinal plane $z = 0$, with the potential now depending explicitly on x and y. In general, most orbits in this potential possess an extra integral (in addition to energy). The majority of orbits can be classified either as *loop* orbits (which have a fixed sense of rotation and never take the star close to the centre) or *box* orbits (which have no fixed sense of rotation and allow the star to pass arbitrarily close to the centre). When the axial ratio of equipotential curves is close to unity, the phase space is filled mostly with loop orbits; as the axial ratio deviates away from unity, box orbits become more common.

These can be illustrated by the technique of surface of section described in Vol. I, Chap. 2, Section 2.8. For definiteness, we study the results of numerically integrating the equations of motion for the potential

$$\phi(x, y) = \frac{1}{2}v_0^2 \ln\left(R_c^2 + x^2 + \frac{y^2}{q^2}\right), \tag{2.83}$$

For a system with $N = 2$, the phase space is four dimensional and any point in the phase space has the coordinates (x, y, p_x, p_y). Conservation of energy restricts the motion to a three-dimensional surface defined by

$$\frac{1}{2}(\dot{x}^2 + \dot{y}^2) + \phi(x, y) = E = \text{constant}. \tag{2.84}$$

(We take the mass of the star to be unity and use the symbols $\dot{x}$ and $\dot{y}$ instead of p_x and p_y.) The intersection of this three-dimensional surface with the surface $y = 0$ defines the "surface of section" for the dynamical problem, which has the coordinates $(x, \dot{x})$. Given x, $\dot{x}$, and $y = 0$, $\dot{y}$ is determined as

$$\dot{y} = +\{2[E - \phi(x, 0)] - \dot{x}^2\}^{1/2}, \tag{2.85}$$

where we have chosen the positive square root by convention.

Given the energy E and suitable initial conditions, the equations of motion can be integrated to give the orbit of the particle. This orbit will repeatedly cut the surface of section at different values of x and $\dot{x}$. If there are no other integrals of motion, we expect these points to be scattered in a two-dimensional region bounded by the curve

$$\frac{1}{2}\dot{x}^2 + \phi(x, 0) \leq E. \tag{2.86}$$

In case the system has some other hidden integrals of motion, the orbit will be confined to a lower-dimensional surface in phase space and hence will form a smooth curve in the x–$\dot{x}$ plane.

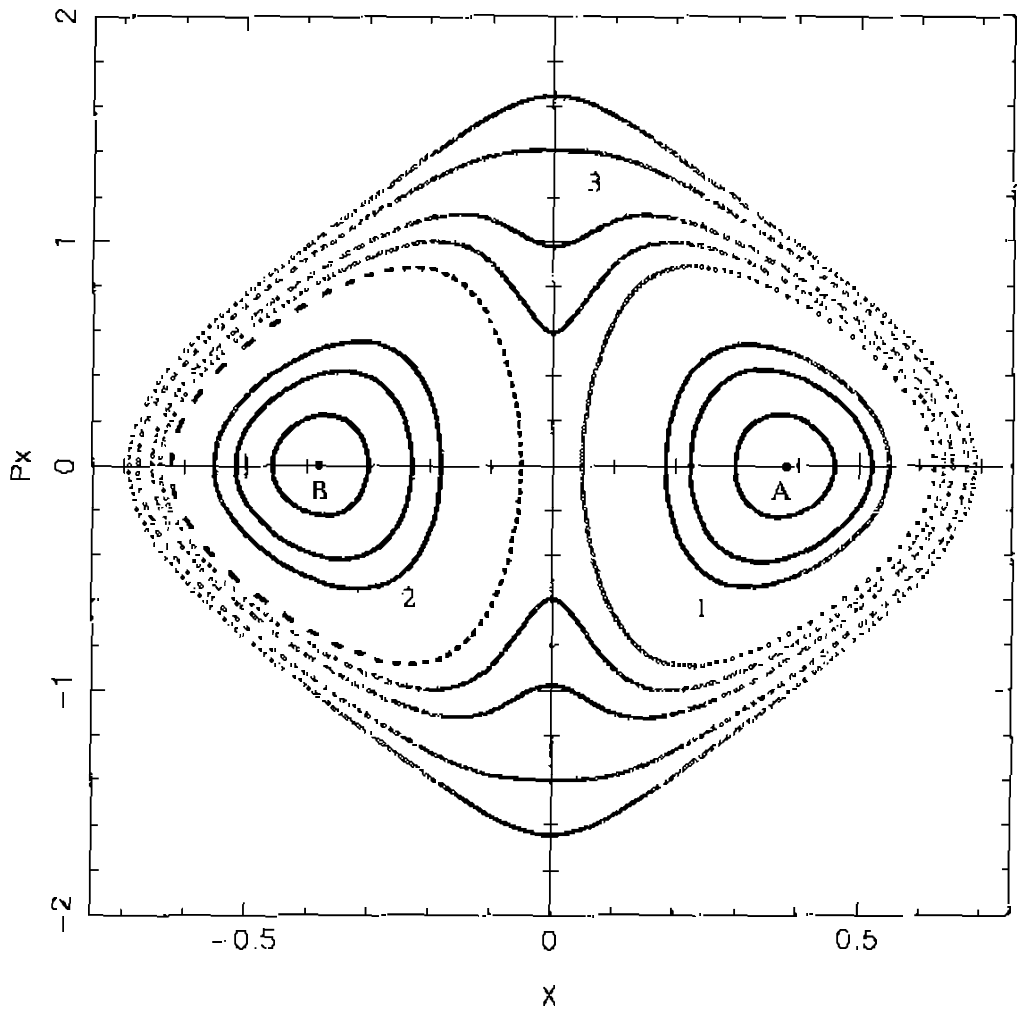
The surface of section for the potential is shown in Fig. 2.4 with $q = 0.9$, $R_c = 0.14$, and the conserved energy $E = -0.337$. The following conclusions can be drawn from the figure.

(1) The intersection of the orbit with the x–$\dot{x}$ plane occurs on reasonably smooth curves rather than on the entire allowed region in the plane. This suggests that there could be some other hidden integrals of motion.

(2) The orbits can be classified as belonging to the two different classes, usually called box orbits or loop orbits. Orbits like 1 or 2 (top panel of Fig. 2.4) revolve around the points A or B, respectively, and have a definite sense of rotation (loop orbits). These orbits do not pass through the origin. In contrast, orbits like 3 enclose both A and B and have no definite sense of rotation (box orbits). These orbits do cross the origin. The behaviour of these orbits in physical space is shown in the lower panel of Fig. 2.4. The two figures on the bottom right show the box orbits that oscillate through the origin with a superposed rotational motion. The two figures on the bottom left show the loop orbits that are confined to an annular region.

(3) As the initial conditions are changed, loop orbits can become more and more eccentric and the box orbits can become less and less elongated.

(4) The preceding discussion was based on a potential with $q = 0.9$. If $q = 1$ the potential depends on only $r^2 = x^2 + y^2$ and possesses an extra conserved quantity, namely the z component of the angular momentum, $J_z = x\dot{y} - y\dot{x} = x\dot{y}$ on the surface of



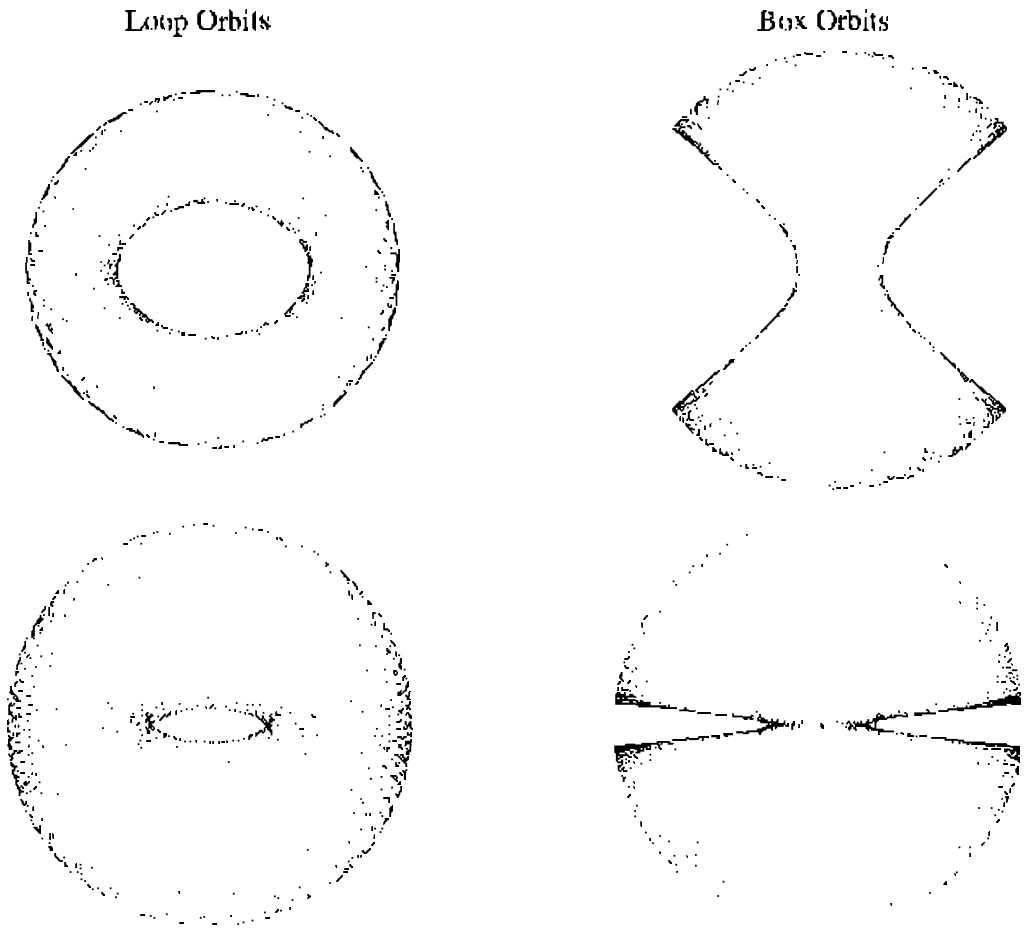


Fig. 2.4. Surfaces of section and orbits in an external potential; see text for discussion.

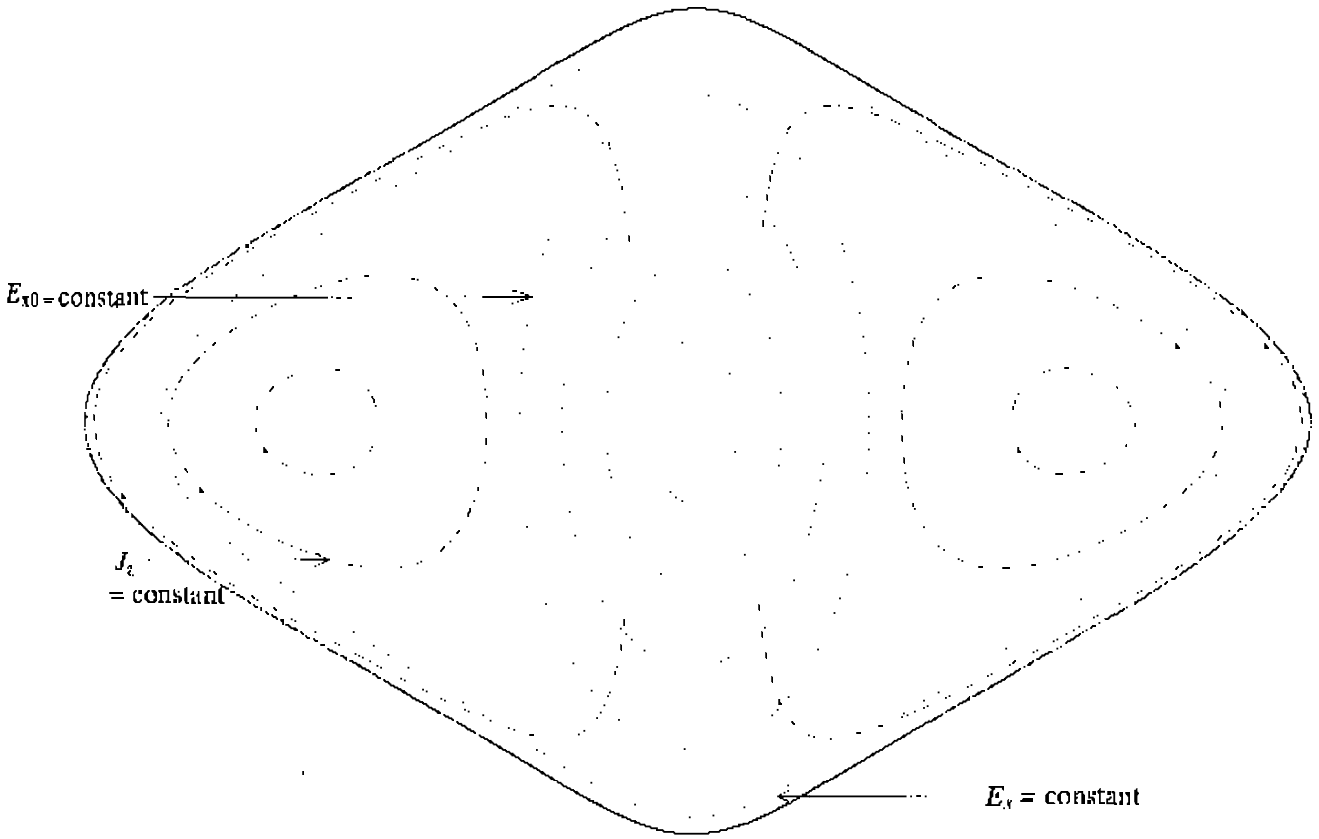


Fig. 2.5. Curves generated by the constancy of $x\dot{y}$ and E_x in the surface of section.

section with $y = 0$. In this case, all generic orbits are loop orbits. For $q \neq 1$, the expression for J_z becomes

$$(J_z)_{y=0} = x\dot{y} = x\left[2E - v_0^2 \ln\left(x^2 + R_c^2\right) - \dot{x}^2\right]^{1/2}. \tag{2.87}$$

Although this quantity is not strictly conserved when $q \neq 1$, it is useful to examine the nature of curves in the x–$\dot{x}$ plane, generated by the constancy of J_z. These are indicated in Fig. 2.5 by dotted–dashed curves. We see that these curves are similar to those corresponding to loop orbits. We may conclude that loop orbits owe their existence to an approximate conservation of a quantity like angular momentum.

(5) On the other extreme we may consider the motion in the potential when $r \ll R_c$, with $\phi(x, y)$ approximated as

$$\phi(x, y) \simeq \frac{v_0^2}{2R_c^2}\left(x^2 + \frac{y^2}{q^2}\right) + \text{constant}. \tag{2.88}$$

This is a double harmonic oscillator (discussed in Vol. I, Chap. 2, Section 2.5) which has an additional integral

$$E_{x0} = \frac{1}{2}\dot{x}^2 + \frac{v_0^2}{2R_c^2}x^2. \tag{2.89}$$

Constancy of E_{x0} will lead to the elliptical orbits shown in the middle of Fig. 2.5. We do not see evidence for such orbits in Fig. 2.4. However, if we generalize E_x to the form

$$E_x = \frac{1}{2}\dot{x}^2 + \phi(x, 0), \tag{2.90}$$

then we get better agreement with the orbits in Fig. 2.4. The outer dotted curve in Fig. 2.5

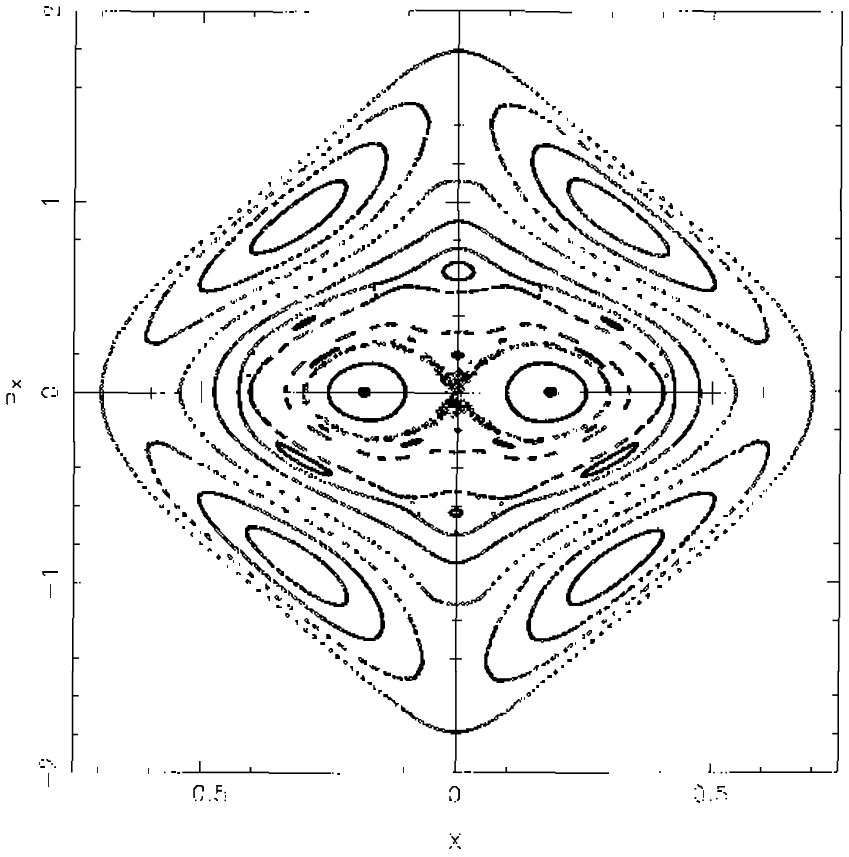


Fig. 2.6. Islands in the surface of section for $q = 0.6$.

corresponds to E_x = constant and is quite similar to the box orbits of Fig. 2.4. Thus box orbits owe their existence to an approximate conservation of the energy corresponding to x-axis motion.

(6) As q is decreased further, larger regions of phase space will be occupied by box orbits at the expense of loop orbits. When the potential is significantly nonaxisymmetric, several "islandlike" regions will form in the x–$\dot{x}$ plane with loop orbits around each of them. (See Fig. 2.6.) These results show that the stellar orbits in real potentials without any symmetries can be fairly complicated.

2.3.5 *Potentials in the Rotating Frame*

For most axisymmetric galaxies that have nonzero rotational motion about the z axis, the potential generated by the distribution of stars will be simplest in the corotating frame. Because such a frame is noninertial, the equations of motion pick up additional terms corresponding to Coriolis and centrifugal forces. If we take the galaxy as rotating with a bulk angular velocity $\Omega_b(>0)$ about the z axis, then the equation of motion in the rotating coordinate system is given by (see Vol. I, Chap. 2, Section 2.3)

$$\ddot{\mathbf{r}} = -\nabla\Phi - 2(\boldsymbol{\Omega}_b \times \dot{\mathbf{r}}) - \boldsymbol{\Omega}_b \times (\boldsymbol{\Omega}_b \times \mathbf{r}). \tag{2.91}$$

Let (R, ϕ) be the polar coordinates in the frame that rotates with the potential so that the equations of motion, in component form, are

$$\ddot{R} - R\dot{\phi}^2 = -\frac{\partial \Phi}{\partial R} + 2R\dot{\phi}\Omega_b + \Omega_b^2 R, \tag{2.92}$$

$$R\ddot{\phi} + 2\dot{R}\dot{\phi} = -\frac{1}{R}\frac{\partial \Phi}{\partial \phi} - 2\dot{R}\Omega_b. \tag{2.93}$$

As for the potential, we take it to be

$$\Phi(R, \phi) = \Phi_0(R) + \Phi_1(R, \phi), \qquad \Phi_1(R, \phi) = \Phi_b(R)\cos(m\phi), \tag{2.94}$$

with $(\Phi_1/\Phi_0) \ll 1$. This form for Φ_1 is suggested by the fact that most galaxies have a barlike structure that rotates with a pattern speed Ω_b. The perturbation Φ_1, having an angular variation $\cos(m\phi)$ with $m = 2$, is the simplest form that takes this effect into account. To study the motion, we assume the trajectories to be

$$R(t) = R_0 + R_1(t), \qquad \phi(t) = \phi_0(t) + \phi_1(t), \tag{2.95}$$

substitute into the equations of motion, and equate the zeroth- and first-order terms. The zeroth-order term for the radial equation will determine the radius R_0 of the centrifugal equilibrium at which the angular velocity is $\Omega_0 = \Omega(R_0)$, where

$$\Omega(R) \equiv \pm\sqrt{\frac{1}{R}\frac{d\Phi_0}{dR}}. \tag{2.96}$$

With a suitable choice for the origin of time, the angular motion – at the lowest order – is given by a steady rotation, $\phi_0 = (\Omega_0 - \Omega_b)t$. In the first-order perturbation equations, we can replace ϕ with ϕ_0 when multiplied by other first-order terms. Elementary algebra now gives these equations as

$$\ddot{R}_1 + \left(\frac{d^2\Phi_0}{dR^2} - \Omega^2\right)_{R_0} R_1 - 2R_0\Omega_0\dot{\phi}_1 = -\left(\frac{d\Phi_b}{dR}\right)_{R_0} \cos[m(\Omega_0 - \Omega_b)t], \tag{2.97}$$

$$\ddot{\phi}_1 + 2\Omega_0\frac{\dot{R}_1}{R_0} = \frac{m\Phi_b(R_0)}{R_0^2}\sin[m(\Omega_0 - \Omega_b)t]. \tag{2.98}$$

Integrating Eq. (2.98), we get

$$\dot{\phi}_1 = -2\Omega_0\frac{R_1}{R_0} - \frac{\Phi_b(R_0)}{R_0^2(\Omega_0 - \Omega_b)}\cos[m(\Omega_0 - \Omega_b)t] + \text{constant}, \tag{2.99}$$

which can be used to eliminate $\dot{\phi}_1$ from the radial equation, leading to

$$\ddot{R}_1 + \kappa_0^2 R_1 = -\left[\frac{d\Phi_b}{dR} + \frac{2\Omega\Phi_b}{R(\Omega - \Omega_b)}\right]_{R_0} \cos[m(\Omega_0 - \Omega_b)t], \tag{2.100}$$

where

$$\kappa_0^2 \equiv \left(\frac{d^2\Phi_0}{dR^2} + 3\Omega^2\right)_{R_0} = \left(R\frac{d\Omega^2}{dR} + 4\Omega^2\right)_{R_0} \tag{2.101}$$

is the square of the epicyclic frequency. [A constant has been ignored in arriving at Eq. (2.100) as it only leads to a shift in the origin of R_1.] Radial equation (2.100) is that of a forced harmonic oscillator with a driving frequency of $m(\Omega_0 - \Omega_b)$ and a natural frequency of κ_0. Its solution is given by

$$R_1(t) = C_1\cos(\kappa_0 t + \psi) - \left[\frac{d\Phi_b}{dR} + \frac{2\Omega\Phi_b}{R(\Omega - \Omega_b)}\right]_{R_0}\frac{\cos[m(\Omega_0 - \Omega_b)t]}{\Delta}, \tag{2.102}$$

where C_1 and ψ are constants and

$$\Delta \equiv \kappa_0^2 - m^2(\Omega_0 - \Omega_b)^2. \tag{2.103}$$

Eliminating t between $R_1(t)$ and $\phi_0(t)$, we can write the trajectory in the form

$$R_1(\phi_0) = C_1\cos\left[\frac{\kappa_0\phi_0}{(\Omega_0 - \Omega_b)} + \psi\right] + C_2\cos(m\phi_0), \tag{2.104}$$

with

$$C_2 \equiv -\frac{1}{\Delta}\left[\frac{d\Phi_b}{dR} + \frac{2\Omega\Phi_b}{R(\Omega - \Omega_b)}\right]_{R_0}. \tag{2.105}$$

Several conclusions can be drawn from this solution. To begin with, note that $R_1(\phi_0)$ is periodic in ϕ_0 if and only if $C_1 = 0$. This will correspond to a closed-loop orbit. Further, the amplitude C_2 diverges (leading to a resonance condition) under two circumstances. The first case (known as *corotation resonance*) occurs when $\Omega_0 = \Omega_b$. The second corresponds to the condition $m(\Omega_0 - \Omega_b) = \pm\kappa_0$ and is known as *Lindblad resonance*. Radii at which these resonances occur are called *inner* and *outer Lindblad radii*. These resonances have a simple interpretation: For a star in circular orbit, there are two natural frequencies of oscillation. Along the radial direction it is κ_0 whereas along the azimuthal direction it is zero as the circular orbit is neutrally stable. The forcing frequency $m(\Omega_0 - \Omega_b)$, seen by the star, is at resonance with the two natural frequencies κ_0 and 0 at the Lindblad resonance and at the corotation resonance, respectively. In most galaxies, the bar structure corresponds to $m = 2$; Fig. 2.7 shows the behaviour of Ω, $\Omega + (1/2)\kappa$ and $\Omega - (1/2)\kappa$ for the isochrone potential. Note that $\Omega - (1/2)\kappa$ is approximately constant.

The perturbation analysis given above breaks down near the resonances, and we need to refine our approximation. We do this most easily by rewriting the perturbation equations assuming that R_1, $\dot{R}_1$ and $\dot{\phi}_1$ are small but not ϕ_1. We

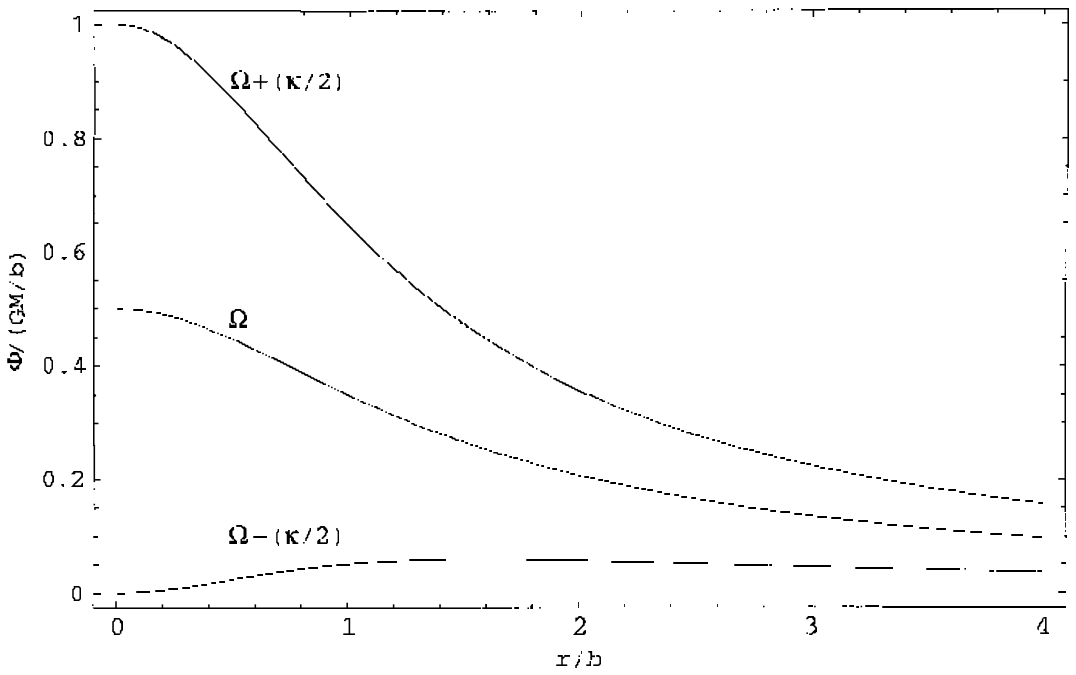


Fig. 2.7. The variation of Ω, $\Omega + (1/2)\kappa$ and $\Omega - (1/2)\kappa$ for the isochrone potential.

shall illustrate this procedure for the resonance at the corotation radius. With the guiding centre located at the corotation radius $[\Omega(R_0) = \Omega_b; \phi_0 = (\pi/2)]$, the perturbation equations now become

$$\ddot{R}_1 + \left(\kappa_0^2 - 4\Omega_0^2\right) R_1 - 2R_0\Omega_0\dot{\phi}_1 = -\frac{\partial \Phi_1}{\partial R}, \tag{2.106}$$

$$\ddot{\phi}_1 + 2\Omega_0 \frac{\dot{R}_1}{R_0} = -\frac{1}{R_0^2}\frac{\partial \Phi_1}{\partial \phi}. \tag{2.107}$$

Let us consider the relative strengths of various terms in this equation. If the nonaxisymmetric perturbation Φ_1 is of the order of a small parameter ϵ^2, then R_1 will be of the order of ϵ and ϕ_1 will be of the order of unity. The time derivatives of any quantity will be smaller by a factor ϵ compared with the original quantity. Therefore, in the radial equation, the three terms on the left-hand side are of the order of ϵ^3, ϵ, and ϵ, respectively, and the right-hand side is of the order of ϵ^2. Retaining only terms of the order of ϵ leads to the equation

$$\left(\kappa_0^2 - 4\Omega_0^2\right) R_1 - 2R_0\Omega_0\dot{\phi}_1 = 0. \tag{2.108}$$

Using this relation to eliminate R_1 from angular equation (2.107) and substituting for the explicit form of Φ_1 with $m = 2$, we get the equation for angular part as

$$\ddot{\phi}_1 = -\frac{2\Phi_b}{R_0^2}\left(\frac{4\Omega_0^2 - \kappa_0^2}{\kappa_0^2}\right)\sin[2(\phi_0 + \phi_1)]. \tag{2.109}$$

Taking $\Phi_b < 0$ and $\phi_0 = (\pi/2)$, we can recast this equation in the form

$$\frac{d^2\psi}{dt^2} = -p^2 \sin\psi, \tag{2.110}$$

where

$$\psi \equiv 2\phi_1, \qquad p^2 \equiv \frac{4}{R_0^2}|\Phi_b(R_0)|\frac{4\Omega_0^2 - \kappa_0^2}{\kappa_0^2}. \tag{2.111}$$

This is precisely the equation for an oscillating pendulum, and its solutions are known in terms of elliptic integrals. The singularity at the corotation radius has now disappeared because of the more careful analysis. The equilibrium point of the pendulum, $\phi_1 = 0$, is at the maximum of the potential Φ_1. If the integral of the motion $E_p = (1/2)\dot{\psi}^2 - p^2\cos\psi$ is less than p^2, the star librates about the equilibrium point, whereas if $E_p > p^2$, the star circulates around the centre of the galaxy. The shape of the orbit is given by

$$R_1 = -\frac{2R_0\Omega_0\dot{\phi}_1}{4\Omega_0^2 - \kappa_0^2} = \pm\left(\frac{R_0\Omega_0}{4\Omega_0^2 - \kappa_0^2}\right)\sqrt{2[E_p + p^2\cos(2\phi_1)]}, \tag{2.112}$$

which reduces to result (2.104) previously obtained in the limit of $E_p \gg p^2$. This result shows that there can be trapped orbits near the corotation radius. (It is possible to perform a similar analysis around the Lindblad radius, although it is mathematically more complicated.)

2.4 Application of the Jeans Equations

The discussion so far has concentrated on orbits of individual stars. We now take up the issue of studying the properties of a collection of stars described by a distribution function $f(\mathbf{x}, \mathbf{v}, t)$. It was mentioned in Vol. I, Chap. 10, Section 10.5, that one way of extracting useful information from the distribution function of stars in a galaxy will be to use the moments of the CBE and relate the different terms to the observed characteristics of a galaxy. It was shown there that the lowest-order moments lead to the equations

$$\frac{\partial\nu}{\partial t} + \frac{\partial(\nu\langle v_i\rangle)}{\partial x_i} = 0, \tag{2.113}$$

$$\nu\frac{\partial(\langle v_j\rangle)}{\partial t} + \nu\langle v_i\rangle\frac{\partial\langle v_j\rangle}{\partial x_i} = -\nu\frac{\partial\Phi}{\partial x_j} - \frac{\partial\left(\nu\sigma_{ij}^2\right)}{\partial x_i}, \tag{2.114}$$

where ν is the spatial density of stars, $\langle\mathbf{v}\rangle$ is the mean velocity, and σ_{ij}^2 is the

velocity dispersion, defined through the relations

$$\nu \equiv \int f\, d^3\mathbf{v}, \qquad \langle v_i \rangle \equiv \frac{1}{\nu} \int f\, v_i\, d^3\mathbf{v},$$
$$\sigma_{ij}^2 \equiv \langle (v_i - \langle v_i \rangle)(v_j - \langle v_j \rangle) \rangle = \langle v_i v_j \rangle - \langle v_i \rangle \langle v_j \rangle. \tag{2.115}$$

Equation (2.113) expresses the conservation of mass, and Eq. (2.114) is the force equation in the stellar-dynamics context. The terms on the left-hand side of Eq. (2.114) give the acceleration; the first term on the right-hand side is the gravitational force and the second term arises from the velocity dispersion of particles. (These are called the *Jeans equations*.) If the system was collisional then this velocity dispersion could have been related to a pressure that in turn could be related to the density by an equation of state, thereby closing the system of equations. For a collisionless system, there is no simple way of closing the moment equations because there is no equation of state.

Because σ_{ij}^2 is a symmetric second-rank tensor, we can always choose the coordinate system such that it is diagonalised to the form $\sigma_{ij} = \text{dia}(\sigma_{11}, \sigma_{22}, \sigma_{33})$. An ellipsoid in this coordinate system with the principal axis along the coordinates and $(\sigma_{11}, \sigma_{22}, \sigma_{33})$ as its semiaxis lengths is called the *velocity ellipsoid*. Note that, in general, σ_{ij}^2 is a function of $\mathbf{x}$ and t; in steady state, the velocity ellipsoid will still depend on the location $\mathbf{x}$. The use of the above equations therefore relies heavily on the additional physical assumptions that we are prepared to make regarding the various averages that appear in these equations. We now consider several such applications of the Jeans equations.

2.4.1 Asymmetric Drift

Observations suggest that stellar populations in our galaxy show a circular velocity v_c that is different from the mean azimuthal velocity $\bar{v}_\phi$. We typically find that this difference is $v_a \equiv (v_c - \bar{v}_\phi) \approx \langle v_R^2 \rangle / D$, with $D \simeq 120$ km s^{-1}. This phenomenon, called *asymmetric drift*, can be understood in terms of the Jeans equations. To begin with, we write the Jeans equations in cylindrical polar coordinates by the standard procedure of multiplying the CBE by velocity components and integrating over the velocities. This gives

$$\frac{\partial(\nu\langle v_R \rangle)}{\partial t} + \frac{\partial\left(\nu\langle v_R^2 \rangle\right)}{\partial R} + \frac{\partial(\nu\langle v_R v_z \rangle)}{\partial z} + \nu\left(\frac{\langle v_R^2 \rangle - \langle v_\phi^2 \rangle}{R} + \frac{\partial \Phi}{\partial R}\right) = 0, \tag{2.116}$$

$$\frac{\partial(\nu\langle v_\phi \rangle)}{\partial t} + \frac{\partial(\nu\langle v_R v_\phi \rangle)}{\partial R} + \frac{\partial(\nu\langle v_\phi v_z \rangle)}{\partial z} + \frac{2\nu}{R}\langle v_\phi v_R \rangle = 0, \tag{2.117}$$

$$\frac{\partial(\nu\langle v_z \rangle)}{\partial t} + \frac{\partial(\nu\langle v_R v_z \rangle)}{\partial R} + \frac{\partial\left(\nu\langle v_z^2 \rangle\right)}{\partial z} + \frac{\nu\langle v_R v_\phi \rangle}{R} + \nu\frac{\partial \Phi}{\partial z} = 0. \tag{2.118}$$

Near the galactic equator ($z = 0$) we can take $(\partial v/\partial z) = 0$, and in steady state we can ignore the time derivatives in Eq. (2.116). Using the definitions $v_c^2 = R(\partial\Phi/\partial R)$, $v_a \equiv (v_c - \langle v_\phi \rangle)$, and defining the velocity dispersion $\sigma_\phi^2 = \langle v_\phi^2 \rangle - \langle v_\phi \rangle^2$, we find that Eq. (2.116) becomes

$$v_a(2v_c - v_a) \simeq 2v_c v_a \simeq \langle v_R^2 \rangle \left[\frac{\sigma_\phi^2}{\langle v_R^2 \rangle} - 1 - \frac{\partial \ln \left(\nu \langle v_R^2 \rangle \right)}{\partial \ln R} - \frac{R}{\langle v_R^2 \rangle} \frac{\partial (\langle v_R v_z \rangle)}{\partial z} \right], \tag{2.119}$$

where the first equality follows from the fact that we expect $v_a \ll v_c$.

Relation (2.119) is typical of the relations that we can obtain from the Jeans equations. To proceed further, we need to introduce a series of reasonable assumptions based on observational inputs. To begin with, let us consider the term $[\partial \ln(\nu \langle v_R^2 \rangle)/\partial \ln R]$. If the disk galaxy has an exponential profile, with a small thickness h_z in the vertical direction, we would expect $h_z \propto \sigma_{zz}^2/\Sigma$ with $\Sigma \propto \exp(-R/R_d)$. This suggests that $\sigma_{zz}^2 \propto \exp(-R/R_d)$. Further, if we take the velocity ellipsoid to be approximately constant in space so that $\sigma_{zz}^2 \propto \sigma_{RR}^2 \equiv \langle v_R^2 \rangle$, we can write

$$\frac{\partial \ln \left(\nu \langle v_R^2 \rangle \right)}{\partial \ln R} \simeq 2 \frac{\partial \ln \nu}{\partial \ln R}. \tag{2.120}$$

To estimate the other derivatives, we consider two extreme possibilities: (1) The principal axes of the velocity ellipsoid are in the direction of cylindrical coordinate axes or (2) the principal axes rotate in such a way as to retain alignment with the spherical coordinates centred on the galactic nucleus. In the first case $\langle v_R v_z \rangle$ is independent of z, and in the second case $\langle v_R v_z \rangle \simeq (\langle v_R^2 \rangle - \langle v_z^2 \rangle)(z/R)$. (This relation is most easily derived if v_R, v_z is expressed in terms of v_r, v_θ and the result is averaged with $\langle v_r v_\theta \rangle = 0$.) Taking the average of the derivatives in these two extreme cases and substituting in relation (2.119) we get the range of possibilities as given by

$$\frac{2v_c v_a}{\langle v_R^2 \rangle} \simeq \left[\frac{\sigma_\phi^2}{\langle v_R^2 \rangle} - \frac{3}{2} - 2 \frac{\partial \ln \nu}{\partial \ln R} + \frac{1}{2} \frac{\langle v_z^2 \rangle}{\langle v_R^2 \rangle} \pm \frac{1}{2} \left(\frac{\langle v_z^2 \rangle}{\langle v_R^2 \rangle} - 1 \right) \right]. \tag{2.121}$$

From the analysis based on Oort constants [see relation (2.81)], we know that $\sigma_\phi^2 \simeq \langle v_z^2 \rangle \simeq 0.45 v_R^2$. Taking the disk of our galaxy to be exponential, with $\nu = \nu_0 \exp(-R/R_d)$, $(R/R_d) = 2.4$, and $v_c = 220$ km s^{-1}, we get the final result,

$$v_a \simeq \frac{\langle v_R^2 \rangle}{(110 \pm 7 \text{ km s}^{-1})}; \tag{2.122}$$

which is in good agreement with observations.

2.4.2 Mass and Velocity Dispersion

Another application of the Jeans equations is in estimating the effects of anisotropic velocity dispersion on other derived quantities. Consider, for example, the problem of determining the mass $M(<R)$ contained within a projected radius R of a galaxy that contains a population of tracers (say, planetary nebulas) distributed in a spherically symmetric manner with a volume number density $n(r) \propto r^{-\gamma}$. Let us assume that the velocity dispersion is anisotropic with $\langle v_\theta^2 \rangle / \langle v_r^2 \rangle = 1 - \beta$; when $\beta < 0$ the system is dominated by tangential anisotropy whereas for $0 < \beta < 1$, the radial anisotropy dominates. (Spherical symmetry, of course, implies that $\langle v_\theta^2 \rangle = \langle v_\phi^2 \rangle$.) The Jeans equations now become

$$\frac{r}{n}\frac{d}{dr}\left(n\sigma_r^2\right) + 2\beta\sigma_r^2 = -r\frac{d\Phi}{dr} = -v_c^2, \qquad \sigma_r^2 = \langle v_r^2 \rangle, \tag{2.123}$$

which can be expressed in the form

$$\left(\frac{d\ln n}{d\ln r} + \frac{d\ln \sigma_r^2}{d\ln r} + 2\beta\right)\sigma_r^2 = -v_c^2. \tag{2.124}$$

If the circular velocity is approximately constant, then the right-hand side of Eq. (2.124) is independent of r; if we further assume that σ_r^2 is also independent of r and use $n \propto r^{-\gamma}$, then $\langle v_r^2 \rangle = v_c^2/(\gamma - 2\beta)$. The mass within a radius r is, of course, $M(<r) = v_c^2 r/G$ when v_c^2 is a constant. However, observations usually give the velocity dispersion $\langle v_l^2 \rangle$ along the line of sight rather than $\langle v_r^2 \rangle$. It is therefore necessary to relate these two quantities. If the radial distance to a point along the line of sight from the centre of the galaxy is r and the angle between the radial direction and the line of sight is $(\pi/2 - \theta)$, then the component v_z along the line of sight is

$$v_z = v_r \sin\theta + v_\theta \cos\theta. \tag{2.125}$$

Squaring and averaging, using $\langle v_r v_\theta \rangle = 0$, $\langle v_\theta^2 \rangle = (1-\beta)\langle v_r^2 \rangle$, we get

$$\langle v_z^2 \rangle = \langle v_r^2 \rangle (1 - \beta\cos^2\theta). \tag{2.126}$$

The integral of this velocity dispersion along the line of sight is given by

$$\langle v_l^2 \rangle = \frac{\int_{-\infty}^{\infty} \langle v_z^2 \rangle n\, dz}{\int_{-\infty}^{\infty} n\, dz}. \tag{2.127}$$

Using Eq. (2.126) and the relations $r = R\sec\theta$, $n(r) \propto R^{-\gamma}\cos^{\gamma}\theta$, and $z = R\tan\theta$, we can reduce this integral to the form

$$\langle v_l^2\rangle = \langle v_r^2\rangle \frac{\int_0^{\pi/2} (1-\beta\cos^2\theta)^{-1}\cos^{\gamma-2}\theta\, d\theta}{\int_0^{\pi/2}\cos^{\gamma-2}\theta\, d\theta}. \tag{2.128}$$

For $\gamma > 1$ this integral can be expressed in terms of beta functions $B(x, y)$, giving

$$\langle v_l^2\rangle = \frac{v_c^2}{(\gamma-2\beta)}\left(1-\beta\frac{B\left[\frac{1}{2},\frac{\gamma+1}{2}\right]}{B\left[\frac{1}{2},\frac{\gamma-1}{2}\right]}\right), \quad (\gamma > 1). \tag{2.129}$$

The integral, however, simplifies for two special cases, $\gamma = 2$ and 3; we then get

$$\langle v_l^2\rangle = \begin{cases} v_c^2\dfrac{[1-(\beta/2)]}{2(1-\beta)} & (\gamma = 2) \\ (v_c^2/3) & (\gamma = 3) \end{cases}. \tag{2.130}$$

We note that, if $\gamma = 3$, the result is independent of β, showing that the velocity anisotropy is not important. However, if $\gamma = 2$, we can only put an upper limit to $M(<r)$ from the observations of $\langle v_l^2\rangle$.

Exercise 2.9

Velocity dispersion and mass estimate: Plot a graph of $\langle v_l^2\rangle/v_c^2$ versus β for different values of γ. Explain why for $\gamma < 3$, $\langle v_l^2\rangle/v_c^2$ increases as β increases, whereas for $\gamma > 3$ it decreases as β increases. Also explain why $\langle v_l^2\rangle/v_c^2$ decreases with increasing γ for $\beta = 0$.

Exercise 2.10

Determining mass and velocity dispersion when $\beta = 0$: Let $I(R)$ and $\sigma_p^2(R)$ be the surface brightness and the line-of-sight velocity dispersion at the projected radius R. Show how these quantities can be used to determine $\nu(r)$ and $\langle v_r^2(r)\rangle$ if the system is spherically symmetric and $\beta = 0$. [Answer. When $\beta = 0$, Jeans equation (2.124) gives

$$M(r) = -\frac{r\langle v_r^2\rangle}{G}\left(\frac{d\ln\nu}{d\ln r} + \frac{d\ln\langle v_r^2\rangle}{d\ln r}\right). \tag{2.131}$$

Further, the projected quantities are related to actual three-dimensional variables by

$$I(R) = 2\int_R^{\infty}\frac{\nu r\, dr}{\sqrt{r^2-R^2}}, \qquad I(R)\sigma_p^2(R) = 2\int_R^{\infty}\frac{\nu\langle v_r^2\rangle r\, dr}{\sqrt{r^2-R^2}}. \tag{2.132}$$

These equations, being Abel integral equations, can be inverted to provide

$$\nu(r) = -\frac{1}{\pi}\int_r^\infty \frac{dI}{dR}\frac{dR}{\sqrt{R^2 - r^2}}, \qquad \nu(r)\langle v_r^2\rangle(r) = -\frac{1}{\pi}\int_r^\infty \frac{d(I\sigma_p^2)}{dR}\frac{dR}{\sqrt{R^2 - r^2}}. \tag{2.133}$$

The right-hand side is known from observations, allowing us to determine $\nu(r)$ and $\langle v_r^2(r)\rangle$. Once these are known, we can compute $M(r)$.]

2.4.3 Rotation of Elliptical Galaxies

The most natural explanation for the elliptical shape of some galaxies could be the flattening that is due to the systematic rotation of the components, much the same way as the ellipsoidal shape of Earth is due to the rotation. The Jeans equation gives a handle on analysing whether this is a feasible explanation.

To do this, we make the assumption that the velocity dispersion is isotropic, allowing us to write $\sigma_{ij}^2 = \sigma^2\delta_{ij}$. The Jeans equation for an axisymmetric galaxy in steady state, with isotropic velocity dispersion and streaming motion in only the azimuthal direction, becomes

$$\frac{\partial(\nu\sigma^2)}{\partial R} - \nu\left(\frac{\bar{v}_\phi^2}{R} - \frac{\partial\Phi}{\partial R}\right) = 0, \tag{2.134}$$

$$\frac{\partial(\nu\sigma^2)}{\partial z} + \nu\frac{\partial\Phi}{\partial z} = 0. \tag{2.135}$$

Integrating Eq. (2.135) from $z = \infty$ to some finite value and assuming that the density vanishes at infinity, we can express the velocity dispersion in terms of the gravitational potential:

$$\sigma^2(R, z) = \frac{1}{\nu}\int_z^\infty \nu\frac{\partial\Phi}{\partial z}\,dz. \tag{2.136}$$

Using this in Eq. (2.134) we can find $\bar{v}_\phi$ as

$$\bar{v}_\phi^2(R, z) = R\frac{\partial\Phi}{\partial R} + \frac{R}{\nu}\frac{\partial}{\partial R}\int_z^\infty \nu\frac{\partial\Phi}{\partial z}\,dz. \tag{2.137}$$

These equations completely determine the velocity dispersion and the rotational streaming velocity in terms of the gravitational potential and density. Given any model for the density distribution (which will determine the gravitational potential), we can use these equations to relate the rotational velocity and the velocity dispersion.

To illustrate the idea, we take the gravitational potential Φ to be given by the logarithmic form of Eq. (2.27), discussed in Subsection 2.3.4:

$$\Phi = \Phi_L(R, z) \equiv \frac{1}{2}v_0^2\ln\left(R_c^2 + R^2 + \frac{z^2}{q_\Phi^2}\right). \tag{2.138}$$

The mass density $(4\pi G)^{-1}\nabla^2\Phi$ corresponding to the potential in Eq. (2.138) is given by

$$\rho_L(R,z) = \left(\frac{v_0^2}{4\pi G q_\Phi^2}\right)\frac{\left(2q_\Phi^2+1\right)R_c^2 + R^2 + 2\left(1-\frac{1}{2}q_\Phi^{-2}\right)z^2}{\left(R_c^2 + R^2 + z^2 q_\Phi^{-2}\right)^2}. \quad (2.139)$$

It is, however, convenient to approximate this density distribution in a form similar to the potential by writing

$$\rho_L \simeq \nu_L(R,z) \equiv v_0^2 R_c^2\left(R_c^2 + R^2 + \frac{z^2}{q_v^2}\right)^{-1}. \quad (2.140)$$

This approximation is valid if $r \gg R_c$ and $1 \gg (1-q_\Phi) \simeq 0.3(1-q_v)$. The fractional error $\delta(R,z) = [\nu_L(R,z) - \rho_L(R,z)]/v_0^2$ for $q_v = 0.5$; $q_\Phi = 0.85$ is only ~15% at worst. The advantage of this approximation lies in the fact that the necessary integral in Eq. (2.137) can be performed in closed form. We get

$$\int_z^\infty \nu_L \frac{\partial \Phi_L}{\partial z}\,dz = \frac{\frac{1}{2}v_0 R_c^2 v_0^2}{\left(R^2+R_c^2\right)\left[(q_\Phi/q_v)^2 - 1\right]}\ln\left[\frac{\left(R^2+R_c^2\right)q_\Phi^2 + z^2}{\left(R^2+R_c^2\right)q_v^2 + z^2}\right], \quad (2.141)$$

which determines σ^2 everywhere. Evaluating the expression at $z = 0$ and using it in Eq. (2.137), we get the rotational velocity on the equatorial plane as

$$\bar{v}_\Phi^2(R,0) = \frac{v_0^2 R^2}{R^2+R_c^2}\left[1 - \frac{2\ln(q_\Phi/q_v)}{(q_\Phi/q_v)^2 - 1}\right]. \quad (2.142)$$

If ϵ_v and ϵ_Φ are the ellipticities of isodensity and isopotential surfaces, respectively, with $q_v = (1-\epsilon_v)$, etc., and $\epsilon_\Phi \simeq 0.3\epsilon_v$, we can expand $\bar{v}_\phi^2$ in the ellipticities to obtain

$$\bar{v}_\Phi^2(R,0) = \frac{v_0^2 R^2}{R^2+R_c^2}[(\epsilon_v - \epsilon_\Phi) + \cdots,]. \quad (2.143)$$

The circular speed v_c on the equatorial plane for a particle moving in this potential is given by

$$v_c = \frac{v_0 R}{\sqrt{R_c^2 + R^2}}. \quad (2.144)$$

A comparison of Eqs. (2.143) and (2.144) now shows that

$$(\bar{v}_\phi/v_c) \approx (\epsilon_v - \epsilon_\Phi)^{1/2} = (0.7\epsilon_v)^{1/2}. \quad (2.145)$$

Because the right-hand side varies only as $\sqrt{\epsilon_v}$, even a small ellipticity requires appreciable rotation speed for its support. For example, if $\epsilon_v = 0.1$, we require

$\bar{v}_\phi \simeq 0.26 v_c$. For this potential, we can easily show from Eq. (2.141) that $\sigma \simeq (v_c/1.4)$ so that we need $\bar{v}_\phi \simeq 0.36\sigma$ for rotational support. Giant ellipticals do not have such large rotations whereas dwarf spheroidals do. This indicates that the assumption of isotropic velocity dispersion is not contradicted by this analysis for low-luminosity ellipticals whereas giant ellipticals must have anisotropic velocity dispersion.

These results can be reinforced by an analysis based on virial theorem derived in Vol. I, Chap. 8, Section 8.2. There it was shown that the following result holds in general:

$$\frac{1}{2}\frac{d^2 I_{jk}}{dt^2} = 2T_{jk} + \Pi_{jk} + W_{jk}, \tag{2.146}$$

with

$$I_{jk} \equiv \int \rho x_j x_k \, d^3\mathbf{x}, \qquad T_{jk} \equiv \frac{1}{2}\int \rho \langle v_j \rangle \langle v_k \rangle \, d^3\mathbf{x}, \qquad \Pi_{jk} \equiv \int \rho \sigma_{jk}^2 \, d^3\mathbf{x}, \tag{2.147}$$

$$W_{jk} \equiv -\int \rho(\mathbf{x}) x_j \frac{\partial \Phi}{\partial x_k} d^3\mathbf{x} = -\frac{1}{2} G \int\int \rho(\mathbf{x})\rho(\mathbf{x}') \frac{(x_j' - x_j)(x_k' - x_k)}{|\mathbf{x}' - \mathbf{x}|^3} d^3\mathbf{x}\, d^3\mathbf{x}'. \tag{2.148}$$

Let us apply this relation to an axisymmetric system like an elliptical galaxy that rotates about the symmetry axis (z axis) and is observed edge-on. Taking the line of sight to the centre of the system as the x axis, it follows from symmetry that

$$W_{xx} = W_{yy}, \qquad W_{ij} = 0, \quad (i \neq j), \tag{2.149}$$

and similarly for Π_{ij} and T_{ij}. Hence the only two nontrivial virial equations in steady state are

$$2T_{xx} + \Pi_{xx} + W_{xx} = 0, \qquad 2T_{zz} + \Pi_{zz} + W_{zz} = 0. \tag{2.150}$$

Taking the ratio of these two, we get

$$\frac{2T_{xx} + \Pi_{xx}}{2T_{zz} + \Pi_{zz}} = \frac{W_{xx}}{W_{zz}}. \tag{2.151}$$

To proceed further, we need to make a series of physically motivated assumptions regarding the various terms. If the only streaming motion is rotation about the z axis, then $T_{zz} = 0$ and $2T_{xx} = (1/2)Mv_0^2$, where M is the mass of the system and v_0^2 is the mass-weighted mean-square speed of rotation. Similarly, $\Pi_{xx} = M\sigma_0^2$, where σ_0^2 is the mass-weighted mean-square random velocity along the line of sight. Further, we relate Π_{zz} to Π_{xx} by introducing a parameter δ that measures the aniostropy of the galaxy's velocity dispersion. We take

$$\Pi_{zz} = (1 - \delta)\,\Pi_{xx} = (1 - \delta)\, M\sigma_0^2. \tag{2.152}$$

Finally, we note that if the isodensity curves are similar concentric ellipsoids, then the ratios like (W_{xx}/W_{zz}) can depend on only the ellipticity ϵ of the surfaces. This result arises from the study of potentials generated by ellipsoidal figures and can be stated in the following form.[3] If the isodensity surfaces are triaxial ellipsoids on which the coordinates (x_1, x_2, x_3) are constrained to satisfy the condition that the combination

$$m^2 \equiv a_1^2 \sum_{i=1}^{3} \frac{x_i^2}{a_i^2} \tag{2.153}$$

is a constant, then we can take the density distribution to be a function of m^2 alone; that is, $\rho(\mathbf{x}) = \rho(m^2)$. It is also convenient to define an auxillary function

$$\psi(m) \equiv \int_0^{m^2} \rho(m^2)\, dm^2 \tag{2.154}$$

in terms of which the potential-energy tensor is given by

$$W_{jk} = -\pi^2 G \frac{a_2 a_3}{a_1^2} \left(\frac{a_j}{a_1}\right)^2 A_j \delta_{jk} \int_0^{\infty} [\psi(\infty) - \psi(m)]^2\, dm. \tag{2.155}$$

The quantities A_j in this expression are functions of axial ratios like a_2/a_1, etc., and the integral is independent of a body's ellipticity and is the same for all components of W_{ij}. Hence, when we take a ratio between any two components of W_{ij}, the result can depend on only the ellipticity of the body. Putting all these together, we get the relation

$$\frac{v_0^2}{\sigma_0^2} = 2(1-\delta)\frac{W_{xx}}{W_{zz}} - 2 \equiv 2(1-\delta) f(\epsilon) - 2, \tag{2.156}$$

where $f(\epsilon) \equiv (W_{xx}/W_{zz})$ can be computed if the model for the galaxy is known. This equation shows that the quantity (v_0/σ_0) depends on only ϵ and δ. It is now possible to plot curves of constant δ in the $[v_0/\sigma_0, \epsilon]$ plane. Given the observed values of these quantities, we can infer the best possible value for σ for a given sample of galaxies. Such an analysis is complicated by the fact that we need to account for the effect of inclination in the sky plane, which is not easy to do for ellipticals. Nevertheless, studies show that most of the low-luminosity ellipticals are clustered around the $\delta = 0$ curve whereas the high-luminosity massive ellipticals do not show any systematic behaviour. This is an indication that giant ellipticals probably have significant and (random) anisotropy in the velocity dispersions whereas low-mass ellipticals are isotropic in the velocity space and are rotationally supported.

For the range of ellipticities encountered in the galaxies, $(W_{xx}/W_{zz}) \equiv f(\epsilon) \approx (a/b)^{0.89}$. If we assume that all the flattening is due to rotation and velocity

anisotropy is negligible ($\delta \approx 0$), Eq. (2.156) gives

$$\frac{v_0}{\sigma_0} = \sqrt{2}\left[\left(\frac{a}{b}\right)^{0.89} - 1\right]^{1/2}. \tag{2.157}$$

For $(b/a) = 0.7$, this requires $(v_0/\sigma_0) \approx 0.8$. In the other extreme case, in which rotation is negligible and the flattening is entirely due to anisotropy, we can set $T_{xx} = T_{yy} = T_{zz} = 0$ in Eq. (2.151) and take $\Pi_{xx} = M\sigma_x^2$ and $\Pi_{zz} = M\sigma_z^2$. In this case, we need $\sigma_z/\sigma_x \approx (b/a)^{0.45} \approx 0.87$ if $(b/a) = 0.7$.

2.5 Stellar Dynamics at Galactic Cores

The dynamical issues discussed so far have dealt with general properties of the galaxies. However, in most galaxies, the central region has special characteristics and is of considerable importance. We have seen in Section 1.7 that the central region of most galaxies might contain a massive black hole. The question arises as to how the black hole will influence the stellar dynamics in the galactic core.

Direct observational evidence for black holes in the centres of galaxies comes from different kinds of observations in several systems. For example, Very Long Baseline Interferometer (VLBI) observations of megamasers in the nucleus of Seyfert galaxy NGC 4258 show individual knots of material revolving around a central source with near-Keplerian velocities. The mass of the compact central object seems to be approximately $3.6 \times 10^7\ M_\odot$, confined within an inner region of 13 pc. Similar conclusions are reached from the velocity field of matter close to the centre, measured through the optical emission lines. In M87, for example, HST observations of the [OII] λ 3727 line at three different positions separated by 0.2″ (with a resolution of 0.03″) allow the measurement of rotation curve up to 5 pc near the centre. The analysis of this data shows that the observed motion is consistent with the presence of a supermassive black hole with $M \simeq (3.2 \pm 0.9) \times 10^9 M_\odot$. High-spatial-resolution HST studies of the central regions of other galaxies have led to similar conclusions. In NGC 6251, a giant E2 galaxy, there exists an extended nuclear disk, the dynamics of which is governed by a central compact object. The mass estimates vary in the range (4–8) $\times 10^8\ M_\odot$, depending on the model used for stellar mass density. Other galaxies, such as NGC 4261, NGC 4374, and NGC 7052, all indicate the existence of central black holes with masses in the range (2–6) $\times 10^8 M_\odot$. Finally, the central masses of AGN can also be measured by reverberation-mapping techniques (see Chap. 8). The central masses estimated from this procedure can be compared with those obtained from the standard photoionisation model of the BLR; they agree within the uncertainties of the measurements.

There are several evolutionary sequences in the core of the galaxy that can lead to the formation of a central black hole. In many of them the gravitational core collapse and dynamical friction (see Vol. I, Chap. 10, Section 10.7) play

an important role. As an example of how such collisional dynamics operates, let us consider the nuclear region of the spiral galaxy M31 that is at a distance of $D = 700$ kpc. At a projected radius of 1 arcsec, stars in the nuclear region have a line-of-sight velocity dispersion of $\sigma = 150\ \text{km s}^{-1}$ and are also rotating about the nucleus at a similar speed of $v_\phi = 150\ \text{km s}^{-1}$. The analysis based on Eq. (2.156) shows that $v_{\text{rot}} = \sigma$ will imply a specific value for eccentricity ϵ. Numerically, we get $\epsilon \approx 0.5$ for $\delta = 0$, implying that $(b/a) = \sqrt{1 - \epsilon^2} = 0.85$, which is fairly spherical. Jeans equation (2.134) now reduces to

$$\frac{\partial \phi}{\partial r} = \frac{v_\phi^2}{r} - \frac{1}{\nu}\frac{\partial(\nu\sigma^2)}{\partial r}. \tag{2.158}$$

Ignoring the variation of σ^2 with r and taking the density profile in the nuclear region to fall as r^{-2} [so that $(\partial \ln \nu / \partial \ln r) = -2$], we find that this reduces to

$$\frac{\partial \phi}{\partial r} = \frac{v_\phi^2}{r} + \frac{2\sigma^2}{r}. \tag{2.159}$$

A rough estimate of mass inside a radius r is given by $M(r) \approx (r^2/G)(\partial\phi/\partial r)$. In this case, we get, using the observed values for v_ϕ and σ,

$$M \simeq \frac{r}{G}\left(v_\phi^2 + 2\sigma^2\right) \simeq 5 \times 10^7 M_\odot. \tag{2.160}$$

The luminosity in the core region of M31 is approximately $3 \times 10^6\ L_\odot$ so that the mass-to-light ratio is ~ 17. The corresponding density profile can now be taken as

$$\rho(r) = 10^5 (M_\odot\ \text{pc}^{-3}) \left(\frac{3.4\ \text{pc}}{r}\right)^2. \tag{2.161}$$

The gravitational two-body relaxation time in the core region is given by (see Vol. I, Chap. 10, Section 10.7)

$$t_R = 0.34 \frac{\sigma^3}{G^2 M \rho \ln \Lambda}, \tag{2.162}$$

where $\ln \Lambda \simeq \ln N \approx 14$, $\sigma \simeq 150\ \text{km s}^{-1}$, and $M = 1\ M_\odot$. This gives

$$t_R \simeq 10^{10} \left(\frac{4 \times 10^5\ M_\odot\ \text{pc}^{-3}}{\rho}\right) \text{yr}. \tag{2.163}$$

If the age of galaxy is taken to be 10×10^9 years, then the two-body relaxation time is of the order of the age of the galaxy at $r \lesssim 1.7$ pc. The discussion in Vol. I, Chap. 10, Section 10.7, shows that such a system will undergo core collapse in a time scale that is typically 10^2 times the relaxation time. Because the relaxation time scales as $\rho^{-1} \propto r^2$, we would expect stars in a region smaller than $(1.7\ \text{pc}/\sqrt{10^2}) = 0.17$ pc to have undergone core collapse.

Another characteristic time scale at the core is that for a massive object with mass M – say a 10 $M_\odot$ blackhole – to sink to the centre. This time scale is given by (see Vol. I, Chap. 10, Section 10.7)

$$t_{\rm sink} \simeq \frac{1}{3\ln\Lambda}\frac{M(<r)}{M}t_{\rm orb} \simeq \frac{1}{3\ln\Lambda}\frac{v^2(r)r}{GM}\frac{2\pi r}{v} \simeq \frac{2\pi}{3\ln\Lambda}\frac{r^2 v}{GM}; \quad (2.164)$$

that is,

$$r = \left(\frac{3\ln\Lambda}{2\pi}\right)^{1/2}\sqrt{\frac{GMt_{\rm sink}}{v}}. \quad (2.165)$$

For $t_{\rm sink} \simeq 10^{10}$ yr, $v(r) = \sqrt{2}\sigma = 210$ km s^{-1}, $\ln\Lambda \approx 15$, and $M = 10\ M_\odot$, this gives a radius of ~4.6 pc. These considerations show that the central regions of the nuclei are indeed dominated by gravitational dynamics.

It is interesting to ask what will be the effect of such a central black hole on a stellar system that interacts with it. The effect will depend on two parameters. The first is the ratio between the mass of the black hole, $M_{\rm BH}$, and the amount of stellar mass, M_*, contained in the central region ($M_* \simeq \rho_0 r_0^3$, where r_0 is of the order of the core radius, say). The second is the ratio between the age of the system and the two-body relaxation time in the central region. Of these, it is usually safe to assume that $M_{\rm BH} \ll M_*$. As regards the time scales, it is necessary to consider the two cases separately.

Let us begin with the situation in which $t \ll t_R$. If the black hole forms by slow accumulation of mass over a time scale that is much larger than the orbital time scale (R_0/σ) of the stars, then the evolution of the distribution function can be estimated in the adiabatic approximation. We have seen in Vol. I, Chap. 2, Section 2.6 that – in the adiabatic limit – the actions J_i remain constant. Hence the distribution function can be thought of as a given function of $\mathbf{J}(E, L)$, where $\mathbf{J}$ denotes the actions as functions of energy and angular momentum. If the initial distribution function is further assumed to be a Maxwellian, then we have

$$f(\mathbf{J}) = \frac{\nu_0}{(2\pi\sigma^2)^{3/2}}e^{-E_i(\mathbf{J})/\sigma^2}, \quad (2.166)$$

where $E_i(\mathbf{J})$ is the energy as a function of the actions in the initial potential. As an example of the application of this result, consider the density contributed by the stars that eventually become bound to the black hole. For these stars, $E_i \ll \sigma^2$, and we can take $f(J) \approx \nu_0/(2\pi\sigma^2)^{3/2}$. Then the density becomes

$$\nu(r) = 4\pi\int_0^{\sqrt{2GM_{\rm BH}/r}} v^2 f(E, L)\,dv = \frac{4\nu_0}{3\sqrt{\pi}}\left(\frac{r_H}{r}\right)^{3/2}, \qquad r_H \equiv \frac{GM_{\rm BH}}{\sigma^2}, \quad (2.167)$$

where the parameter r_H represents the radius of influence of the hole. Note that the black hole has introduced a weak cusp ($\nu \propto r^{-3/2}$) in the density profile.

The situation in the case of $t \gg t_R$ is more complicated. A naive application of Maxwellian distribution in the energy will give a density of $\nu \propto \exp[-\Phi(r)/\sigma^2] \propto \exp(r_H/r)$, which diverges exponentially near the origin. This result, of course, is unrealistic because any star having a close encounter with the black hole will be tidally shredded and swallowed by the black hole. If these processes lead to a steady-state density distribution $\nu(r) = \nu_0(r_H/r)^s$, then equilibrium would require an energy balance between the potential energy of the stars that are swallowed by the hole and the flux of positive energy that must flow out through the cusp. Because $v^2 \propto r^{-1}$, the relaxation time will vary as $t_R \propto (v^3/\nu) \propto r^{(s-3/2)}$. The number of stars $N(r)$ within a radius r scales as $N(r) \propto r^{3-s}$ whereas the energy of a star at a given radius r scales as $E(r) \propto r^{-1}$. Hence the rate of flow of energy through a radius r scales as $[N(r)E(r)/t_R] \propto r^{(7/2-2s)}$. If this flow has to be independent of radius r, then we must have $s = 7/4$, which is a slightly milder cusp compared with the one previously obtained. Numerical simulations show that this result is indeed true.

Finally, let us consider the density of stars that are *not* bound to the black hole. Far away from the black hole, we may assume that the test particles have a Maxwellian velocity distribution

$$f_0(v) = \frac{\nu_0}{(2\pi\sigma^2)^{3/2}} e^{-v^2/2\sigma^2} \tag{2.168}$$

with a velocity dispersion σ^2. For unbound particles the distribution function is of the form $f(E) \propto \exp(-E/\sigma^2)$, with $E > 0$. The number of particles that are not bound at any given radius r is given by

$$\nu(r) = \int_{v_{\rm esc}(r)}^{\infty} 4\pi v^2 \, dv \, f(E), \tag{2.169}$$

where $v_{\rm esc}(r) = (2GM_{\rm BH}/r)^{1/2}$ is the escape velocity at the location r. It is convenient to again define a scale length $r_H = (GM_{\rm BH}/\sigma^2)$, in terms of which the integral is easy to evaluate, and we get

$$\nu(r) = \nu_0 \left\{ 2\sqrt{\frac{r_H}{\pi r}} + e^{r_H/r} \left[1 - \mathrm{erf}\left(\sqrt{\frac{r_H}{\pi r}} \right) \right] \right\}, \tag{2.170}$$

where

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\, dt \tag{2.171}$$

is the error function. Using the behaviour of $\mathrm{erf}(x)$ near the origin, we easily see that the second term in the braces in Eq. (2.170) also behaves as $r^{-1/2}$ near the origin so that $\nu(r) \propto r^{-1/2}$ in the vicinity of the black hole.

2.6 Spiral Structure

We next consider the spiral pattern seen in the disk galaxies, which has attracted a considerable amount of theoretical attention. Visual appearance may suggest that such a pattern arises because the material making up the spiral arm winds up because of the differential rotation of the galaxy. As we shall subsequently see, this cannot be a possible explanation as this will lead to a very tightly wound spiral pattern in many cases. A more likely explanation is that a spiral structure is a quasi-stationary density wave in which a particular wave pattern stays in place even though the actual material forming the wave pattern does not reside in the location of the spiral. Such an idea is related to the stability of the differentially rotating disk, and we will see that the problem is closely connected with the resonances described earlier in Subsection 2.3.5.

Let us first show that the actual winding up of the material that is due to differential rotation, if that were the cause for the spiral pattern, is counterproductive. To see this, we consider the equation describing the location of a set of stars that were along a straight line $\phi = \phi_0$ at $t = 0$. At any given later time, the location of this strip of stars is given by $\phi(R, t) = \phi_0 + \Omega(R)t$ if they rotate differentially. The pitch angle i of the arm at any given radius is defined as the angle between the tangent to the arm and a circle of constant radius (see Fig. 2.8) so that $\cot i \equiv |R(\partial\phi/\partial R)|$. For our case, this gives $\cot i = Rt|(d\Omega/dR)|$. If ΔR is the separation between adjacent arms at a given angle ϕ, then we must have

$$2\pi = |\Omega(R + \Delta R) - \Omega(R)|t \tag{2.172}$$

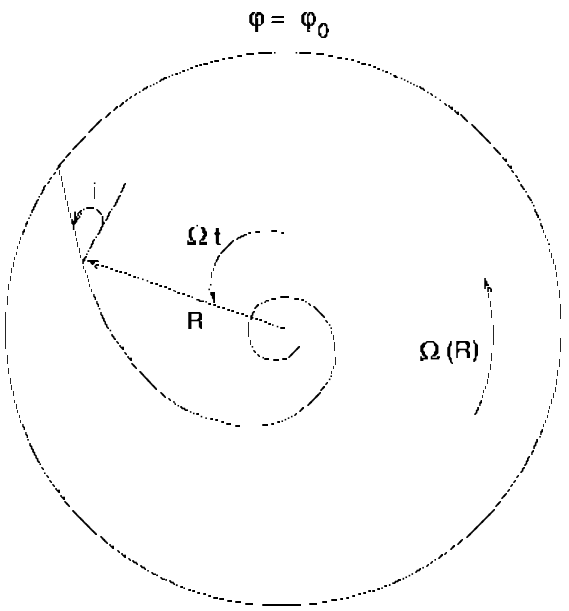


Fig. 2.8. The winding problem and the pitch angle.

so that, for $\Delta R \ll R$,

$$\Delta R = \frac{2\pi R}{\cot i}. \tag{2.173}$$

For a typical galaxy, with $\Omega R \simeq 220\,\mathrm{km\,s^{-1}}$, $R \simeq 10\,\mathrm{kpc}$, and $t \simeq 10^{10}\,\mathrm{yr}$, we get $i = 0.25^\circ$ and $\Delta R \approx 0.28\,\mathrm{kpc}$. The actual observed separations are considerably larger; in other words, the spiral is not as tightly wound as suggested by this analysis.

It is possible, however, to come up with a density enhancement of stars along a spiral without leading to overwinding. To understand this idea in its simplest context, let us consider the orbit of stars in the epicycle approximation described in Section 2.3. If we concentrate on a specific guiding centre orbit at some radius R_0, it is easy to see that, in general, the orbit will not be closed in the inertial frame because the angular velocity $\Omega(R_0)$ and the epicyclic frequency $\kappa(R_0)$ will not be commensurate. On the other hand, if we view the orbit in a suitably chosen rotating frame (rotating with an angular pattern speed Ω_p), then the orbit can be closed provided that $\Omega' = (\Omega - \Omega_p)$ and κ are commensurate. We now note that, for a reasonable class of density profiles, the quantity $[\Omega(R) - \kappa(R)/2]$ is fairly constant at all R (see, e.g., Fig. 2.7). This suggests that if we take the pattern speed to be $\Omega_p = (\Omega - \kappa/2)$ so that $\Omega' = \kappa/2$, all the orbits will close. If we now arrange the closed orbits (which will in general be ellipses) with the position angle of the major axis increasing systematically, then stars will be concentrated along a spiral pattern (see Fig. 2.9). What is more, this pattern will persist in time as long as the mutual interaction of the stars is ignored. Note that stars are not winding up in this picture and the enhancement along the spiral is, in

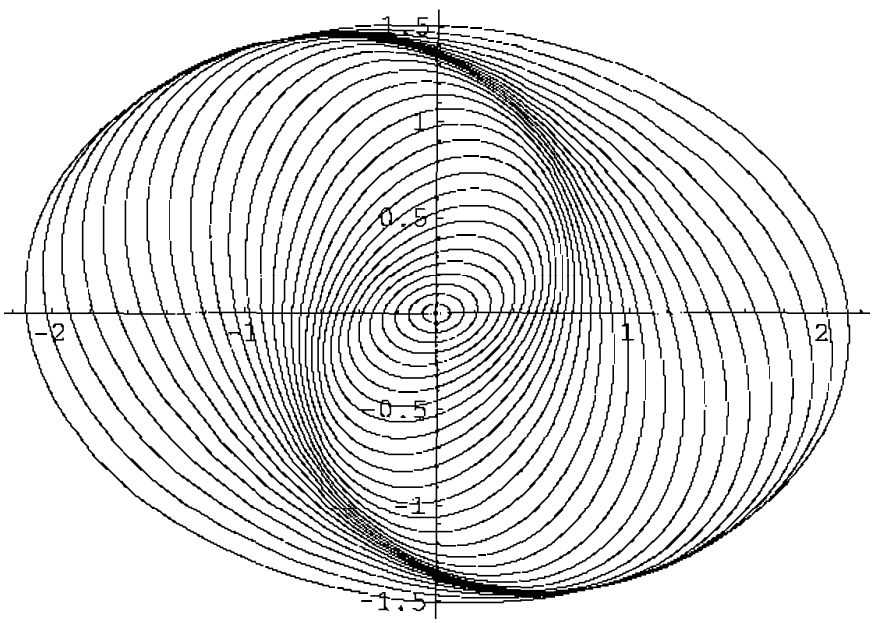


Fig. 2.9. Kinematic spiral wave generated by nested ellipses.

some sense, kinematical. In the nonrotating, inertial frame, the spiral pattern will rotate at the pattern speed.

A more sophisticated and realistic approach to a spiral structure treats it as a density wave related to an instability in the rotating disk. To be rigorous, such an analysis needs to be carried out by treating the contents of the disk as a collisionless system of stars. It turns out, however, that somewhat similar results are obtained even if the material in the disk is treated as gaseous. Because the latter analysis is mathematically simpler, we shall describe it here and briefly mention the corresponding results for a stellar system.[3]

This analysis has three ingredients. First, we assume that there exists a surface-density wave capable of describing a spiral and we determine its gravitational potential by using Poisson's equation. Next, we determine how the motion of stars is affected by this gravitational potential. Finally, we match the modified orbital pattern of the stars to the initially assumed spiral-density wave, thereby obtaining a self-consistent model. The key assumption we shall use is that the response of the gravitational field to the perturbed density field is local and that the long-range effects are negligible. This approximation is reasonable, provided the spiral is not very loosely wound, and we need to confirm that this is indeed the case, at the end of the calculation, for consistency.

We begin by taking the surface density of a (zero-thickness) disk as the sum of two terms: $\Sigma_0(R)$ represents the unperturbed surface density and $\Sigma_1(R, \phi, t)$ represents the perturbation, where

$$\Sigma_1(R, \phi, t) = H(R, t)e^{i[m\phi + f(R,t)]}. \tag{2.174}$$

In this expression, $H(R, t)$ is a smooth function of radius that gives the amplitude of the spiral pattern. The location of m spiral arms ($m = 2, 3, \ldots,$) is defined by the relation

$$m\phi + f(R, t) = \text{constant}. \tag{2.175}$$

If we introduce a radial wave number by

$$k(R, t) \equiv \frac{\partial f(R, t)}{\partial R}, \tag{2.176}$$

then the sign of k determines whether the arms are leading or trailing. If the galaxy rotates in the direction of increasing ϕ, then $k < 0$ implies leading arms, whereas $k > 0$ implies trailing arms. The pitch angle is now given by the relation $\cot i = |kR/m|$ and the condition for tight a spiral is that $\cot i \gg 1$. Observations show that the average pitch angle varies with the galaxy type and the median pitch angle is $\langle i \rangle \simeq 13^\circ$, giving $\langle \cot i \rangle \approx 4$. This shows that the tight-coupling approximation is satisfied only marginally and hence the agreement of the results derived below to actual observed features is somewhat fortuitous.

We now need to determine the gravitational potential that is due to this surface-density distribution. In general this can be obtained only as an integral involving a

two-dimensional Green's function to the Poisson equation. However, if the tight-winding approximation is true, then the sinusoidal oscillations of the surface density lead to a near-complete cancellation of the contribution from distant parts of the pattern to the local potential and the perturbed potential at a given location is essentially determined by the local density. In this case, we expand $f(R, t)$ in a Taylor series around R_0 to get

$$\Sigma_1(R, \phi, t) \simeq \Sigma_a e^{ik(R-R_0)}, \qquad \Sigma_a = H(R_0, t)e^{i[m\phi_0 + f(R_0,t)]}, \qquad k = k(R_0, t). \tag{2.177}$$

Because the radial dependence is also sinusoidal, the gravitational potential is trivial to determine in Fourier space. We get

$$\Phi_1(R, \phi, t) \simeq \Phi_a e^{ik(R-R_0)}, \qquad \Phi_a = -\frac{2\pi G \Sigma_a}{|k|}. \tag{2.178}$$

We are now free to set $R_0 = R$, thereby obtaining the gravitational potential at any location as

$$\Phi_1(R, \phi, t) = -\frac{2\pi G}{|k|} H(R, t)e^{i[m\phi + f(R,t)]} + \mathcal{O}(|kR|^{-1}). \tag{2.179}$$

We can obtain the corresponding inverse relation by differentiating relation (2.179) with respect to R and ignoring the slow variation of H:

$$\Sigma_1(R, \phi, t) = \frac{i\ \mathrm{sign}(k)}{2\pi G} \frac{d}{dR} \Phi_1(R, \phi, t) + \mathcal{O}(|kR|^{-1}). \tag{2.180}$$

We next determine the response of the matter in the disk to this perturbed gravitational potential. We will assume that the matter is confined to the $z = 0$ plane and that the pressure p that acts on the disk is related to the density by an equation of state $p = K\Sigma^\gamma$. The corresponding specific enthalpy is given by $h = [\gamma/(\gamma - 1)]K\Sigma^{\gamma-1}$ (see Vol. I, Chap. 5, Section 5.6). The Euler equations, in component form, are given by

$$\frac{\partial v_R}{\partial t} + v_R \frac{\partial v_R}{\partial R} + \frac{v_\phi}{R} \frac{\partial v_R}{\partial \phi} - \frac{v_\phi^2}{R} = -\frac{\partial \Phi}{\partial R} - \frac{1}{\Sigma} \frac{\partial p}{\partial R}, \tag{2.181}$$

$$\frac{\partial v_\phi}{\partial t} + v_R \frac{\partial v_\phi}{\partial R} + \frac{v_\phi}{R} \frac{\partial v_\phi}{\partial \phi} - \frac{v_\phi v_R}{R} = -\frac{1}{R} \frac{\partial \Phi}{\partial \phi} - \frac{1}{\Sigma R} \frac{\partial p}{\partial \phi}. \tag{2.182}$$

We can simplify the right-hand sides of these equations by using the specific enthalpy as we have

$$-\frac{\partial \Phi}{\partial R} - \frac{1}{\Sigma} \frac{\partial p}{\partial R} = -\frac{\partial \Phi}{\partial R} - \gamma K \Sigma^{\gamma-2} \frac{\partial \Sigma}{\partial R} = -\frac{\partial(\Phi + h)}{\partial R}. \tag{2.183}$$

We now solve Eqs. (2.181) and (2.182) in linear perturbation theory by taking the variables to be $v_R = v_{R1}$, $v_\phi = v_{\phi 0} + v_{\phi 1}$, $h = h_0 + h_1$, and $\Phi = \Phi_0 + \Phi_1$, where the subscript '1' denotes the first-order perturbations. In the lowest order,

we assume the orbits to be circular with

$$\frac{v_{\phi 0}^2}{R} = \frac{d}{dR}(\Phi_0 + h_0) \simeq \frac{d\Phi_0}{dR} \tag{2.184}$$

where the second relation is valid because the interstellar sound speed ($v_s \approx 10\ \mathrm{km\ s^{-1}}$) is negligible compared with the circular speed ($v_c \approx 220\ \mathrm{km\ s^{-1}}$); the derivative of h_0 is smaller than the term retained by the factor $(v_s/v_c)^2$. The lowest-order perturbed equations now become

$$\frac{\partial v_{R1}}{\partial t} + \Omega\frac{\partial v_{R1}}{\partial \phi} - 2\Omega v_{\phi 1} = -\frac{\partial}{\partial R}(\Phi_1 + h_1), \tag{2.185}$$

$$\frac{\partial v_{\phi 1}}{\partial t} + \left[\frac{d(\Omega R)}{dR} + \Omega\right] v_{R1} + \Omega\frac{\partial v_{\phi 1}}{\partial \phi} = -\frac{1}{R}\frac{\partial}{\partial \phi}(\Phi_1 + h_1). \tag{2.186}$$

We can solve these equations by assuming that all the perturbed variables have dependences of the form $q_{a1}(R, \phi, t) = q_a(R)\exp[i(m\phi - \omega t)]$. Making these substitutions, we can solve for the velocity and enthalpy perturbations in terms of other quantities and obtain

$$v_{Ra} = -\frac{i}{\Delta}\left[(m\Omega - \omega)\frac{d}{dR}(\Phi_a + h_a) + \frac{2m\Omega}{R}(\Phi_a + h_a)\right], \tag{2.187}$$

$$v_{\phi a} = \frac{1}{\Delta}\left[-2B\frac{d}{dR}(\Phi_a + h_a) + \frac{m(m\Omega - \omega)}{R}(\Phi_a + h_a)\right], \tag{2.188}$$

$$h_a = \gamma K \Sigma_0^{\gamma-2}\Sigma_a \equiv \frac{v_s^2\Sigma_a}{\Sigma_0}, \tag{2.189}$$

$$\Delta \equiv \kappa^2 - (m\Omega - \omega)^2, \tag{2.190}$$

with

$$\kappa^2(R) = R\frac{d\Omega^2}{dR} + 4\Omega^2 = -4B\Omega,$$

$$B(R) = -\frac{1}{2}\left[\frac{d(\Omega R)}{dR} + \Omega\right] = -\Omega - \frac{R}{2}\frac{d\Omega}{dR}. \tag{2.191}$$

So far we have considered only the Euler equations that govern the velocity response of the matter to the perturbed potential. Velocity perturbation induces a density perturbation through the continuity equation, which – to the lowest order – becomes

$$\frac{\partial \Sigma_1}{\partial t} + \frac{1}{R}\frac{\partial}{\partial R}(R\Sigma_0 v_{R1}) + \Omega\frac{\partial \Sigma_1}{\partial \phi} + \frac{\Sigma_0}{R}\frac{\partial v_{\phi 1}}{\partial \phi} = 0 \tag{2.192}$$

or, equivalently,

$$i(m\Omega - \omega)\Sigma_a + \frac{1}{R}\frac{d}{dR}(R\Sigma_0 v_{Ra}) + \frac{im\Sigma_0}{R}v_{\phi a} = 0. \tag{2.193}$$

Equations (2.187), (2.188), (2.189), and (2.193) provide four equations for the five variables (Σ_a, v_{Ra}, $v_{\phi a}$, h_a, Φ_a). The last relation needed to close the system is the Poisson equation connecting the potential to the density, which has already been solved in Eq. (2.179). This relation shows that Σ_a and because of Eq. (2.189) – h_a contain exactly the same factor $\exp[if(R)]$ as the potential Φ_a. Because this factor varies rapidly with R, we can drop terms involving the derivative with respect to R in Eqs. (2.187) and (2.188); we also write $d(\Phi_a + h_a)/dR \cong ik(\Phi_a + h_a)$ and $d(R\Sigma_0 v_{Ra})/dR \cong ik(R\Sigma_0 v_{Ra})$ as all these replacements involve only a fractional error of $\mathcal{O}(|kR|^{-1})$. This gives

$$v_{Ra} = \frac{(m\Omega - \omega)k(\Phi_a + h_a)}{\Delta}, \qquad v_{\phi a} = -\frac{2Bik(\Phi_a + h_a)}{\Delta}, \tag{2.194}$$

$$(m\Omega - \omega)\Sigma_a + k\Sigma_0 v_{Ra} = 0. \tag{2.195}$$

Simple algebraic manipulation now leads to the condition

$$\Sigma_a\left(1 + \frac{k^2 v_s^2}{\Delta} - \frac{2\pi G\Sigma_0|k|}{\Delta}\right) = 0, \tag{2.196}$$

which provides the dispersion relation connecting ω and k:

$$(m\Omega - \omega)^2 = \kappa^2 - 2\pi G\Sigma|k| + k^2 v_s^2. \tag{2.197}$$

We now consider the implications of this dispersion relation, beginning with the simple case of $m = 0$. This dispersion relation for $m = 0$ contains, as a further special case, the rigidly rotating fluid disk for which $\kappa = 2\Omega$. In this case, the dispersion relation becomes

$$\omega^2 = 4\Omega^2 - 2\pi G\Sigma_0|k| + k^2 v_s^2. \tag{2.198}$$

The stability requires $\omega^2 \geq 0$, and – if there is no rotation, $\Omega = 0$ – then all the modes with $|k| < k_J \equiv (2\pi G\Sigma_0/v_s^2)$ are unstable. This is the standard Jeans instability (which has been discussed in Vol. I, Chap. 8, Section 8.5) but applied to the two-dimensional case. One important difference between Jeans instability in three-dimensional homogeneous media and a two-dimensional disk is the following. In the three-dimensional case, in the absence of pressure support ($v_s = 0$) and rotation ($\Omega = 0$), the instability grows as $\exp(\gamma t)$, where $\gamma^2 = 4\pi G\rho_0$, which is independent of the wavelength of the perturbation. In a two-dimensional disk, without pressure support and rotation, the corresponding growth rate is $\gamma^2 = 2\pi G\Sigma_0|k|$, which increases for small wavelengths.

This instability cannot be suppressed by rotation. In fact, when the sound speed is zero, all the modes with $|k| > (2\Omega^2/\pi G\Sigma_0)$ are unstable with a diverging growth rate at short wavelengths. It is interesting to observe that neither rotation nor pressure, by itself, can stabilise a sheet; a rotating sheet without

pressure support is unstable at small wavelengths whereas a nonrotating sheet with pressure support is unstable at large wavelengths. Interestingly enough, the existence of *both* rotation and pressure can stabilise the sheet. When both effects are present, an examination of quadratic equation (2.198) shows that the sheet is stable at all wavelengths if the following condition is satisfied:

$$\frac{v_s \Omega}{G\Sigma_0} \geq \frac{\pi}{2} \cong 1.57. \tag{2.199}$$

Let us now return to the disk rotating differentially so that $\Omega(R)$ depends on the radius but still keep $m = 0$. The marginal stability determined by $\omega^2 = 0$ leads to the quadratic equation

$$\kappa^2 - 2\pi G\Sigma|k| + k^2 v_s^2 = 0. \tag{2.200}$$

It is easy to see that stability now requires that

$$Q \equiv \frac{v_s \kappa}{\pi G \Sigma} > 1, \tag{2.201}$$

which is similar to condition (2.199) previously obtained. The region of stability can be more conveniently expressed in terms of Q and a critical wavelength defined by $\lambda_{\text{crit}} \equiv (4\pi^2 G\Sigma/\kappa^2)$. The region of marginal stability has the equation $Q = 2[\zeta(1-\zeta)]^{1/2}$, where $\zeta = (\lambda/\lambda_{\text{crit}})$ (see Fig. 2.10). In a solar neighbourhood, $\kappa \approx 32$ km s^{-1} kpc^{-1}, $\Sigma \approx 60\ M_\odot$ pc^{-2}, and $v_s = 30$ km s^{-1}, giving $Q \approx 1.2$. Observationally, therefore, we have no direct evidence for

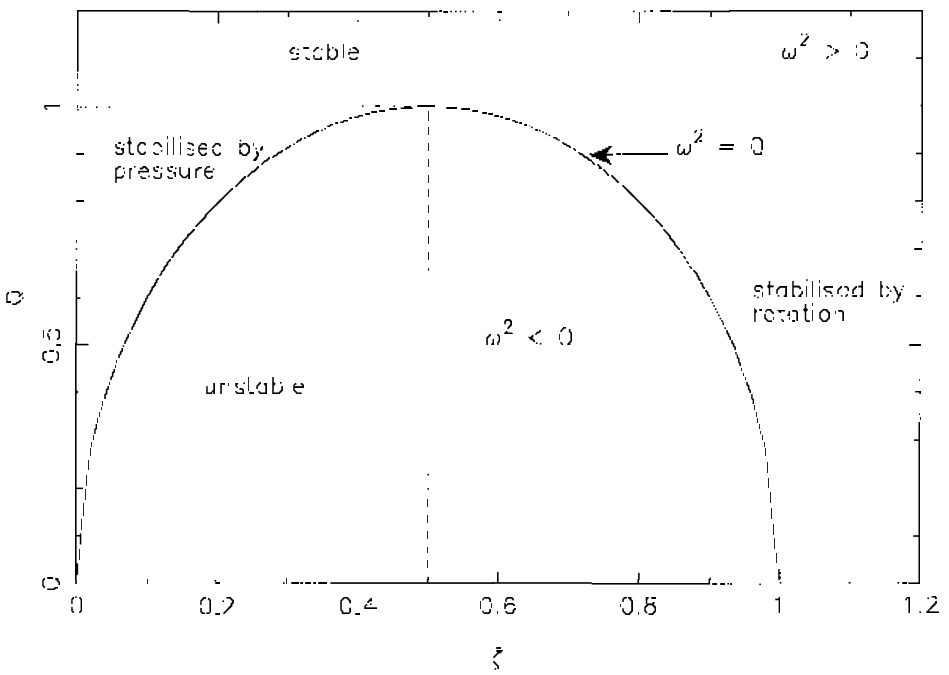


Fig. 2.10. Regions of stability in an axisymmetric differentially rotating gaseous disk.

the occurrence of local instability in the solar neighbourhood, although the uncertainties in the observations do not allow us to firmly rule out values of Q equal to or marginally lower than unity.

To analyse the case with $m \neq 0$, which can lead to spiral patterns with a pattern speed of $\Omega_p \equiv (\omega/m)$, it is convenient to define a dimensionless frequency, $s = [m(\Omega_p - \Omega)/\kappa]$, such that corotation resonance is at $s = 0$ and the Lindblad resonances are at $s = \pm 1$. The actual radii at which these resonances occur, of course, will depend on the functional form of $s(R)$. {For example, in a Mestel disk, $s = (m/\sqrt{2})[(R/R_{CR}) - 1]$, where R_{CR} is the corotation radius and the Lindblad resonances are at $(R/R_{CR}) = 1 \pm (\sqrt{2}/m)$.} The dispersion relation provides a general curve connecting s to (k/k_{crit}) [where $k_{crit} = (2\pi/\lambda_{crit})$] for a given value of Q:

$$\frac{|k|}{k_{crit}} = \frac{2}{Q^2}\left\{1 \pm [1 - Q^2(1 - s^2)]^{1/2}\right\}. \tag{2.202}$$

The forms of this curve for different values of Q are shown in Fig. 2.11. It is obvious that the dispersion relation is the same for leading ($k < 0$) and trailing ($k > 0$) waves. Hence, for a given value of wave frequency ω and disk location R, there can exist four kinds of waves: short trailing [ST; $k > 0$ with a plus sign in Eq. (2.202)]; long trailing [LT; $k > 0$ with a minus sign in Eq. (2.202)]; short leading [SL; $k < 0$ with a plus sign in Eq. (2.202)]; and long leading [LL; $k < 0$ with a minus sign in Eq. (2.202)]. All the long waves become infinitely long at the outer and inner Lindblad resonances corresponding to $s = \pm 1$. At these turning points, the character of the wave changes from sinusoidal to exponentially

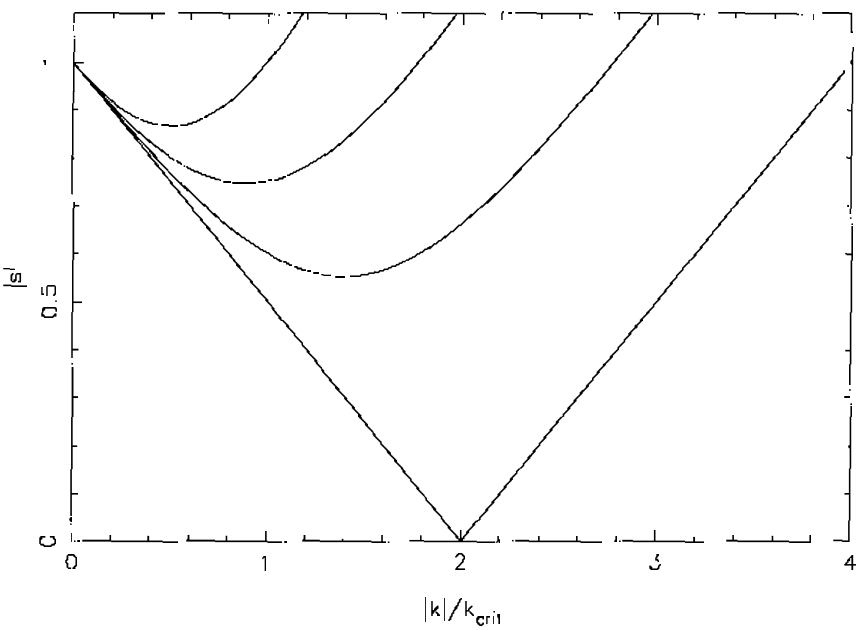


Fig. 2.11. The dispersion relation in Eq. (2.202) is plotted for different values of Q. The Q values (from bottom to top) are 1, 1.2, 1.5, and 2.

damped. Hence in galaxy models in which $\Omega \pm (\kappa/m)$ vary monotonically with R, long waves can propagate only in the region between the Lindblad resonances. When $Q > 1$, there is also a forbidden region around the corotation resonance in which the waves are exponentially damped. This is easy to see from the fact that the term within the square root in Eq. (2.202) vanishes when $s^2 = 1 - Q^{-2}$. For $Q \gg 1$, this condition is satisfied at the Lindblad resonances ($s = \pm 1$); if we lower Q throughout the disk, this point moves towards the corotation resonance ($s = 0$). For $Q \gg 1$, therefore, the long waves are excluded from the entire region between the Lindblad resonances, whereas for $Q \gtrsim 1$ they can propagate everywhere except in a narrow region surrounding the corotation region.

In summary, it may be stated that the following aspects of the dispersion relation are fairly generic: (1) In all disks, with $Q > 1$, there is a forbidden region around the corotation radius, and (2) between the corotation ($s = 0$) and the Lindblad resonances ($s = \pm 1$), there are two branches of the dispersion relation. The long wave begins at $k = 0$, $s = 1$, with $|k|$ increasing as s decreases; near $s = \pm 1$, the dispersion relation is independent of Q and varies as $|k| = k_{\text{crit}}(1 - s^2)$. The short-wave branch has the limit $|k| \to (2\pi G\Sigma/v_s^2)$ as $|s| \to 1$ and $|k|$ decreases as s decreases.

The entire preceding discussion, including the stability analysis, used a gaseous disk rather than the distribution function for stars in the disk. It turns out, however, that the basic results are the same even for a collisionless set of particles. The stability criterion is modified only slightly and reads as

$$Q \equiv \frac{\sigma_R \kappa}{3.36\, G\, \Sigma} > 1 \qquad \text{(for stars)} \tag{2.203}$$

for a stellar-dynamical disk, where σ_R is the velocity dispersion. The dispersion relation near Lindblad resonances is the same for both gaseous and stellar disks whereas the short-wave branch has $|k| \to \infty$ as $|s| \to 1$ in the stellar disk.

Given the dispersion relation for a given mode, we can construct propagating wave packets by adding different wave numbers with the propagation determined by the dispersion relation. For such waves, physical quantities like energy, angular momentum, etc., are propagated at the group velocity given by

$$v_g = \frac{\partial \omega(k, R)}{\partial k} = \text{sign}(k) \frac{|k| v_s^2 - \pi G \Sigma}{\omega - m\Omega}. \tag{2.204}$$

In this equation, $v_s(R)$, $\Sigma(R)$, and $\Omega(R)$ are determined by the properties of the unperturbed disk and k is determined as a function of ω through the dispersion relation. As the wave propagates, k changes, leaving ω fixed. It is obvious that a wave packet localised around R will propagate outwards if $v_g(R) > 0$ and inwards if $v_g(R) < 0$. As an example, let us consider the wave propagation in a Mestel disk with $Q = 1.5 =$ constant. Figure 2.12 shows the dispersion relation for $m = 2$ in the R–k space. At point A, a packet of leading waves on the short

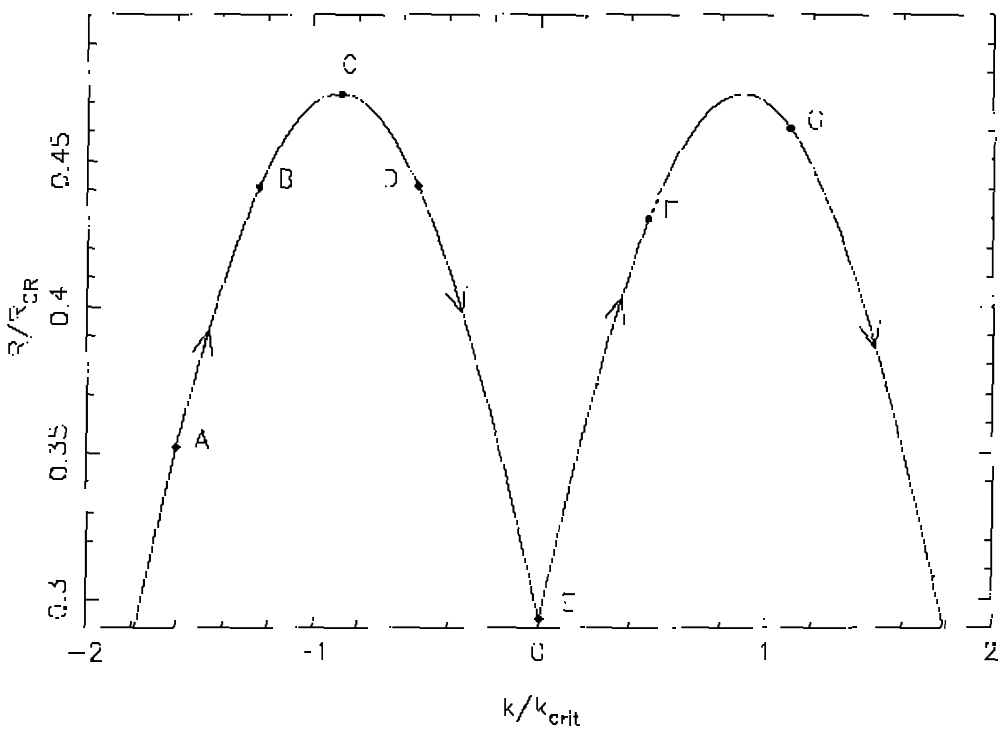


Fig. 2.12. Propagation of a wave packet in a Mestel disk with $Q = 1.5$.

branch has positive group velocity and the wave packet propagates outwards (towards increasing R) with $|k|$ decreasing. Eventually the wave reaches the edge of forbidden region (at point C) where the group velocity changes sign, and it starts propagating inwards as a leading wave on the long branch. This change of direction is completely analogous to the reflection of wave packets by potential barriers in quantum mechanics. Eventually the wave will reach the inner Lindblad resonance (E) at which it reflects off the resonance and propagates out as a trailing wave (F). (The WKB theory is suspect at the turning point; however, a more rigorous analysis confirms this result.) It is reflected at the forbidden zone again and propagates inwards as a ST branch wave.

This entire process causes a tightly wound leading wave to first become a loosely wound leading wave, then a loosely wound trailing wave, and finally a trailing wave that becomes more and more tightly wound. This is reminiscent of the winding up of material arms, and the rate of winding can be worked out with the group velocity and dispersion relation. It is easy to show that the rate of change of pitch angle i of a trailing-wave packet is given by

$$\frac{d}{dt}(\cot i) = \left.\frac{\partial \cot i}{\partial R}\right|_{\omega} v_g, \tag{2.205}$$

which, for a Mestel disk with $\kappa \propto R^{-1}$, $\Omega \propto R^{-1}$, $\Sigma \propto R^{-1}$, and $v_s =$ constant, will give

$$\frac{d}{dt}(\cot i) = \frac{\omega}{m} = \Omega_p. \tag{2.206}$$

In comparison, the material arms will wind up at a rate

$$\frac{d}{dt}(\cot i) = \left| R\frac{d\Omega}{dR} \right| = \Omega, \tag{2.207}$$

as $\Omega \propto R^{-1}$ for the Mestel disk. Clearly waves outside the corotation wind up faster whereas waves inside the corotation wind up slower.

The net result of the preceding analysis therefore is that – at the level of linear perturbation theory based on the WKB approximation – any tightly wound disturbance will wind up fast and disappear into the Lindblad resonance within a few rotation times. Hence linear analysis based on the WKB approximation *alone* cannot sustain a spiral pattern for a very long time. To understand how these patterns are sustained, we need to do a more careful analysis of what happens to the waves near the turning points. Such a study suggests that there are mechanisms that could lead to amplification of the wave, thereby possibly providing a consistent picture. We shall now briefly describe some of the ideas in this direction.

The travelling-wave packets previously described carry energy and angular momentum with a speed equal to the group velocity. Let us consider what happens to a LT wave carrying (-1) unit of angular momentum towards the corotation radius from the inner side. If $Q > 1$, then a ST wave carrying angular momentum ϵ will be transmitted and a wave with angular momentum $-(1+\epsilon)$ will be refracted back. (Similar results hold for energy as well, and the process conserves energy and angular momentum at the reflection point.) Such a result – which is completely analogous to transmission and reflection of wave packets in quantum mechanics – shows that we can achieve a multiplicative factor of $(1+\epsilon)$ in energy and angular momentum. If there is an inner Q barrier associated with, say, a galactic bulge, then the inwardly propagating ST wave will be reflected back into an outwardly propagating LT spiral wave. This wave will now undergo transmission and reflection at the corotation radius, with the transmitted ST wave carrying an angular momentum $\epsilon(1+\epsilon)$ while the ST reflected wave carries an amount $(1+\epsilon)^2$. This can now result in a feedback cycle, and the process can reinforce the original incident wave, provided an eigenvalue condition is satisfied. Thus we can arrive at a growing normal mode that amplifies by a factor $(1+\epsilon)$ per feedback cycle (see Fig. 2.13).

There is another process that can lead to similar effects if the Q barrier rises sufficiently fast compared with the wavelength near the corrotation radius.[4] In that case, a SL wave incident upon the corotation radius from inside can lead not only to the refraction of a LL wave but also to the reflection of a ST wave. The latter (which arises through a nonadiabatic process and cannot be discussed in the WKB approximation) can lead to the transmission of $X(\gg 1)$ units of angular amplitude into the outer region and an amount $-(1+X)$ units inwards in the

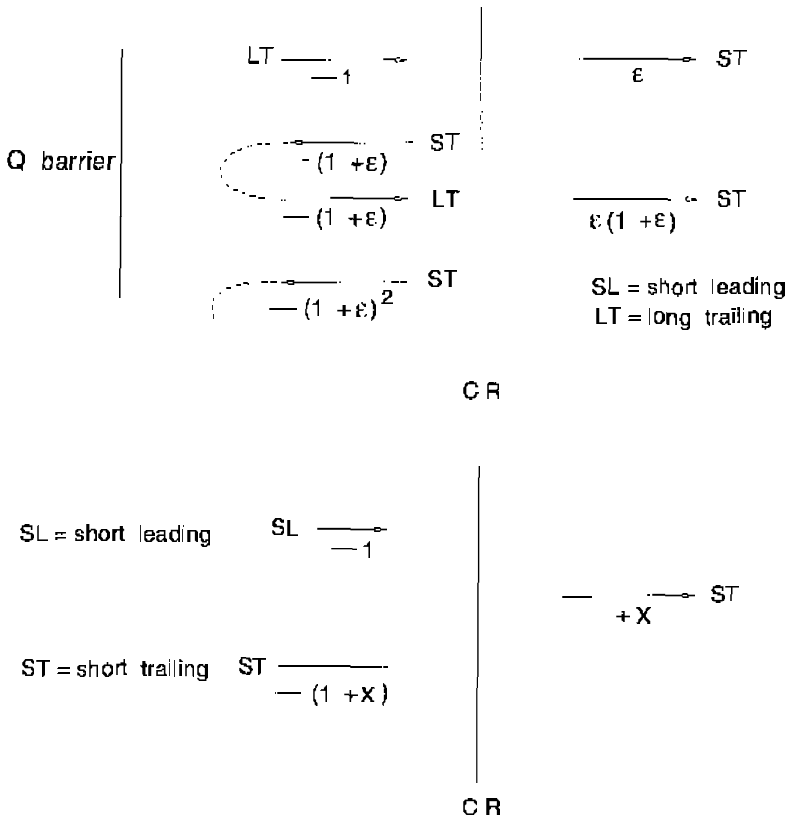


Fig. 2.13. Feedback mechanisms capable of inducing spiral instability.

form of a ST wave. This process, called *swing amplification*, can often dominate the generation of a large-amplitude fluctuation. The physical reason for the swing amplification is a resonance between the epicyclic motion and the rotating spiral feature. To see this, we note that, for a material spiral arm with $\phi(R, t) = \phi_0 + \Omega(R)t$, the winding angle satisfies the relation $\cot i = -Rt(d\Omega/dR) \equiv 2At$. This will give the rate of change of pitch angle as

$$\frac{di}{dt} = \frac{2A}{1 + 4A^2t^2}. \tag{2.208}$$

For a tightly wound arm, (di/dt) is slow, but as the arm swings from leading to trailing branch, it reaches a maximum rotation rate of $2A$. This maximum rate is comparable with the average angular speed κ of stars around their epicycles; for example, $2A = \Omega$ and $\kappa = \sqrt{2}\Omega$ for a Mestel disk. Further, both the unwinding of the arm and the rotation of the stars around their epicycles are in the same sense (opposite to the direction of rotation) leading to a temporary resonance between the epicyclic motion and the rotating spiral feature. This enhances the effect of gravitational force from the spiral on the stellar orbit and could lead to rapid growth of the amplitude representing the spiral arm.

2.7 Warps

Disk galaxies, when observed edge on in the visible band, show a reasonably flat planar structure. The HI observations – which extend to much larger radii than the optical observations – however, often show noticeable warps at the edges. These are most easily seen in the external galaxies that are seen edge-on; for example, almost all the spiral galaxies in the local group are warped.

One possible explanation for the warps is the following. If the disk is embedded in a halo such that the symmetry axes of the disk and the halo are misaligned, then gaseous material in the galactic disk will settle at a configuration in which the net torque from the halo and the disk on each ring of material vanishes. At each radius, there is a unique plane (which in general is different at different radii) on which the net torque vanishes. This makes the gas reach a nonplanar configuration.

The determination of this surface is straightforward if both the disk and the halo are assumed to be axisymmetric and some simple gravitational potential is chosen for the halo. We choose the geometry such that the symmetry axis of the disk is along the z axis and the intersection of the equatorial planes of the disk and the halo is along the y axis. The inclination between the two symmetry axis will be taken to be ν (see Fig. 2.14, left). We take the halo potential to be a simple quadratic in the inclined coordinates given by the form

$$\Phi_h = \frac{v_0^2}{2R_c^2}\left(x'^2 + y^2 + \frac{z'^2}{q^2}\right) + \text{ constant}, \tag{2.209}$$

where q is a free parameter and $x' = x\cos\nu + z\sin\nu$ and $z' = z\cos\nu - x\sin\nu$. We now consider a ring of radius r_0 that has an inclination i and whose symmetry axis (like that of the halo) is in the x-z plane. Expressing Φ_h in terms of x, y, z and taking the gradients, we can calculate the components of the acceleration $\mathbf{g}_h$ that is due to the halo as a function of $\mathbf{x}$. We get

$$(g_{hx}, g_{hy}, g_{hz}) = \frac{v_0^2}{R_h^2}(x, y, z) - \frac{v_0^2(q^{-2}-1)(x\sin\nu - z\cos\nu)}{R_h^2}(\sin\nu, 0, \cos\nu). \tag{2.210}$$

This will lead to a torque on the material at the angular position between ϕ and $d\phi$ in this ring, given by

$$d\mathbf{N}_h = \frac{d\phi}{2\pi}\mathbf{x}(\phi) \times \mathbf{g}_h[\mathbf{x}(\phi)], \tag{2.211}$$

where

$$\mathbf{x}(\phi) = (r\cos\phi\cos i, r\sin\phi, r\cos\phi\sin i). \tag{2.212}$$

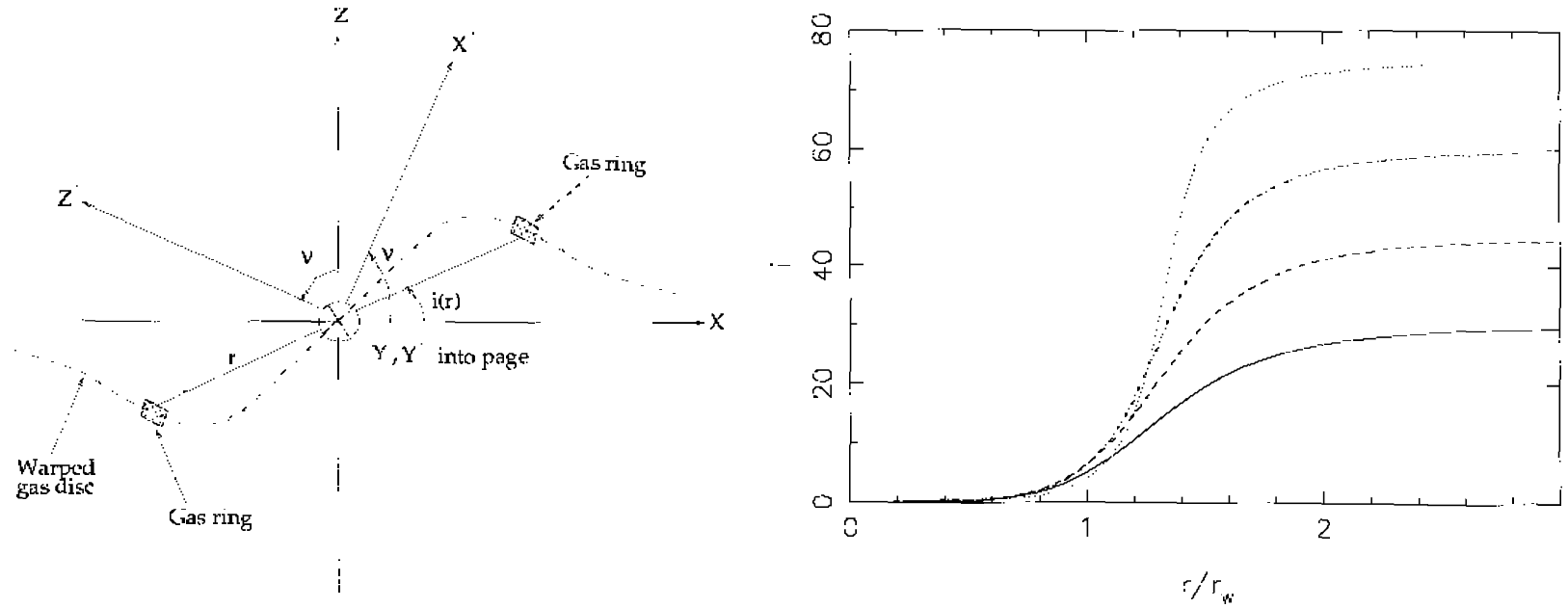


Fig. 2.14. The geometry determining the structure of the warp. The solid, dashed, dotted–dashed, and dotted curves in the right panel are for $\nu = 30°, 45°, 60°, 75°$, respectively.

By symmetry only the y component of the torque will survive. Its total magnitude can be calculated by integration over all ϕ to give

$$N_{\text{halo}} = \frac{v_0^2 r_0^2}{4R_c^2}\left(\frac{1}{q^2} - 1\right)\sin 2(i - v). \tag{2.213}$$

We next need to estimate the force that is due to the disk. If the typical size of the optical disk is $r_{\max}$ then, at distances $r \gg r_{\max}$, the potential that is due to the disk can be expressed in the form (see exercise 2.7)

$$\Phi_D(R, z) = -\frac{GM_D}{\sqrt{R^2 + z^2}} - \frac{G}{4}\frac{(R^2 - 2z^2)}{(R^2 + z^2)^{5/2}}\int_0^\infty 2\pi R'^3 \Sigma(R')\, dR'. \tag{2.214}$$

Given this potential, we can compute the force and the torque that are due to the disk on the ring of material. This is also directed along the y axis and has a magnitude

$$N_{\text{disk}} = \frac{3G}{8r_0^2}\left[\int_0^\infty 2\pi R'^3 \Sigma(R')\, dR'\right]\sin 2i. \tag{2.215}$$

The equilibrium shape of the gas is decided by the surface on which $N_{\text{halo}} + N_{\text{disk}} = 0$. For an exponential disk with $\Sigma(R) = \Sigma_0 \exp(-R/R_d)$ and total mass $M_D = 2\pi\,\Sigma_0 R_d^2$, this equation leads to the condition

$$\frac{v_0^2 r^5}{9GM_D R_c^2 R_d^2}\left(\frac{1}{q^2} - 1\right) \equiv \left(\frac{r}{r_w}\right)^5\left(\frac{1}{q^2} - 1\right) = \frac{\sin 2i}{\sin 2(v - i)}. \tag{2.216}$$

The right-hand side of Fig. 2.14 shows the shape of this surface for $v = 30°$, $45°$, $60°$, $75°$, $q = 0.9$. The x axis is in units of

$$r_w = \left(\frac{9GM_D R_c^2 R_d^2}{v_0^2}\right)^{1/5} q^{2/5}(1 - q^2)^{-1/5}, \tag{2.217}$$

which defines the characteristic scale for the warp. We note that for $r \ll r_w$, the equilibrium surface coincides with plane of the *disk*, whereas beyond $r = 2r_w$, it shifts to the equatorial plane of the *halo*.

Exercise 2.11
Filling in blanks: Prove Eqs. (2.213) and (2.215).

2.8 Chemical Evolution of Galaxies

The discussion so far treated galaxies as a chemically homogeneous medium interacting through different dynamical processes. Because stellar evolution leads to the synthesis of chemical elements that are distributed within the galaxy through supernova bursts, it is obvious that the chemical content of the galaxy

will vary as stars evolve. This chemical evolution of a galaxy is a complex process because of the coupling between the stars and the ISM and the time dependence of the phenomena arising from stellar and galactic evolutions. At a fundamental level, this problem needs to be addressed in terms of local conservation laws for mass, momentum, and energy for the ISM and stars.[5] For example, the equation for mass conservation will read as

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \rho \mathbf{v} = s_{\text{mass}}, \tag{2.218}$$

where the right-hand side takes into account conversion of interstellar gas into the stars as well as the addition of mass into the ISM from the stars through stellar winds and supernova explosions. We could write similar equations for conservation of energy and momentum. Given the details of stellar-evolution processes as well as the coupling between stars and the ISM, we can in principle solve these equations numerically, thereby obtaining the spatial and temporal variations of different properties (for example, the metallicity) in a galaxy. This procedure unfortunately is not practical because of mathematical complexity as well as the uncertainties in the input physics. Even if we succeeded in solving the detailed equations, it would be difficult to compare the results with observations at the same level of detail. Hence chemical evolution of the galaxy is usually studied in a coarse-grained manner in terms of parameters defined globally over a galaxy.

Some of the key limitations of this coarse-grained approach can be easily seen from the following argument. To ignore the spatial variations of physical processes (say, evolution of metallicity), we need to assume that homogenisation takes place at a time scale much shorter than the evolution time scale we are interested in. The relevant time scale to mix the constituents over 1 kpc, say, by physical processes operating in the ISM (with characteristic speeds of ~ 10 km s^{-1}) will be $\sim 10^8$ yr, which is comparable with the evolution time scale we are interested in. Similar estimates can be made for time scales for momentum and energy transfer as well. An object of mass m moving through interstellar gas of density ρ_{ISM} with a velocity v will accrete mass at the rate $\dot{M} = \rho_{\text{ISM}} v(\pi r_G^2)$, where $r_G = 2Gm/v^2$ is the radius over which the gravitational influence of the mass would be felt in the ISM. This gives

$$\frac{dM}{dt} = \rho_{\text{ISM}} \frac{4\pi (GM)^2}{v^3} \approx 5 \times 10^{11} n \text{ gm s}^{-1} \approx 8 \times 10^{-15} n \, M_\odot \text{ yr}^{-1}. \tag{2.219}$$

The characteristic time scale is given by $\tau_{\text{acc}} = m/\dot{M}$. This is very large ($\gtrsim 10^{14}$ yr) for a stellar-mass object moving through an ISM with $n \simeq 1$ cm^{-3}. However, in a giant molecular cloud (GMC), with $n = 10^4$ cm^{-3} and $v = 5 \times 10^4$ cm s^{-1}, we get $\tau_{\text{acc}} \simeq 10^6$ yr. Thus mass accretion could be important in GMCs where star formation takes place.

This accretion process leads to a momentum transfer that takes place at the same time scale τ_{acc}. In addition to this, stellar winds also act as a source of

momentum in the ISM. The stellar winds exert a ram pressure of $P_w = \rho_w v_w^2$. Taking the mass-loss rate from a star as $(dm/dt) = 4\pi r^2 \rho_w v_w \simeq 10^{-6}\, M_\odot\ \mathrm{yr}^{-1}$ and $v_w \simeq 100\ \mathrm{km\ s}^{-1}$, we get $P_w \simeq 5 \times 10^{26}/r^2\ \mathrm{dyn\ cm}^{-2}$. The gas pressure of ISM is $P_{\mathrm{ISM}} \simeq 10^{-16} nT\ \mathrm{dyn\ cm}^{-2}$. This shows that $P_w = P_{\mathrm{ISM}}$ at a distance of $r \approx 7$ pc from the star. In other words, this effect can be important at the typical interstellar spacing of $r \simeq 1$ pc, even in the vicinity of a star with a moderate mass-loss rate of $(dm/dt) \simeq 2 \times 10^{-8}\, M_\odot\ \mathrm{yr}^{-1}$. Hence we cannot ignore the input of momentum into the ISM in regions having stars with appreciable mass loss.

Although these results show that the coarse-grained approach has limitations, the absence of detailed knowledge forces us to model the average evolution of the galaxy by making several assumptions to simplify the analysis. We begin with a general description of this approach.

Integrating mass conservation law (2.218) over a given volume $\mathcal{V}$ will lead to an equation of the form

$$\frac{\partial G}{\partial t} + \nabla \cdot (G\mathbf{v}) = \int_{\mathcal{V}} s_{\mathrm{mass}}\, d^3x + I, \qquad I = -\int_{\Sigma} \rho \mathbf{v} \cdot d\Sigma, \quad (2.220)$$

where I denotes the net mass flowing through the external boundary Σ of the region $\mathcal{V}$. The first term on the right-hand side can be expressed in the form

$$\int_{\mathcal{V}} s_{\mathrm{mass}}\, d^3x = -F + E, \quad (2.221)$$

where F is the rate of conversion of the ISM into stars and E is the rate of ejection of mass from the stars to the ISM. If we now further assume that the system is homogeneous, then we can ignore the spatial derivative on the left-hand side.

Although ignoring the spatial derivatives will make the equation depend only on time, it will remain in general nonlocal in time for the following reason. We note that the SFR F is given by

$$F = \int_0^\infty B(m, t)\, dm, \quad (2.222)$$

where $B(m, t)$, usually called the *birthrate function*, is the rate at which gas is converted at time t into stars of mass m measured in units of solar mass; $m = M/M_\odot$. The ejection rate E can be expressed in the form

$$E(t) = \int_0^\infty \int_0^t B(m, t-a)\, R(a, m)\, dm\, da, \quad (2.223)$$

where $R(a, m)$ is the rate at which a star of mass m and age a returns the mass to the ISM. It is obvious that the term $(E - F)$ is nonlocal in time.

These equations can be simplified if further assumptions are made with the inputs from stellar-evolution theory. Most of the stars return a negligible amount of mass to the ISM for most of their lifetime, the major exception being the massive stars with very short lifetimes. In their case, we can make an approximation

that the mass ejection takes place at the end of their life at $a = \tau_m$. This will allow us to approximate the return function by $R(a, m) \approx \delta(a - \tau_m)R(m)$, so that

$$E(t) \approx \int_0^\infty B(m, t - \tau_m)R(m)\,dm. \tag{2.224}$$

With this approximation, $E(t)$ and $F(t)$ can be determined from stellar evolution theory, allowing us to integrate the relevant equations.

It is, however, conventional to make another related approximation known as *instantaneous recyling*. If the rate of change of birthrate $B(m, t)$ is small compared with the lifetime of the stars, then we can write $B(m, t - \tau_m) \approx B(m, t)$. Further, we take the function $B(m, t)$ to be separable into the variables m and t in the form $B(m, t) = \Lambda(t)\xi(m)$, where $\xi(m)$ is the *initial mass function* (IMF). We now have $E(t) = f\,\Lambda(t)$, where f is a constant given by

$$f = \int_0^\infty R(m)\,\xi(m)\,dm. \tag{2.225}$$

There is a considerable amount of uncertainty in the form of the IMF $\xi(m)$ (see Vol. II, Chap. 3, Section 3.6). One of the simplest characterisations of this function is in the form $\xi(m) = \xi_0 m^{-\alpha}$, with $\xi_0 \simeq 1.5 \times 10^{-2}\,\mathrm{pc}^{-3}$ and $\alpha \simeq 1.33$. Very often we use the normalised IMF $\phi(m)$ with

$$\int_{m_L}^{m_U} m\,\phi(m)\,dm = 1 \tag{2.226}$$

with some lower and upper limits on the mass, $m_L \simeq 0.1$ and $m_U \simeq 100$. The sensitivity of the pure power-law functions to the limits of integration can be avoided if a suitable series of power laws is used in different ranges. One such fitting function has the form

$$m\phi(m) = Am^B, \tag{2.227}$$

with the constants A and B having different values for different ranges as follows: $A = 0.93$, $B = 0.15$ for $0.1 \le m \le 0.5$; $A = 0.46$, $B = -0.85$ for $0.5 \le m \le 1$; $A = 0.46$, $B = -2.4$ for $1.0 \le m \le 3.16$; and $A = 0.21$, $B = -1.7$ for $3.16 \le m \le 100$. An approximate single power-law fit for this range is given by $m\phi(m) = 0.17m^{-1.35}$. Either of these will lead to 31% of the mass fraction in the interval $0.1 < m < 0.5$, another 31% in the range $0.5 < m < 1$, a mass fraction of 0.26 in the range $1 < m < 3.16$, and a mass fraction of 0.12 in the range $3.16 < m < 100$.

If the mass fraction locked up in stars and stellar remnants is α, then the return fraction is $R = 1 - \alpha$. The locked-up fraction can be computed as

$$\alpha = 1 - F_M(> m_\tau) + \int_{m_\tau}^{m_U} m_{\mathrm{rem}}(m)\,\phi(m)\,dm, \tag{2.228}$$

where $F_M(> m_\tau)$ is the fraction of stars with masses greater than a critical mass

m_τ, which we evaluate by equating the lifetime $\tau(m)$ to the age of the galaxy. In other words, $1 - F_M(> m_\tau)$ gives the fraction of mass that is locked up in long-lived stars and the last term of Eq. (2.228) gives the mass that is locked up in stellar remnants. The latter can be estimated as

$$m_{\rm rem} = \begin{cases} 0.11\, m + 0.45, & (m \le 6.8) \\ 1.5, & (m > 6.8) \end{cases} . \tag{2.229}$$

If we take $m_\tau = 1$, most of the IMFs will give $\alpha \simeq (0.7\text{–}0.8)$.

Given this background, an illustrative model for the chemical enrichment of a galaxy can be constructed along the following lines. Let us consider a region of galaxy containing a mass $M_g(t)$ in gas and $M_s(t)$ in stars, and let us assume that gas with metallicity Z_f is added or ejected from that region at a rate $\dot{M}_f$, where an overdot denotes the time derivative. Conservation of mass implies that

$$\dot{M}_g + \dot{M}_s = \dot{M}_f . \tag{2.230}$$

The conservation of heavy elements requires

$$d(ZM_g) = p\,dM_s - Z\,dM_s + Z_f\,dM_f . \tag{2.231}$$

The left-hand side is the change in the heavy elements in the gaseous phase; the first term on the right-hand side gives the nucleosynthesis yield that is due to processing in stars. We take this yield to be proportional to the processed stellar mass parameterised by a yield parameter p. The second term on the right-hand side is the amount of heavy elements locked in stars and the last term arises from sources or sinks present in the region. This equation assumes (as previously explained) that stellar processes are efficient in changing the heavy-element abundance instantaneously (i.e., at a time scale shorter than any other timescale in the problem) and that everything is spatially homogeneous. Dividing this equation by dt and using conservation of mass (2.230), we get

$$\dot{Z}M_g - p\dot{M}_s = \left(Z_f - Z\right)\dot{M}_f . \tag{2.232}$$

To proceed further, we need to make some specific assumptions regarding the sources and sinks. As the first example, let us consider a model in which the gas is blown out of the region at a rate that is a constant fraction ν of the SFR so that $\dot{M}_f = -\nu\dot{M}_s$. From mass conservation (2.230) we get $\dot{M}_g = -(1+\nu)\dot{M}_s$, which integrates to give

$$M_g(t) = M_g(0) - (1+\nu)M_s(t). \tag{2.233}$$

Using this in Eq. (2.232) with the assumption $Z_f = Z$, we get

$$Z = \frac{p}{1+\nu}\ln\frac{M_g(0)}{M_g(t)} = \frac{p}{1+\nu}\ln\left\{1 + (1+\nu)\left[\frac{1}{f_g(t)} - 1\right]\right\}, \tag{2.234}$$

where $f_g \equiv M_g/(M_g + M_s)$ is the fraction of gas.

The model shows that the metallicity increases as the gas is converted into stars and star formation effectively ceases when all the gas is exhausted, which

occurs when $M_s = M_g(0)/(1+\nu)$. Hence the mean stellar metallicity at late times is given by

$$\bar{Z}_s = \frac{1+\nu}{M_g(0)} \int_0^{M_g(0)/(1+\nu)} Z\, dM_s = \frac{p}{1+\nu}. \tag{2.235}$$

The results from Eqs. (2.234) and (2.235) allow us to make some simple deductions that could be compared with observations. Note that the mean metallicity $\langle Z \rangle$ will approach the true yield p (if $\nu = 0$) or $p(1+\nu)^{-1}$ (if $\nu \neq 0$) when the gas fraction goes to zero. It is also clear that the ratio of metallicities for two elements will be the same as the ratio of their yields: $Z_i/Z_j = p_i/p_j$. There have been several attempts to test these results against observations with a mixed level of success. Although there is some evidence suggesting that these predictions are true, there are also significant discrepancies in the case of some elements.

This model also predicts a logarithmic relation between the metallicity $Z(t)$ and the gas fraction $M_g(t)$, which can be tested if we can identify a population of objects with similar physical conditions but different amounts of gas fraction. This was attempted in a study of the oxygen abundance in a class of irregular and blue compact galaxies. The results show that the model is in broad agreement with the observations, although the error bars are fairly large. Similar results were obtained in the study of metallicities of F- and G-type stars as a function of their ages.

To get actual numbers, we need to estimate ν and p, which turns out to be fairly difficult. We can attempt to construct a simplified model for ν along the following lines. We have seen in Vol. 1, Chap. 6, Section 6.12, that the cooling rate of a gas decreases with temperature in the range $(10^4 - 3 \times 10^6)$ K, indicating an instability. Hence we would expect the ISM of a galaxy to contain gas with temperatures $T < 10^4$ K and $T > 3 \times 10^6$ K. The one-dimensional velocity dispersion of the hotter component of gas will be $\sigma_h \approx 200$ km s^{-1}. For comparison, stars at the centres of dwarf spheroidals have a dispersion of $\sigma_* \approx (5\text{-}10)$ km s^{-1} whereas those at the centre of giant ellipticals have $\sigma_* \gtrsim 200$ km s^{-1}. Hence, in a dwarf spheroidal, the hot gas is likely to flow freely out of the system although in a giant elliptical we would expect the gravitational field to be able to confine the hot component of the ISM. To obtain a quantitative relationship between σ_* and Z in a giant elliptical, we can equate the rate at which ejected gas gains potential energy to the rate at which a supernova injects energy into the ISM. This would suggest that

$$\left| \Phi \frac{dM_g}{dt} \right| \simeq f\sigma_*^2 \nu \frac{dM_s}{dt} = E_{\rm SN} \frac{dM_s}{dt}, \tag{2.236}$$

where we have taken the gravitational potential to be $\Phi = -f\sigma_*^2$, with f denoting a numerical factor. Assuming that one core-collapse supernova injects $\sim 10^{51}$ ergs

into the ISM and that one such event occurs for 200 $M_\odot$ of star formation, we get $E_{SN} \simeq 2.5 \times 10^{15}$ ergs gm^{-1}. This gives

$$\nu = \frac{0.6}{f}\left(\frac{200 \text{ km s}^{-1}}{\sigma_*}\right)^2, \tag{2.237}$$

suggesting that ν will increase from a value of less than 1 for giant ellipticals to an order of a few for dwarf spheroidals. Combining Eq. (2.237) with Eq. (2.235) and using $\nu \propto \sigma_*^{-2}$, we would expect a relation of the form

$$\bar{Z}_s = \frac{p}{1 + (\sigma_0/\sigma_*)^2}, \tag{2.238}$$

where σ_0 is a constant. Observationally, there are some data giving the abundance (based on MgII) for systems with different σ_* that seems to be in broad qualitative agreement with Eq. (2.238). There is no direct confirmation of flattening of $\bar{Z}$ for large σ_*, although this could depend on the relationship between the MgII line strength and total mean metallicity.

A similar analysis based on Eq. (2.235) can also be used to estimate the radial gradient of the metallicity in ellipticals. We have

$$\frac{d \ln \bar{Z}_s}{d \ln r} = -\frac{r}{1+\nu}\frac{d\nu}{dr} = \frac{\nu}{1+\nu}\frac{v_c^2}{\Phi} = \frac{E_{SN}}{E_{SN} - \Phi}\frac{v_c^2}{\Phi}, \tag{2.239}$$

where we have used $\nu = E_{SN}/|\Phi(r)|$ and $v_c = \sqrt{r d\Phi/dr}$. For plausible values of Φ and E_{SN}, this gives a gradient of -0.2, which is in broad agreement with observation.

The second parameter that appears in the preceding model is the average yield p, which, in principle, could be estimated from stellar-evolution theory. From our discussion in Vol. 2, Chap. 3, Section 3.6, we know that stars of different initial masses produce different amounts of metals and lock up different amounts of mass in the final stage. To a very good approximation, the mass of heavy elements $M(Z)$ ejected during the evolution of a star of mass $M = mM_\odot$ is given by

$$M(Z) \approx M\left(0.4 - \frac{4}{m}\right) \qquad (10 \lesssim m \lesssim 25) \tag{2.240}$$

for M in the range of 10–25 $M_\odot$. After approximately 5×10^9 yr, stars with mass $M < 1.2\ M_\odot$ will still be in the main sequence; stars with mass $1.2\ M_\odot < M < 7\ M_\odot$ will leave white dwarf remnants of mass $0.6\ M_\odot$; and stars of mass $7\ M_\odot < M < 25\ M_\odot$ will leave neutron star remnants of mass $1.4\ M_\odot$. The extreme limits are more uncertain, but for simplicity we will assume that (1) no brown dwarfs are formed and (2) stars with $M > 25\ M_\odot$ leave black hole remnants of mass $M/2$. Given a mass function of stars $\psi(m)\,dm$, we can then

easily determine the total yield of heavy elements as

$$Y = \int_{10}^{\infty} dm\, \psi(m) M \left(0.4 - \frac{4}{m}\right). \tag{2.241}$$

The mass fraction α that is locked up in stars is given by Eq. (2.228). The parameter p is just the ratio $p = (Y/M_{\rm lock})$. The actual numerical value, of course, depends sensitively on the asymptotic behaviour of $\psi(m)$. If we take the Miller–Scalo IMF with

$$\psi = Cm^{-1.9} 10^{-0.28(\log m)^2}, \tag{2.242}$$

we find that $Y \approx 0.161C$ and $M_{\rm lock} \approx 2.67C$, giving $p = 0.06$. On the other hand, the Salpeter IMF gives $p \simeq 0.065/\alpha \simeq 0.08$. If we take $\nu \simeq 2$ from Eq. (2.237) then $p/(1+\nu) \simeq 0.02$. Equation (2.234) suggests that $Z = Z_\odot \simeq 0.02$ when $M_g \simeq e^{-1} M_g(0) \simeq 0.37\, M_g(0)$.

The *simplest* model of chemical evolution, of course, will correspond to $\nu = 0$, which is usually called the *closed-box model*. Equation (2.234) with $\nu = 0$ shows that

$$f_s \equiv \frac{M_s(z)}{M} = 1 - \exp(-Z/p), \tag{2.243}$$

where $f_s \equiv M_s(Z)/M$ is the fraction of mass in stars when the metallicity is Z. It follows that

$$\frac{df_s}{d\log z} \propto z e^{-z}, \qquad z = \frac{Z}{p}, \tag{2.244}$$

implying that the number of stars with a given metallicity has a characteristic distribution that has a maximum at $z = 1$ or $Z = p$. We can test this result by studying the distribution of stars with different abundances for, say, [O/H]. Observations suggest that the theory fits the observations but with $p = 10^{-1.1}\, Z_\odot$. This modification can be easily incorporated into a model with $\nu \neq 0$, which allows for a mass loss at a rate proportional to the SFR. In such a case, the distribution in relation (2.244) will change to the form $ze^{-(1+\nu)z}$, which is equivalent to using an effective yield of $p/(1+\nu)$.

Also note that, in the $\nu = 0$ model, the metallicity increases as the gas fraction decreases. Such a conclusion, of course, directly contradicts the existence of globular clusters that have very little gas but the stars have a metal-poor composition. Similarly dwarf spheroidal galaxies contain very little gas, although their stars have metal abundances which are 30–100 times lower than the galactic bulge. It is likely that, in both the cases, stars formed out of well-mixed initial gas and any residual gas were expelled from the system. This suggests that the flow of material in and out of the system needs to be taken into account for a more realistic result.

The most serious difficulty faced by the closed-box model is in the case of long-lived G-dwarf stars in a solar neighbourhood. Observations show that the fraction of low-metallicity G-dwarf stars is significantly lower than that predicted by this model. This problem, however, can be solved by a suitably modelled infall of gas (although it is not clear whether this is the best possible approach). One of the simplest models that solves the G-dwarf problem is based on a postulate that relates the gas mass to the star mass in the system by an (ad hoc) quadratic expression:

$$g(s) = \left(1 - \frac{s}{M}\right)\left(1 + s - \frac{s}{M}\right). \tag{2.245}$$

Here $g(s)$ is the gas mass, s is the mass in stars, and M is a constant parameterising the relation. All masses are measured in units of the initial mass, and M can be interpreted as the final mass in stars when $g = 0$. From Eq. (2.231) for the evolution of metallicity (in a slightly modified notation with $Z_f = 0$),

$$\frac{d}{ds}(Zg) = p - Z, \tag{2.246}$$

we can easily obtain the solution for $z = Z/p$ as

$$z(s) = \frac{Z}{p} = \left[\frac{M}{1 + s - (s/M)}\right]^2 \left[\ln\frac{1}{1 - s/M} - \frac{s}{M}\left(1 - \frac{1}{M}\right)\right]. \tag{2.247}$$

Correspondingly, the characteristic distribution in metallicity now becomes

$$\frac{ds}{d\ln z} = \frac{z[1 + s(1 - 1/M)]}{(1 - s/M)^{-1} - 2z(1 - 1/M)}. \tag{2.248}$$

Figure 2.15 shows this function for different values of M. Note that, compared with the original model of $M = 1$, this class of models with $M > 1$ allows for a relative enhancement of higher-metallicity stars with respect to G dwarfs.

Exercise 2.12

Chemical evolution with inflow: Another extreme form of evolution corresponds to assuming that the primordial gas flows into the region at a rate that is a constant fraction of the SFR. In this case $\dot{M}_f = +\nu\dot{M}_s$ and $Z_f = 0$. (a) Show that this implies that the amount of mass contained in stars with metallicity less than a particular value Z scales as

$$F(Z, Z_1) \equiv \frac{M_s(< Z)}{M_s(< Z_1)} = \frac{M_g(0) - M_g(< Z)}{M_g(0) - M_g(< Z_1)} = \frac{1 - q}{1 - q_1}, \qquad q = \frac{M_g(< Z, t)}{M_g(0)}. \tag{2.249}$$

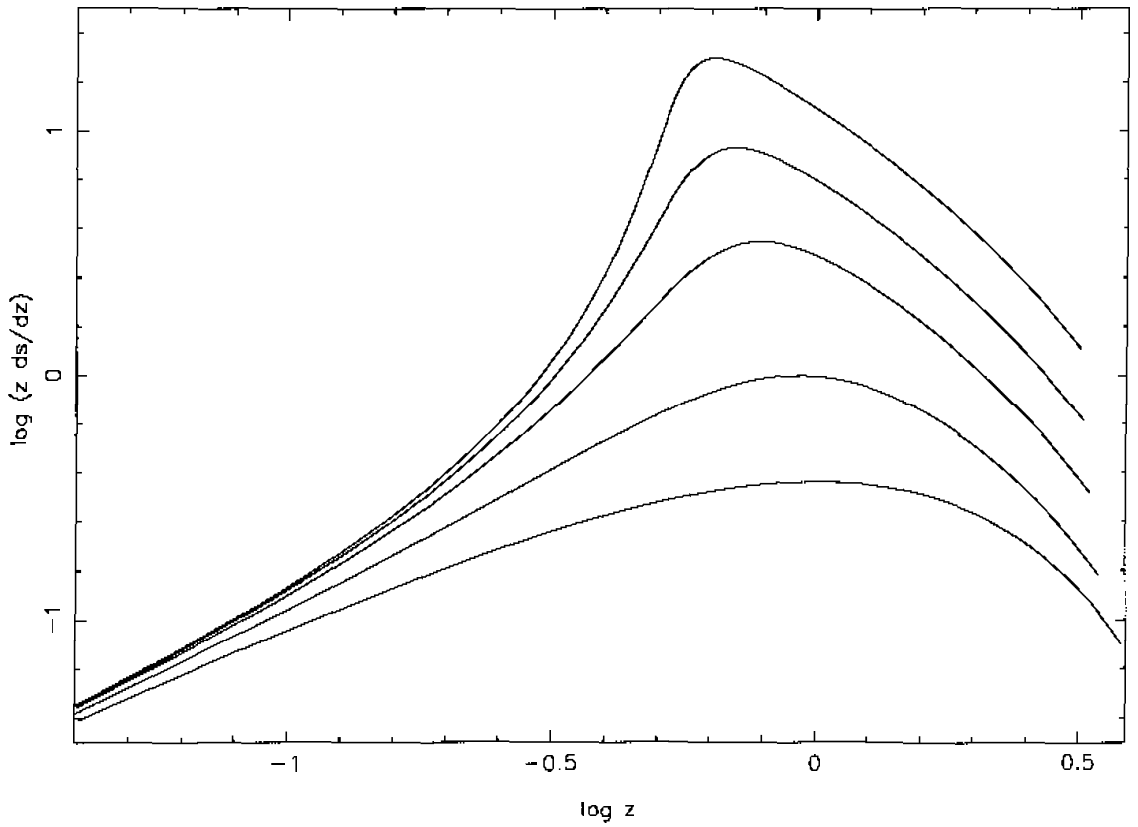


Fig. 2.15. Abundance distribution function for parameterised model with in-fall for $M = 1, 2, 5, 10, 20$ from bottom to top.

Express this result in the form

$$F(Z, Z_1) = \frac{1}{1 - q_1} \left\{ 1 - \left[1 - \frac{Z}{Z_1} \left(1 - q_1^{\nu/(1-\nu)} \right) \right]^{(1-\nu)/\nu} \right\}. \tag{2.250}$$

Also show that q_1 is related to the gas fraction by

$$q_1 = \left[1 + (1 - \nu) \left(\frac{1}{f_{g1}} - 1 \right) \right]^{-1}. \tag{2.251}$$

(b) Observationally it is found that the F- and G-dwarf stars with $Z < (1/6)Z_1$ contribute ~2%; that is, $F[Z = (1/6)Z_1, Z_1] \simeq 0.02$. (In the solar neighbourhood we can take $f_{g1} = 0.1$ for $Z_1 = 0.03 = 1.5\ Z_\odot$.) Can ν be adjusted to get the correct abundance of G dwarfs?

Finally, we describe a general result connecting the mass of the heavy elements produced, the luminosity of the galaxy, and the index of the IMF. As we saw in Chap. 1, Section 1.6, most of the present-day luminosity comes from stars whose original masses were in a narrow range of values around the characteristic mass M_{GB} of stars on the present giant branch, whereas the heavy elements generated were all contributed by stars more massive than some critical mass M_Z. Assuming that the massive stars reinject a fraction p of the original mass into the ISM in

the form of heavy elements, we find that the mass of heavy elements injected into the ISM will be

$$M_h = \int_{M_Z}^{\infty} pM \frac{dN}{dM}\, dM. \tag{2.252}$$

Taking a simple IMF with $(dN/dM) = K(M/M_\odot)^{-\alpha}$ for $M < M_{\max}$ and 0 for $M > M_{\max}$, we get

$$M_h = pKM_\odot^2 \int_{M_Z/M_\odot}^{M_{\max}/M_\odot} m^{(1-\alpha)}\, dm = \frac{pKM_\odot^2}{\alpha - 2}\left[\left(\frac{M_\odot}{M_Z}\right)^{\alpha-2} - \left(\frac{M_\odot}{M_{\max}}\right)^{\alpha-2}\right]. \tag{2.253}$$

Because $\alpha > 2$, we can ignore the second term if $M_{\max} \gg M_Z$. Combining this with relation (1.55) of Chap. 1, Section 1.6, we get the mass of the heavy elements injected into the ISM per unit luminosity generated as

$$\frac{M_h}{L} = \frac{pM_\odot^2}{\alpha - 2}\left[\frac{25\ \text{Gyr}}{E_{\text{GB}}(M_{\text{GB}})}\right]\left(\frac{M_Z}{M_\odot}\right)^{2-\alpha}\left(\frac{M_{\text{GB}}}{M_\odot}\right)^{\alpha-3.5}. \tag{2.254}$$

Inserting the typical numbers $p \simeq 0.1$, $\alpha \simeq 2.4$, $M_{\text{GB}} \simeq 0.85\, M_\odot$, and $E_{\text{GB}} \simeq 2.9 \times 10^{10}\, L_\odot$ yr, we get

$$\frac{M_h}{M_\odot} \simeq 0.26\left(\frac{M_Z}{M_\odot}\right)^{2-\alpha}\frac{L}{L_\odot}. \tag{2.255}$$

The value for M_Z depends on the elements we are interested in. For carbon, nitrogen, and oxygen, $M_Z \simeq 2\, M_\odot$, whereas for elements produced in a core-collapse supernova, $M_Z \simeq 8\, M_\odot$. Because $(2 - \alpha) \approx -0.4$, the result is not very sensitive to the value of M_Z. We find that (M_h/L) varies in the range $0.11 - 0.20\ M_\odot/L_\odot$ for most elements.

2.9 Galaxy Interactions and Mergers

So far we have been treating galaxies as isolated noninteracting systems. It is, however, fairly easy to see that galaxy interactions cannot, in general, be ignored in the study of their dynamics. Even in the case of field galaxies, with a mean separation $l \simeq 1$ Mpc and relative velocities of $v \simeq 200$ km s^{-1}, the corresponding time scale is $t \simeq (l/v) \simeq 5 \times 10^9$ yr. This suggests that most galaxies would have undergone some form of gravitational interaction with another galaxy during the lifetime of the universe.[6] The conclusion is strengthened when we take into account the fact that galaxies are clustered and a fraction of them occur in high-density regions which will decrease l and increase v.

The actual gravitational interaction of two galaxies is an enormously complicated affair that leads to a wide variety of effects. Even in the simplest contexts in which each galaxy is treated as being made up of point particles interacting

only through gravity (thereby ignoring the entire gaseous phase of the ISM in the galaxy), there are several parameters that play a role in determining the outcome, and, in general, we must resort to numerical simulations to understand these processes. Given the density profile of the two galaxes, when they are far away from each other (thereby determining the individual masses, velocity dispersion, and gravitational potential energies) and the trajectories of the centres of mass of the galaxies under their mutual influence, we can identify a few cases in which analytic understanding is possible. We shall first describe these situations.

The first case corresponds to a situation in which the mass of one of the galaxies is significantly higher than that of the other. In this situation, the lighter galaxy will move in the gravitational potential of the heavier one and will, in general, lose energy and angular momentum because of dynamical friction, thereby merging with the heavier galaxy. This process can be effective not only in the case of galaxies but even for globular clusters orbitting in the halo of the galaxies. Mergers also occur between objects of comparable mass – usually called galactic cannibalism – but those cases are difficult to treat analytically.

The second situation in which some progress can be made corresponds to high-speed encounters that can be defined by the following criterion. If the colliding systems have masses M_1 and M_2, radii R_1 and R_2, impact paramter b, and relative velocity V, then the time scale for the encounter could be estimated as $t_{\rm enc} \simeq \max(R_1, R_2, b)/V$. If the internal-velocity dispersions are σ_1 and σ_2, then the crossing times for stars in the galaxies will be $t_i = R_i/\sigma_i$ with $i = 1, 2$. If $t_i \gg t_{\rm enc}$, we can assume that the stars have moved very little during the time of encounter. This translates into the condition $V \gg (\sigma_i/R_i)\max(R_1, R_2, b)$, which can be satisfied at high speeds.

It is straightforward to show that, when this condition is satisfied, we can also approximate the motion of the two centres of colliding galaxies as one of uniform velocity. Focussing the attention on one system (of mass M_1, which we will call the perturbed system), we can compute the change in the kinetic energy of the stars that is due to the passage of the second system of mass M_2 and trajectory $\mathbf{X}(t) \simeq (0, b, Vt)$. This was done in Vol. II, Chap. 10, Section 10.7, and we found that the change of velocity of the αth star is given by

$$\Delta \mathbf{v}_0 = -\int_{-\infty}^{\infty} \left[\nabla\Phi(\mathbf{r}_\alpha, t) - \frac{1}{M_1}\sum_\beta m_\beta \nabla\Phi(\mathbf{r}_\beta, t) \right] dt, \qquad (2.256)$$

where $\Phi(\mathbf{r}, t)$ is the gravitational potential that is due to the perturber, which is treated as a known quantity. The potential energy of the stars does not change in this *impulse approximation*, whereas the kinetic energy of the perturbed system increases. The actual estimate of the change in the energy simplifies in two limiting cases. The first one was studied in Vol. II, Chap. 10, Section 10.7, under the tidal approximation in which it was assumed that R_i/b is a small quantity. In such distant encounters, the amount of energy transferred varies

as b^{-4}. The other limiting case is a penetrating encounter corresponding to a head-on collision with $b \to 0$. A more general case can be treated (to a good approximation) as an interpolation between the two.

Immediately after the encounter, neither system is in virial equilibrium as their potential energies are unchanged but their kinetic energies have increased. As the systems relax back to virial equilibrium, the total kinetic energy actually *decreases*, as can be seen by the following argument. Originally, in virial equilibrium, the kinetic energy T_0 and the total energy E_0 are related by $T_0 = -E_0$. If the encounter increases the kinetic energy by δT, the final energy is $E_1 = E_0 + \delta T$, and the final kinetic energy in the final virialised configuration will be

$$T_1 = -E_1 = -(E_0 + \delta T) = T_0 - \delta T. \tag{2.257}$$

Thus the final relaxation process has to decrease the kinetic energy from $T_0 + \delta T$ to $T_0 - \delta T$. This is usually accomplished by the system expanding after the encounter in some form, either by the ejection of more energetic components or by a gentler redistribution of mass. We will discuss the issue of mergers of systems with widely different masses in Subsection 2.9.1 and will study some of the effects of penetrating encounters in Subsection 2.9.2.

One of the key effects of interaction between galaxies seems to be an enhancement in the SFR – often in the form of star bursts. Because it is not possible to view a system both before and after interaction (except, of course, in simulations), our conclusions are necessarily based on statistical analysis of interacting and noninteracting samples of galaxies with respect to the SFR. Interacting systems are often seen to show evidence of young stars, an excess of type II supernova outbursts, evidence for line emission that indicates an overabundance of young OB stars, and other indications of enchanced SFR. Although the evidence is not completely conclusive, it is generally believed that the gravitational perturbation of one galaxy on another can induce instabilities in either or both of them involving density enhancements in gas and a consequent increase in SFR.

Mergers of galaxies also lead to striking visual signatures that include tails of material, polar rings, and shells of matter, or counterrotating stellar or gaseous subsystems. Detailed numerical simulations have been able to reproduce most of these features as a result of strong gravitational encounter between two galaxies, although the detailed modelling and statistical study do not seem to lead to any conclusions that are definitive and useful beyond the morphological aspects. In Subsections 2.9.1–2.9.3 we shall briefly describe the simple analytic results and some results from numerical simulations.

2.9.1 Galactic Cannibalism

One of the most dramatic effects of galaxy–galaxy interaction is a phenomena called *galactic cannibalism*, in which a large galaxy gobbles up smaller

neighbours over a period of time. The chance for this phenomenon to occur over a Hubble time is fairly large and is enhanced by the fact that galaxies cluster around each other. The number density $n(r, L)\,dL$ of galaxies near another galaxy varies as (see Chap. 1, Section 1.13)

$$n(r, L)\delta L = \phi(L)\left[\left(\frac{r_0}{r}\right)^{1.8} + 1\right]\delta L, \tag{2.258}$$

where $\phi(L)$ is the galaxy luminosity function and $r_0 \approx 5h^{-1}$ Mpc. If the orbiting galaxies slowly spiral towards the central galaxy, then the inward current that is due to dynamical friction will be $dJ(r, L) = 4\pi r^2 v_r n(r, L)\,dL$. The total number of galaxies eaten up by a large galaxy in one Hubble time $t_H \equiv H_0^{-1}$ is obtained by integration of $n(r, L)$ from $r = 0$ to a radius $r_i(t_H, L)$ at which $t_{\text{fric}} = t_H$. Taking $r \ll r_0$ and using the results of Section 2.5 (see Eq. (2.165)), we get

$$dN(L) = 4\pi\, dL \int_0^{r_i} n(r, L)r^2\, dr = 9.5\left(\frac{GQL_* \ln\Lambda}{H_0 r_0^2 v_c}\right)^{0.6} r_0^3 \phi(L)\, dL, \tag{2.259}$$

where we have taken the mass of the galaxy to be $M = QL$, with $Q \approx 12h(M_\odot/L_\odot)$ being the mass-to-light ratio of a typical galaxy. We can obtain the total amount of luminosity gained by a galaxy by integrating this expression with the Schecter luminosity function:

$$\begin{aligned} L_{\text{gain}} &= \int_0^\infty L\,dN(L) \approx 9.5\left(\frac{GQL_* \ln\Lambda}{H_0 r_0^2 v_c}\right)^{0.6} r_0^3 n_* L_* \int_0^\infty x^{1.6+\alpha} e^{-x}\, dx \\ &= 9.5\left(\frac{GQL_* \ln\Lambda}{H_0 r_0^2 v_c}\right)^{0.6} r_0^3 n_* L_* \Gamma(2.6+\alpha). \end{aligned} \tag{2.260}$$

For typical values of $v_c \approx 300$ km s^{-1}, $r_0 \approx 5h^{-1}$ Mpc, $\ln\Lambda = 3$, and the Schecter luminosity function, we find that L_{gain}/L_* is approximately 10%–15%. This shows that galactic cannibalism is a significant effect in increasing the intrinsic luminosity of the galaxies.

Although it is difficult to obtain direct observational signatures for determining whether a galaxy has undergone merging in its past, some insights can be gained from scaling arguments if we assume that the light distribution before and after merging satisfies some universal functional form: $\Sigma(r)/\Sigma_0 = F(r/r_0)$. For constant mass-to-light ratios, the total mass and binding energy of the galaxy will then scale as $M \propto \Sigma_0 r_0^2$, $E \simeq (GM^2/r_0) \propto \Sigma_0^2 r_0^3$. Mergers that conserve both mass and energy can be characterised by two extreme forms of behaviour. The first one corresponds to a massive galaxy's merging with several low mass galaxies with small binding energies, which will increase M at constant E, giving $\Sigma_0 \propto M^{-3}$. The second one involves the merging of nearly equal-mass galaxies in which case the binding energy increases in proportion to the mass, giving $\Sigma_0 \propto M^{-1}$. An appropriate projection of the fundamental plane of elliptical galaxies to get a relation between Σ_0 and M suggests that the actual

scaling is roughly midway between the extreme models with $\Sigma_0 \propto M^{-2}$. Although this certainly does not *prove* the merger history of elliptical galaxies, it is consistent with the conjecture that elliptical galaxies have undergone a significant amount of mergers.

It is also possible to obtain simple scaling relations for the growth of total luminosity of a merging galaxy within a given aperture scale of observation r. If r_0 is small so that $r_0 \lesssim r$, we observe the full luminosity of the galaxy and $L \propto \Sigma_0 r_0^2 \propto M$ in both the extreme cases. However, at late times, r_0 is large so that $L \propto \Sigma_0 r^2$ (where r is the chosen scale of observation), which increases as M^{-1} or M^{-3} in the two extremes. Thus, for observations at fixed aperture, the magnitude of a cannibal galaxy can be very nearly constant for large parts of its history.

At a still smaller scale, mergers affect the globular clusters in a galaxy. The globular clusters that are distributed around the galaxy in an approximately spherical manner will experience dynamical friction when they move through the halo. If we take the density distribution in the halo to be $\rho(r) = (v_c^2/4\pi G r^2)$, then the general formula derived in Vol. II, Chap. 10, Section 10.5, will give the frictional force on a cluster of mass M moving with speed v_c at a radius r as

$$F = -\frac{4\pi(\ln\Lambda)G^2M^2\rho(r)}{v_c^2}\left[\mathrm{erf}(1) - \frac{2}{\sqrt{\pi}}e^{-1}\right] = -0.428(\ln\Lambda)\frac{GM^2}{r^2}. \tag{2.261}$$

(The Coloumb logarithm is $\ln\Lambda \simeq 10$ in this case.) Because the frictional force is tangential, it causes the cluster to lose the angular momentum (per unit mass) L at the rate

$$\dot{L} = \frac{Fr}{M}. \tag{2.262}$$

Assuming that the cluster continues to orbit at speed v_c as it spirals to the centre, we can take $\dot{L}$ to be equal to $v_c\dot{r}$. This leads to

$$r\frac{dr}{dt} = -0.428\frac{GM}{v_c}\ln\Lambda, \tag{2.263}$$

which can be easily integrated to find the time for the cluster to fall to the centre from some initial radius r_i; we get

$$t_{\mathrm{fric}} = \frac{1.17r_i^2v_c}{(\ln\Lambda)GM} = \frac{2.64\times10^{11}}{\ln\Lambda}\left(\frac{r_i}{2\ \mathrm{kpc}}\right)^2\left(\frac{v_c}{250\ \mathrm{km\ s^{-1}}}\right)\left(\frac{10^6\,M_\odot}{M}\right)\mathrm{yr}. \tag{2.264}$$

As an example, consider the globular clusters around M31 that have masses of approximately $M = 5\times10^6\,M_\odot$ and a circular speed of $v_c \approx 250$ km s^{-1}. Equation (2.264) shows that if $\ln\Lambda = 10$, massive clusters whose original orbits were at $r_i < 3$ kpc would have already spiralled into the centre of the galaxy. In other words, centres of galaxies like M31 could have accreted mass through the gobbling up of such clusters.

Another application of dynamical friction, in the context of interacting systems, can be illustrated by the Magellanic cloud, the centre of mass of which is orbitting our galaxy, at a distance of $r \approx 60$ kpc. Taking $M \approx 2 \times 10^{10}\, M_\odot$, we find that Eq. (2.164) gives

$$t_{\text{fric}} = \frac{1.0 \times 10^{10}}{\ln \Lambda} \left(\frac{r}{60\ \text{kpc}}\right)^2 \left(\frac{v_c}{220\ \text{km s}^{-1}}\right) \left(\frac{2 \times 10^{10}\, M_\odot}{M}\right) \text{yr}. \quad (2.265)$$

For a Magellanic cloud, $\ln \Lambda \simeq 3$, and this relation suggests that the clouds will spiral to the centre of our galaxy in a fraction of Hubble time. There are two caveats to this conclusion that could provide an escape from this conclusion: (1) The argument assumes that the halo of our galaxy reaches up to a large Magellanic cloud (LMC); current observations neither confirm nor disprove this assumption strongly, but if the halo density falls more rapidly than r^{-2} near the location of the LMC, the time scale will increase. (2) The clouds were assumed to be on a circular orbit, which may not be necessarily true. If the orbit is eccentric then the time scale will depend on the location on the orbit and could be smaller or larger than our estimate, depending on whether the clouds are currently near the apocentre or the pericentre. Determining the orbital motion requires knowledge of the proper motions of the clouds, which are not well determined.

2.9.2 Galaxy Collisions

It was previously mentioned that the analysis of the energy transfer in the encounter between two galaxies simplifies in two limiting cases. In the first case, we assume that the effect of a perturber on a system can be calculated by studying the tidal effect in a Taylor series expansion in R/b, where R is the size of the system and b is the impact parameter. This analysis was performed in Vol. II, Chap. 10, Section 10.7, where it was shown that energy transfer varies as $\Delta E \propto b^{-4}$. In the case of galaxies, this acts as an effect that tidally transfers energy.

More dramatic effects occur in the case of penetrating encounters corresponding to $b = 0$. The previous approach is obviously invalid in this case, but we can obtain the necessary results along the following lines. Let us choose the coordinates such that the galaxy is centred at the origin ($R = 0, z = 0$) of the cylindrical coordinate system and the trajectory of the perturber is along the z axis given by $[R = 0, z = Z_p(t) = Vt]$. If the perturber can be treated as a spherical system with a potential $\Phi(r)$, the R component of the perturber's gravitational field at (R, z) is $(-R/r)(d\Phi/dr)$, where $r = \sqrt{(Z_p - z)^2 + R^2}$. This is the only component that is relevant because, by symmetry, the velocity increment $\Delta \mathbf{v}$ is generated towards only the perturber's line of motion. We may neglect any changes in the velocity parallel to the perturber's trajectory as equal and opposite impulses occur during the approaching and the receding phases of the perturber.

The magnitude of the velocity change now becomes

$$\Delta v_R = -\int_{-\infty}^{\infty} \frac{d\Phi}{dr}\frac{R}{r}\,dt = -\frac{2R}{V}\int_0^{\infty} \frac{d\Phi}{dr}\frac{dZ_p}{r}, \tag{2.266}$$

corresponding to a net gain of energy:

$$\Delta E = \pi \int_0^{\infty} [\Delta v_R(R)]^2\, \Sigma(R) R\, dR. \tag{2.267}$$

For example, if both the galaxies are described by a Plummer model (see Subsection 2.2.1) with $\Phi = -GM/\sqrt{r^2 + a^2}$, then the net energy gained is $\Delta E = (G^2M^3/3V^2a^2)$. This complements the result for distant tidal encounters, $\Delta E = (4G^2M_2^2M_1\langle r^2\rangle/3b^4V^2)$. In fact, any other encounter can be treated as a smooth interpolation between these two.

One of the results of a head-on impact of galaxies is the formation of ring-like structures seen in several disk galaxies. Consider two galaxies modelled as singular isothermal spheres with circular velocities v_c that collide with each other with a speed $V \gg v_c$. The change in the radial velocity of any given star located at a distance R from the line of motion is now given by Eq. (2.266) with $\Phi = v_c^2 \ln r$, so that

$$\Delta v_R = -2R\frac{v_c^2}{V}\int_0^{\infty} \frac{dZ_p}{R^2 + Z_p^2} = -\frac{\pi v_c^2}{V}. \tag{2.268}$$

We now estimate the change in the trajectory of a star located in the stellar disk that is embedded in the target system. Taking the intruder's trajectory to be normal to the plane of the disk, we can calculate the change in trajectory by using the epicyclic approximation. Originally, the stars at R_0 execute small harmonic radial oscillations with frequency $\kappa(R_0) = \sqrt{2}(v_c/R_0)$. Taking the position of the star to be at $R(t, R_0)$ at time t, we must have $R(t, R_0) = R_0 - a\sin(\kappa t)$, where the amplitude a has to be fixed such that the velocity gained at $t = 0$ is given by expression (2.268). This gives

$$R(R_0, t) = R_0\left[1 - \frac{\pi v_c}{\sqrt{2}\,V}\sin\left(\frac{\sqrt{2}\,v_c t}{R_0}\right)\right]. \tag{2.269}$$

Figure 2.16 shows a plot of $R(R_0, t)$ in which the caustic structure and the enhancement of density along the lines of constant slope are apparent. This, in turn, suggests the propagation of a ring of material away from the galaxy at nearly constant speed.

If the unperturbed disk had a surface density $\Sigma_0(R_0)$, the new surface density is given by

$$\Sigma(R) = \Sigma_0(R_0)\left[\frac{R}{R_0}\left|\frac{\partial R}{\partial R_0}\right|\right]^{-1} = F(\tau)\Sigma_0(R_0), \tag{2.270}$$

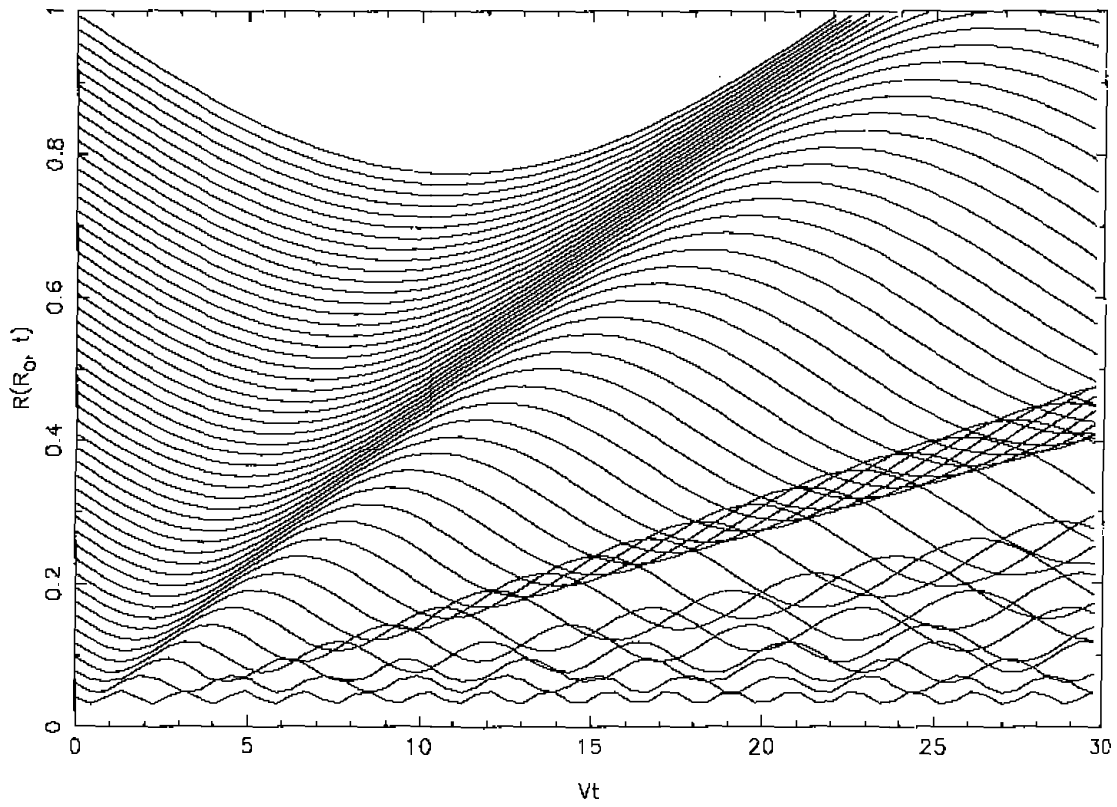


Fig. 2.16. Perturbation in a penetrating encounter that leads to ring galaxies.

where

$$\tau \equiv \frac{\sqrt{2}v_c t}{R_0}, \quad F(\tau) \equiv \left(1 - \frac{\pi v_c}{\sqrt{2}V}\sin\tau\right)^{-1} \left|1 - \frac{\pi v_c}{\sqrt{2}V}(\sin\tau - \tau\cos\tau)\right|^{-1}. \tag{2.271}$$

We note that the new surface density is amplified by a factor $F(\tau)$ that depends on the ratio (t/R_0). This amplification has distinct maxima; if, for example, $(v_c/V) \approx 0.1$, then the maximum occurs at $\tau_m \approx 3.1$. At any given time t, the density is sharply amplified around the radius $R_m(t) = \sqrt{2}(v_c t/\tau_m)$, thereby producing a ringlike enhancement. Numerical simulations show that this result is fairly robust and is valid even when the perturber's trajectory is slightly inclined to the plane of the disk.

2.9.3 Numerical Simulations

The interaction and merger of galaxies have been traditionally studied by numerical simulations for several decades now. Although the quality of simulations and the level of details as regards input physics have increased significantly, clear quantitative results that allow us to classify the interactions in some systematic manner do not exist. We shall discuss some of the features of numerical simulations of galaxy encounters in this subsection.

If stars are treated as rigid hard spheres of small radius, interacting only through Newtonian gravity, then setting up a numerically feasible algorithm to study the encounter between two galaxies is relatively straightforward. However, the parameter space that needs to be explored in any such simulation is fairly large. To begin with, we need to specify the distribution function $f_i(\mathbf{x}, \mathbf{v}, 0)$, $i = 1, 2$, at $t = 0$ for the two galaxies. Choosing the coordinates and the Lorentz frame such that the centre of mass of the first galaxy is at rest at the origin at $t = 0$, we can describe the initial conditions for the centre of the second galaxy in terms of the position and velocity $(\mathbf{X_0}, \mathbf{V_0})$. Given these parameters, the integration of the equations of motion will allow the study of future evolution.

The simplest situation will correspond to spherically symmetric distributions of matter in the two galaxies, which are identical. In this case, there is no preferred orientation (for example, spin axis) for the individual galaxies, and we can choose the initial separation vector between the two centres of mass to be along the x axis; the initial velocity of the second galaxy can be chosen to lie in the x–y plane. The centre of masses will move broadly along a Kepler-like orbit, with tidal perturbations affecting the individual movement of the stars. If the two galaxies are gravitationally bound, then the transfer of energy contained in the centre-of-mass motion into the internal kinetic energy of the stars could eventually lead to a merger of the galaxies with the ejection of high-energy stars in the form of a halo. Even in this case, the situation can become complicated at late times, and a diverse class of results can be obtained if the radial distributions of matter in the two galaxies are not identical.

More parameters enter the discussion when the galaxies are not spherical. An extreme case will involve the interaction between two galaxies that are infinitesimally thin disks at $t = 0$. It is again convenient to keep the centre of mass of the first galaxy at the origin and choose a frame in which the first galaxy is at rest at $t = 0$. We can choose the x axis along the direction of the centre of mass of the second disk galaxy at $t = 0$. The second galaxy now has a distinct initial velocity as well as a direction of internal orbital angular momentum. One of these, say, the initial velocity vector, could be taken to lie in the x–y plane. This leaves two more angular coordinates that denote the direction of initial angular momentum of the second galaxy as arbitrary, and the results of the encounter, of course, will depend on this parameter.

Given the amount of arbitrariness in the parameters specifying the initial condition, it is not surprising that a wide variety of morphological features emerge in these simulations. Although these are visually striking and often bear a good resemblence to actual galactic systems observed in the sky, they are not very useful in the study of actual physical processes. In fact, most of the morphological features that emerge in these simulations are purely kinematic and can be reproduced in a fairly simple analysis that does not require genuine N-body simulations. We first describe this approach and some of its results in order to

illustrate the importance of kinematics (or, more precisely, "reduced dynamics") in this particular context.

We model the encounter between two disk galaxies that were infinitesimally thin at $t = 0$ along the following lines: (1) All the mass in either of the galaxies will be taken to reside at the centres of the respective galaxies, with the masses being m_1 and m_2. (2) In each of the galaxies, we will distribute stars in circular Keplerian orbits at different radii in the range $r_{\min} < r < r_{\max}$, with the minimum and the maximum radii as well as the total number of stars being arbitrary for the two galaxies. Within each galaxy the initial distribution is uniform in the sense that the radii of the rings are uniformly spaced between $r_{\min}$ and $r_{\max}$ and the stars at any given radii are spaced uniformly on the circle. (3) The stars are treated as test particles and are affected by only the masses of the galaxies; that is, the stars do not interact with each other gravitationally. In such a situation, the dynamics can be separated into two parts. The two galaxies move in a Keplerian orbit under their mutual gravitational attraction. Any given star moves in the time-dependent potential generated by the two point masses, m_1 and m_2. The motion of each star is thus equivalent to a reduced version of the three-body problem. A simple numerical code that uses leapfrog integration can now generate wide varieties of morphology for different initial conditions.[7]

As the first example, let us consider the kinematic encounter between two spiral galaxies with the following initial conditions: (1) Both the galaxies are taken to be zero-thickness disks and coplanar with the stars rotating in circular Keplerian orbits around central masses. (2) The target galaxy has the mass of 5 units (one unit of mass corresponds to $2 \times 10^{10}\ M_\odot$ in these simulations; hence the actual mass is $10^{11}\ M_\odot$), and the second, the perturbing galaxy, has a mass that is one quarter of that of the target galaxy. (3) There are 1,000 stars distributed equally among 20 rings with radii varying between 10 and 30 units in the target galaxy; there are 250 stars distributed equally among 5 rings varying in radius between 2 and 7 units in the second galaxy. (The unit of distance is chosen such that 1 unit corresponds to 0.5 kpc.) The stars in both the galaxies are rotating in a counterclockwise direction. (4) The target galaxy is located at the origin with zero velocity at $t = 0$, and the perturber has the position (30, −30, 0) and velocity (0, 0.34, 0.34) in the chosen units. (The physical velocity corresponding to 1 unit is 400 km s^{-1}.) Figure 2.17 shows the configuration we obtain by integrating essentially a set of time-dependent three-body problems. The unit of time corresponds to 1.2×10^6 yr. (The units of time, length, and mass are chosen such that $G = 1$ in the numerical code.) It is clear that purely kinematic perturbation – which does not involve any self-gravity of the stars – can lead to striking spiral patterns similar to, for example, the ones seen in the whirpool galaxy.

In the preceding example, the angular momentum of the Kepler motion of the galaxies is in the same direction as the spin angular momentum of the individual stars in the galaxies (called *prograde motion*). This leads to a resonancelike

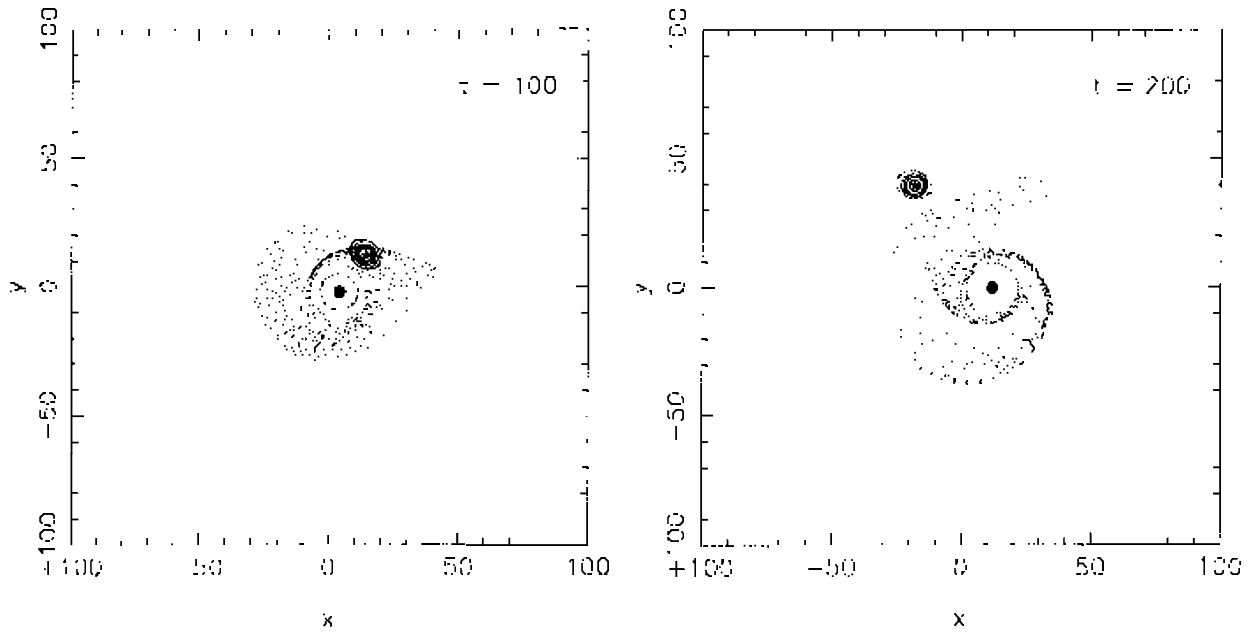


Fig. 2.17. Kinematic simulation of galaxy encounters that lead to a configuration similar to that of a whirpool galaxy.

condition, making the perturbing effect somewhat strong. To see this effect more clearly, we compare it with the retrograde initial condition in which the orbital angular momentum is in a direction opposite to the spin angular momentum. This effect is illustrated more dramatically in Fig. 2.18, in which the initial conditions differ only in the case of the position of the perturbing galaxy. In this case, we have taken the perturbing galaxy to have the same mass as that of the target galaxy

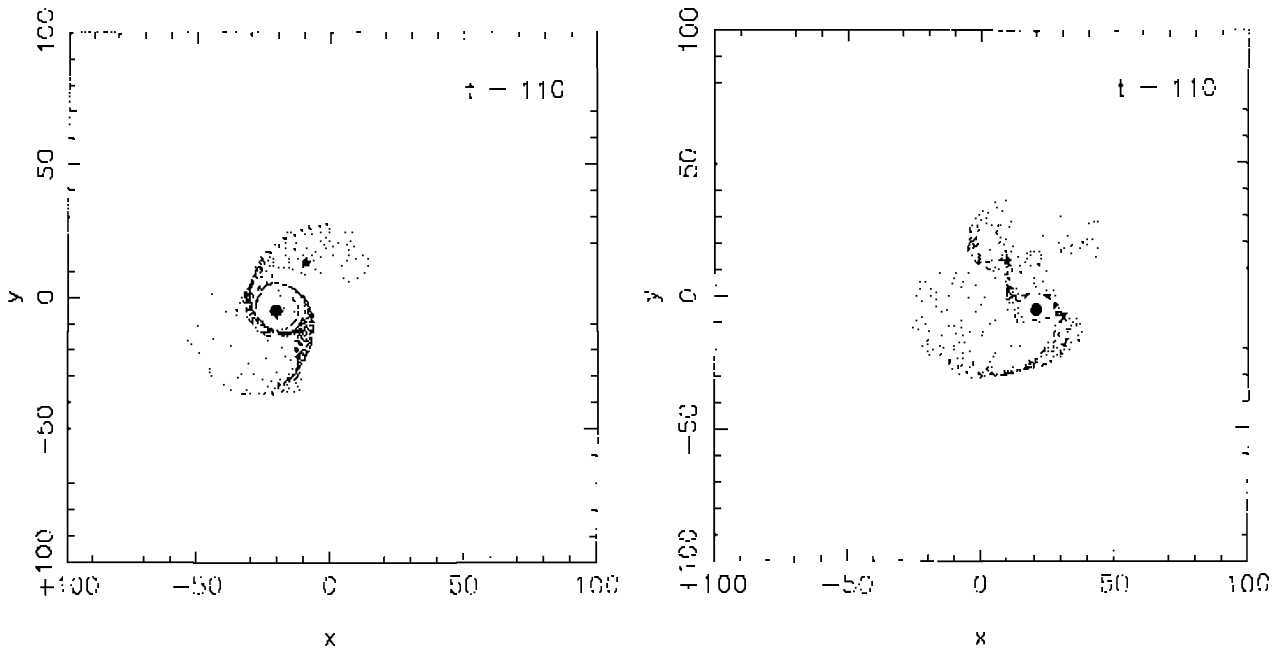


Fig. 2.18. Kinematic simulations of galactic encounters of two equally massive galaxies in retrograde (left) and prograde (right) motions. The stars in the perturber are not shown for simplicity. The retrograde initial conditions in the left frame lead to somewhat less disruption.

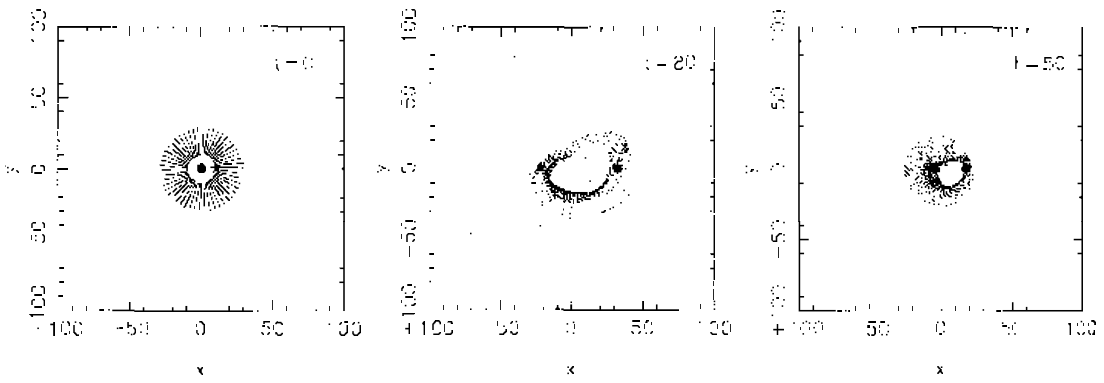


Fig. 2.19. Kinematic simulation of a plunging galaxy in a near-head-on encounter.

(compared with the previous case shown in Fig. 2.17, in which the perturber had a mass that was 0.25 times that of the target).

Finally, Fig. 2.19 illustrates, through kinematic simulations, the effect that was previously discussed. In this case, the target galaxy has exactly the same parameters as those in the previous case whereas the perturber has the initial position of (10, 0, 35) and initial velocity of (0, 0, −1) so that it can plunge through the target galaxy. The masses of the galaxies are taken to be equal. It is clear from the figure that a ringlike structure can form and propagate outwards.