Eur. J. Phys. 2 (1981) 91–102. Printed in Northern Ireland

Regularity and chaos in classical mechanics, illustrated by three deformations of a circular 'billiard'

M V Berry†

Institute for Theoretical Physics, University of Utrecht, The Netherlands

Received 13 April 1981

Abstract The motion of particles bouncing inside a curve B is employed to illustrate different types of orbit in classical mechanics. In the 'phase space' whose two coordinates are s, the position around B, and p, the direction of impact, orbital dynamics is a discrete area-preserving mapping between successive bounces; an orbit may be zero dimensional (i.e. closed, and so returning repeatedly to a finite number of points), one dimensional (eventually filling an 'invariant curve') or two dimensional (eventually filling an area chaotically). When B is a circle, s, p space is covered with invariant curves and no closed orbits are isolated. Different deformations of a circle generate very different orbits: stadia give ergodic motion (almost all orbits explore almost all s, p values) with extreme unpredictability (chaos), ellipses give motion entirely confined to invariant curves whose topology is organised by two isolated closed orbits, a family of ovals gives (generic) motion in which phase space is intricately divided into chaotic areas and areas covered with invariant curves. The nature of the motion is determined by whether the closed orbits are stable, unstable or neutrally stable.

Résumé On utilise le mouvement de particules rebondissant à l'intérieur du domaine limité par une courbe fermée B pour illustrer différents types d'orbites telles que les décrit la mécanique classique.

Dans 'l'espace des phases', associé ici aux deux coordonnées s, la position le long de B et p, le direction d'impact, la dynamique orbitale définit une transformation discrète, conservant les aires, entre rebonds successifs; une orbite peut être de dimension nulle (orbite fermée repassant indéfiniment par un nombre fini de points), de dimension un, (associée à une 'courbe invariante'), ou de dimension deux, (associée à une aire dont tous les points sont visités aléatoirement au cours du mouvement).

Quand B est un cercle, l'espace (s, p) est couvert par l'ensemble des courbes invariantes, et il n'apparaît pas d'orbite fermée isolée. Des déformations diverses du cercle B conduisent à des orbites elles-mêmes très diverses: les courbes 'en stade' donnent lieu à des mouvements ergodiques, (presque toutes les orbites explorent presque toutes les valeurs des coordonnées s et p), d'un caractère très chaotique et imprévisible; les ellipses donnent lieu à des mouvements limités à des courbes invariantes dont la topologie est définie par deux orbites fermées isolées; une famille d'ovales conduit à des mouvements pour lesquels l'espace des phases se répartit de manière complexe en zones couvertes par des courbes invariantes et zones correspondant à des orbites bidimensionnelles. La nature du mouvement est déterminée par le caractère stable, instable ou indifférent des orbites fermées.

1. Introduction

In recent years the study of classical mechanics has revived, and major advances are being made. These concern the qualitative behaviour of systems over long times: sometimes motion is predictable in the sense that slight changes in initial conditions result in only slightly different motion, sometimes motion is unpredictable in the sense that slight changes in initial conditions result in radically different motion, and sometimes both sorts of motion can coexist in the same system for different initial condi-

† Permanent address: H H Wills Physics Laboratory, Tyndall Avenue, Bristol BS8 1TL, England.

0143-0807/81/020091+12$01.50

tions. Although the 'new mechanics' bears on important issues as diverse as the long-term stability of the solar system and the degree to which Newtonian mechanics is really deterministic, its principles are not widely known. My purpose here is to illustrate them with an example simple enough to be presented at the undergraduate level whilst possessing all the features of the general case.

The example is the 'billiard problem' of a point particle moving freely in the region of the plane bounded by a closed curve B (the 'billiard') and being reflected elastically at impacts with B, according to the law: 'the angle of reflection equals the angle of incidence' (figure 1). At any instant the particle's state is determined solely by its position and direction of motion, because elasticity of collision implies conservation of energy and hence speed; for this simple system, dynamics is elementary geometry. Natural questions are: after many bounces, has the particle visited the neighbourhood of every point within B? Has it travelled in almost every direction? These are questions of ergodic theory, most familiar in the context of many-particle systems, where the assumption that all configurations and momenta, compatible with the total energy, are eventually explored lies at the foundation of statistical mechanics. For billiards which have only two degrees of freedom, the answers will depend delicately on the shape of B.

A brief guide to the extensive literature on modern mechanics, and billiards in particular, is given in §8.

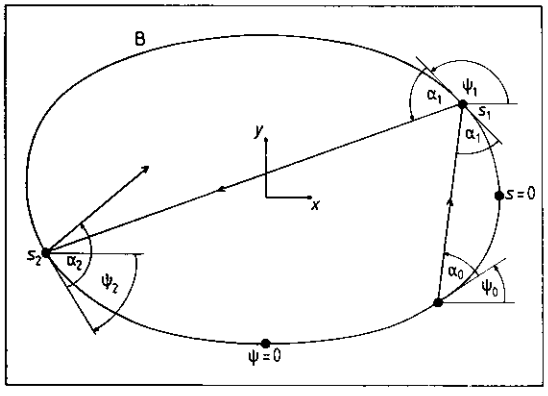

Figure 1 Billiard geometry and coordinates.

2. Billiard mapping

Between impacts with B, the particle moves in straight lines. An orbit may therefore be completely specified by giving the sequence of its positions and directions immediately after each impact. The position round B will be parametrised either by arc length s or by the direction ψ of the 'forward' (i.e. anticlockwise-pointing) tangent, measured from the origins shown in figure 1. The angle ψ gives a unique parametrisation of B provided B has no inflections, and this will be assumed henceforth. To relate the s and ψ parameters, B can be defined by giving its radius of curvature R as a function of ψ. Then

$$R(\psi)=\frac{\mathrm{d}s}{\mathrm{d}\psi} \qquad \text{i.e.} \qquad s(\psi)=\int_{\pi/2}^{\psi}\mathrm{d}\psi' R(\psi'). \quad (1)$$

The direction of the orbit after impact will be labelled by its angle α with respect to the forward tangent (figure 1), or by the tangential momentum p, defined by

$$p\equiv\cos\alpha. \quad (2)$$

For calculating orbits the ψ, α description is more convenient, but for theoretical purposes the s, p description is preferable.

An orbit, then, consists of the succession of number pairs $\{s_n, p_n\}$ corresponding to the nth bounce, and is generated by specifying an initial state s_0, p_0 (figure 1). This discrete dynamics is a *mapping* M of the 'phase space' with coordinates s, p and is symbolised by

$$\begin{pmatrix} s_{n+1} \\ p_{n+1} \end{pmatrix} = M \begin{pmatrix} s_n \\ p_n \end{pmatrix}. \quad (3)$$

(The 'bounce mapping' is usually non-linear and so M cannot usually be represented by a 2×2 matrix.) If L is the length of B, phase space can be restricted to the rectangle $-1\leq p\leq +1$, $0\leq s<L$, but since s is a periodic coordinate ($s+L$ is equivalent to s) its true topology is that of a circular ribbon. In terms of the variables s, p (but not ψ, α) M is *area preserving*, i.e.

$$\frac{\partial(s_1, p_1)}{\partial(s_0, p_0)} = \det\begin{Bmatrix} \partial s_1/\partial s_0 & \partial s_1/\partial p_0 \\ \partial p_1/\partial s_0 & \partial p_1/\partial p_0 \end{Bmatrix} = 1 \quad (4)$$

this is proved in appendix 1.

There are three ways in which the orbit generated by infinitely many iterations of M can be explored in phase space.

(i) A finite set of N points s_0, p_0; s_1, p_1; ...; s_{N-1}, p_{N-1} may be encountered repeatedly, corresponding to orbits that *close* after N bounces. Symbolically, such a closed orbit satisfies

$$\begin{pmatrix} s_{n+N} \\ p_{n+N} \end{pmatrix} = M^N \begin{pmatrix} s_n \\ p_n \end{pmatrix} = \begin{pmatrix} s_n \\ p_n \end{pmatrix} \quad (5)$$

so that each of its N points is a *fixed point* of the mapping M^N.

(ii) The iterates of s_0, p_0 may fill a smooth curve in phase space, called an *invariant curve* because the whole curve maps onto itself under M (although its individual points do not map onto themselves). This behaviour occurs, for example, if the dynamics is

integrable in the sense that there exists a constant of motion in the form of a function $F(s, p)$ satisfying

$$F(s_1, p_1) = F(s_0, p_0) \tag{6}$$

in which case the invariant curves are the contours of $F(s, p)$.

(iii) The iterates of s_0, p_0 may fill an area in phase space. This happens when the orbit, unrestricted by the existence of any conserved quantity, evolves in a *chaotic* manner whose detail is sensitively dependent on the values of s_0 and p_0.

All three types of orbit will be encountered in the billiards to be considered later.

In terms of ψ, α, the mapping equations can be found with reference to figure 1 as follows. The slope of the trajectory segment beginning at ψ_0, α_0 is given by the quotient of the x and y increments around the curve between ψ_0 and ψ_1. These increments can be found using

$$\mathrm{d}x/\mathrm{d}s = \cos\psi \qquad \mathrm{d}y/\mathrm{d}s = \sin\psi \tag{7}$$

which together with (1) give

$$\begin{aligned} x(\psi_1) - x(\psi_0) &= \int \cos\psi \,\mathrm{d}s = \int \cos\psi \frac{\mathrm{d}s}{\mathrm{d}\psi}\,\mathrm{d}\psi \\ &= \int_{\psi_0}^{\psi_1} R(\psi)\cos\psi \,\mathrm{d}\psi \qquad (8) \\ y(\psi_1) - y(\psi_0) &= \int_{\psi_0}^{\psi_1} R(\psi)\sin\psi \,\mathrm{d}\psi. \end{aligned}$$

The slope is then

$$\left(\int_{\psi_0}^{\psi_1} R(\psi)\sin\psi \,\mathrm{d}\psi\right)\left(\int_{\psi_0}^{\psi_1} R(\psi)\cos\psi \,\mathrm{d}\psi\right)^{-1} = \tan(\psi_0 + \alpha_0) \tag{9}$$

and this equation determines ψ_1 (and hence s_1) given ψ_0 (or s_0) and α_0 (or p_0). α_1 (and hence p_1) is now determined by another slope relationship, namely

$$\psi_1 - \alpha_1 = \psi_0 + \alpha_0 \qquad \text{i.e.} \qquad \alpha_1 = \psi_1 - \psi_0 - \alpha_0. \tag{10}$$

These two mapping equations are well suited to rapid computer iteration. Conversion to the variables s, p is trivial using (1) and (2).

3. Stability of closed orbits

In what follows an important role will be played by the closed orbits, which satisfy (5). These may be stable or unstable in the sense that an orbit starting at $s_0 + \delta s_0$, $p_0 + \delta p_0$, where δs_0 and δp_0 are small, may after many bounces remain near the closed orbit or may deviate increasingly from it. After N iterations, when s_0 and p_0 have returned to their initial values, the deviations δs_N and δp_N of the nearby orbit will be

$$\begin{pmatrix} \delta s_N \\ \delta p_N \end{pmatrix} = m_N \begin{pmatrix} \delta s_0 \\ \delta p_0 \end{pmatrix} \tag{11}$$

where m_N is a 2×2 matrix with unit determinant whose precise form for billiard mappings is given in appendix 1.

Orbital stability depends on the eigenvalues of m_N. These are $\lambda_\pm$, given in terms of the trace of m_N by

$$\lambda_\pm = \tfrac{1}{2}\{\mathrm{Tr}\, m_N \pm [(\mathrm{Tr}\, m_N)^2 - 4]^{1/2}\}. \tag{12}$$

After j traversals of the closed orbit (i.e. Nj iterations of M), the deviations $(\delta s_{Nj}, \delta p_{Nj})$ can be written as a linear combination of $\lambda_\pm^j$ times eigenvectors of m_N, i.e.

$$\begin{pmatrix} \delta s_{Nj} \\ \delta p_{Nj} \end{pmatrix} = A\lambda_+^j \begin{pmatrix} \delta s_+ \\ \delta p_+ \end{pmatrix} + B\lambda_-^j \begin{pmatrix} \delta s_- \\ \delta p_- \end{pmatrix}. \tag{13}$$

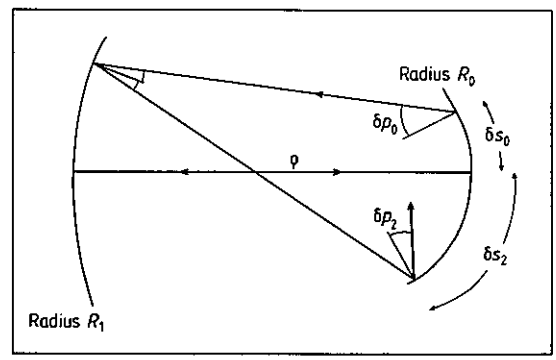

Figure 2 Deviation from the simplest closed orbit.

There are three possibilities. Firstly, if

$$|\mathrm{Tr}\, m_N| < 2 \qquad \text{(stable)} \tag{14}$$

it follows from (12) that $\lambda_\pm$ are complex conjugates on the unit circle, so that

$$\lambda_\pm^j = \mathrm{e}^{\pm \mathrm{i}j\beta} \tag{15}$$

where β is a 'stability angle'. In this case the deviations (13) oscillate about zero as j increases, and remain bounded, so that the orbit is *stable*. Secondly, if

$$|\mathrm{Tr}\, m_N| > 2 \qquad \text{(unstable)} \tag{16}$$

it follows from (12) that $\lambda_\pm$ are real and reciprocals of each other, so that

$$|\lambda_\pm|^j = c^{\pm j\gamma} \tag{17}$$

where γ is an 'instability exponent'. In this case the positive exponent guarantees that almost all deviations grow exponentially so that the orbit is *unstable*. And thirdly, in the exceptional case that

$$|\mathrm{Tr}\, m_N| = 2 \qquad \text{(neutral)} \tag{18}$$

both eigenvalues are +1 or −1 and a further analysis shows that the deviations grow linearly so that in this case the orbit has neutral stability.

The simplest example is the 'diametral' two-bounce orbit (figure 2) with impacts on opposite sides of B at normal incidence ($\alpha_0 = \alpha_1 = \frac{1}{2}\pi$, $p_0 = p_1 = 0$). If the radii of curvature R are the same at both impacts, and if the length of each trajectory segment is ρ, then it follows from general formulae in appendix 1 that the deviation matrix m_2 is

$$m_2 = \begin{Bmatrix} 2(\rho/R-1)^2-1 & 2\rho(1-\rho/R) \\ (2/R)(\rho/R-1)(2-\rho/R) & 2(\rho/R-1)^2-1 \end{Bmatrix}. \tag{19}$$

From (14) and (16), the stability conditions are

$$\frac{\rho}{2R} - 1 \begin{cases} >0 & \text{instability} \\ <0 & \text{stability.} \end{cases} \tag{20}$$

The same conditions follow in this special case from more elementary arguments involving the focusing or defocusing consequent upon repeated reflections between concave mirrors. Instability can have dramatic consequences, as we shall see in §5.

4. Circular billiards

When B is a circle, the radius of curvature $R(\psi)$ is independent of ψ. Elementary geometry, or trigonometry based on (9) and (10), shows that each orbit consists of a succession of chords (figure 3(*a*)) making equal angles α with B. Circular billiard motion is therefore restricted by the simple conservation law

$$p = \text{constant} \tag{21}$$

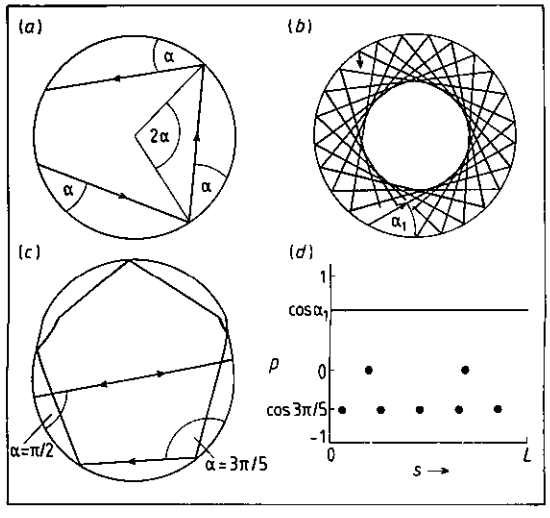

Figure 3 Billiards in a circle: (*a*) basic orbit geometry, (*b*) typical orbit (never closing), (*c*) two closed orbits, (*d*) phase space trajectories for orbits in (*b*) and (*c*).

the system is integrable and phase space s, p is covered with invariant curves parallel to the s direction. This excludes the third type of orbit discussed in §2, namely those filling an area. But as will now be explained, the other two types of orbit do occur.

A typical value of α will be an irrational submultiple of π, and generates an orbit that never repeats but continually hits B at different points s_n, eventually filling an annulus within B as shown for the orbit labelled α_1 in figure 3(*b*). In phase space the iterates s_n, p fill the invariant curve $p = \cos\alpha_1$ (figure 3(*d*)).

But if α is a rational submultiple of π, i.e.

$$\alpha = \pi K/N \qquad K, N \text{ mutually prime integers} \tag{22}$$

then the orbit closes after N bounces, as shown for $\alpha = 90°$ and $\alpha = 36°$ in figure 3(*c*). In the phase plane the iterates repeatedly return to N points on the line $p = \cos \pi K/N$ (figure 3(*d*)).

An important feature of these closed orbits is the fact that they are not isolated: a continuous family of new ones can be produced by rotation relative to B. In phase space the different N-point orbits are related by translation in the s direction, so that the complete family of closed orbits will fill the invariant curve $p = \text{constant}$. Associated with this non-isolation of the closed orbits in a circle is the fact that the orbits have *neutral stability* in the sense of equation (18). This is easily verified for the diametral orbit ($N = 2$) by using equation (20), because in a circle the separation ρ between the ends of a diameter is $2R$.

5. First deformation: stadia

In the stadium (figure 4), B consists of two semicircular arcs with radius R joined by tangential straight lines with length η. As $\eta \to 0$, the stadium

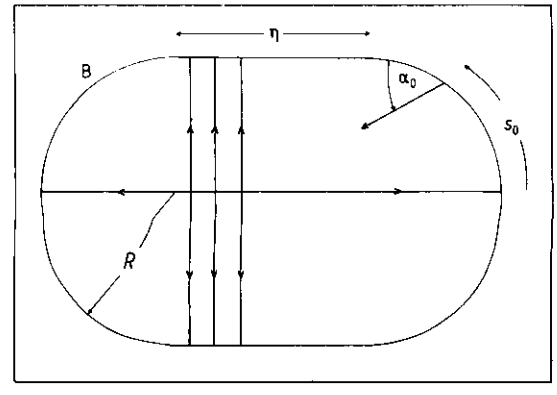

Figure 4 The stadium, with initial conditions s_0, α_0 for the orbit mapped on figure 5, a family of non-isolated two-bounce orbits and an isolated unstable two-bounce orbit.

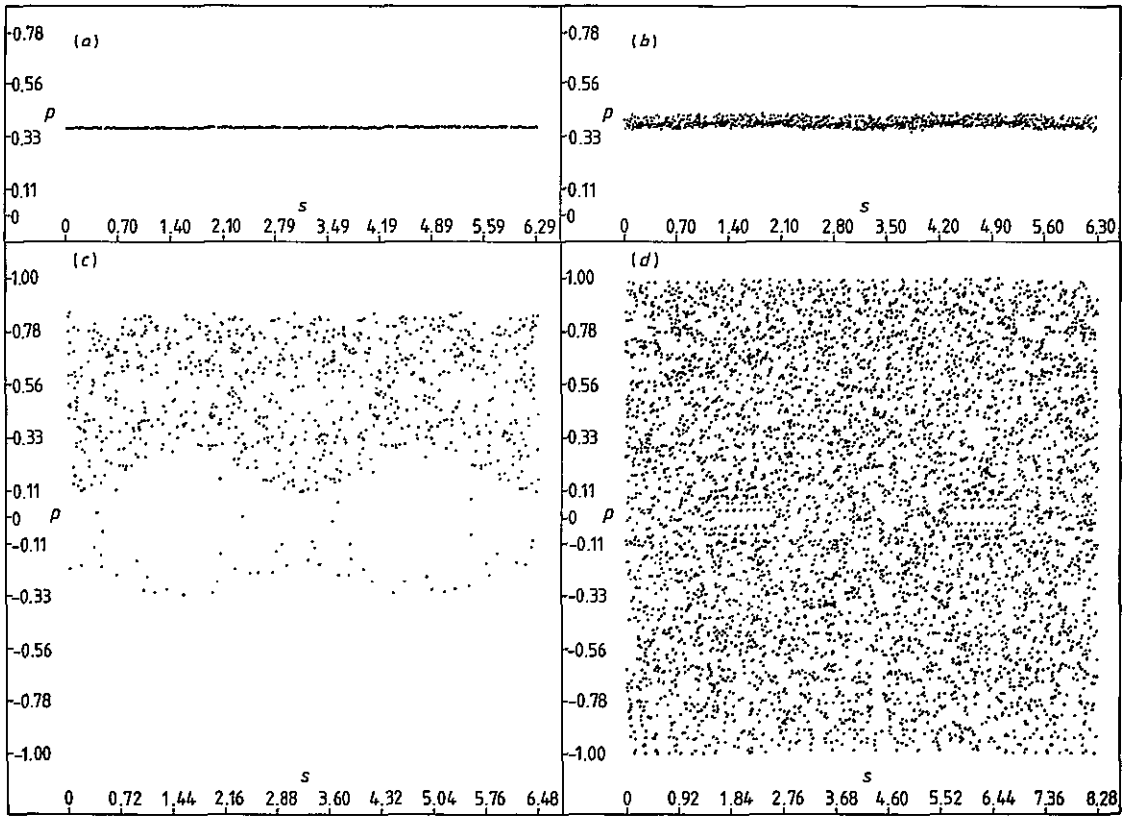

Figure 5 Stadium billiard mapping with $s_0 = 1$, $p_0 = 1/e$. $R = 1$ and (*a*) $\eta = 0.001$, 900 bounces, (*b*) $\eta = 0.01$, 900 bounces, (*c*) $\eta = 0.1$, 900 bounces, (*d*) $\eta = 1$, 5000 bounces.

degenerates into the circle, and it might be thought that for small η the mapping M would generate invariant curves similar to those for the circular billiard. In fact the opposite is the case, because it has been proved (see §8) that the stadium is an *ergodic* billiard, meaning that for almost every initial condition s_0, p_0, the iterates s_n, p_n will come arbitrarily close to *every* point in phase space as $n \to \infty$.

To illustrate this astonishing theorem, figure 5 shows iterates under the mapping M from the 'typical' initial conditions $s_0 = 1$, $p_0 = 1/e$ (i.e. $\alpha_0 = 68.4°$) (figure 4) for stadia with $R = 1$ and increasing values of η. Figure 5(*a*) shows 900 iterations for $\eta = 0.001$. A slight thickening of the line shows that points are beginning to deviate from the line $p = p_0$ which is an invariant curve when $\eta = 0$. Figure 5(*b*) shows 900 iterations for $\eta = 0.01$, and it is now obvious that the points are exploring an area rather than a curve. In figure 5(*c*), showing 900 iterations for $\eta = 0.1$, this area has expanded to include most of the phase space with $p > 0$ and some of the phase space with $p < 0$. Similar behaviour is expected, and observed, if η is held fixed and the orbit followed for more bounces.

It is very instructive to follow the process of area filling in more detail, by increasing the deformation to $\eta = 1$ and the number of iterations to 5000, thus generating the mapping of figure 5(*d*). It is clear that the points do fill the phase space uniformly except for two small 'holes' centred on $p = 0$, $s = \frac{1}{4}L$ and $p = 0$, $s = \frac{3}{4}L$. The existence of these holes (which can also be seen in figure 5(*c*)) does not violate the ergodic theorem, because they get smaller as $n \to \infty$ and eventually disappear, leaving the whole s, p plane filled with points.

The holes are connected with the family of two-bounce non-isolated closed orbits formed by perpendicular impacts on the straight sections of B (figure 4). In phase space this family forms two invariant curves in the form of straight line segments with $p = 0$ and $\frac{1}{2}\pi R < s < \frac{1}{2}\pi R + \eta$ and $\frac{3}{2}\pi R + \eta < s < \frac{3}{2}\pi R + 2\eta$. An orbit near one of these closed orbits (i.e. with small angular deviation δp_0), will 'resonate' in a zigzag path for many bounces before striking one of the semicircles and getting lost in the chaos. But according to the ergodic theorem such orbits must, on average, spend equal

times in equal areas of phase space. Therefore the existence of 'resonances' near the invariant curves implies that after leaving a resonance the orbit spends a long time (i.e. many bounces) avoiding the resonance region. Appendix 2 gives a quantitative theory of this effect. In particular, an orbit starting far from the invariant curves will probably avoid their neighbourhood for a long time. Several cycles of resonance and avoidance can be seen in figure 5(*d*) (the resonances—zigzag paths—are the lines of dots inside the holes). This phenomenon shows that ergodicity does not preclude strong position-dependent correlations between bounces.

Away from the holes, points s_n, p_n jump erratically and apparently randomly over phase space. To quantify this randomness we study the long diametral closed orbit in figure 4, which has $N=2$, $s_0=0$, $s_1=\frac{1}{2}L$, $p_0=p_1=0$. This is isolated, and it is unstable according to the criterion (20) because

$$\frac{\rho}{2R}-1=\frac{2R+\eta}{2R}-1=\frac{\eta}{2R}>0. \tag{23}$$

Now consider a mis-aimed orbit that starts out from $s_0=0$ but has a small angular deviation δp_0. According to (13), its evolution (diverging unstably away from the long diameter) depends on the larger eigenvalue λ. After $2j$ bounces the particle will be travelling in a direction

$$\delta p_{2j}\approx\lambda^j\delta p_0. \tag{24}$$

It is reasonable to claim that when the deviation reaches 1 rad the orbit has lost all memory of the closed orbit near which it began. This loss of memory occurs after n^* bounces where, from (24)

$$n^*=2\frac{\log(1/\delta p_0)}{\log\lambda}. \tag{25}$$

As a concrete example, let us choose a stadium with $\eta=R$ as in figure 5(*d*). Then (19) and (12) give

$$\mathrm{Tr}\,m_2=14 \qquad \lambda=7+\sqrt{48}=13.93. \tag{26}$$

In a typical computer about 14 digits can be stored, so let us take $\delta p_0=10^{-14}$. Then (25) gives

$$n^*=19.58 \tag{27}$$

so that in spite of the careful aim the particle bounces irrecoverably away from the closed orbit after only 20 impacts. Moreover each increase of one decimal digit in the precision with which δp_0 is specified will only enable the orbit to be predicted for about two extra bounces.

In such systems the extreme natural instability must therefore soon outstrip the precision available in any computer (unless special programming techniques are employed to increase the number of stored digits in proportion to n). For pictures such as figure 5(*d*), involving thousands of bounces, this implies that the individual points bear no relation to those that would have been generated by the exact solution of the billiard problem from s_0, p_0. Strangely enough, this does not mean that the computations are worthless, because it has also been proved that there exist initial conditions close to s_0 p_0 whose exact orbits lie arbitrarily close to the approximate one computed from s_0 p_0.

What this example shows is that the existence of causal dynamics in which initial conditions determine the trajectory of systems for all time is consistent with instabilities preventing any particular trajectory from being calculated, even approximately, by any practical (or, in a finite universe, conceivable) means.

6. Second deformation: ellipses

Quite different orbits are generated by deforming the circular billiard into ellipses rather than stadia. In terms of a parameter λ, the equations

$$\begin{aligned} x&=a\cosh M\cos\lambda \\ y&=a\sinh M\sin\lambda \end{aligned} \tag{28}$$

determine an ellipse whose eccentricity ϵ is

$$\epsilon=(\cosh^2 M)^{-1} \tag{29}$$

and whose foci lie at $x=\pm a$, $y=0$. λ is related to the direction parameter ψ (figure 1) by

$$\tan\psi=\frac{dy}{dx}=\frac{dy/d\lambda}{dx/d\lambda}=-\tanh M\cot\lambda. \tag{30}$$

The radius of curvature is

$$R(\psi)=\frac{a\cosh M\sinh M}{(\cosh^2 M\sin^2\psi+\sinh^2 M\cos^2\psi)^{3/2}}. \tag{31}$$

Billiard motion in an ellipse is integrable: no matter what the value of the eccentricity ϵ is, there exists a constant of motion $F(s,p)$ restricting the orbits to invariant curves in s, p space. This contrasts with the stadium, which for arbitrarily small deformations η generated ergodic motion. The existence of the conserved quantity depends on a geometric fact whose proof is elementary but tedious: each orbit will repeatedly touch a conic confocal with B; an ellipse is touched if s_0, p_0 is such that the first segment of the orbit does not pass between the foci, and a hyperbola is touched if the segment does pass between the foci. The two types of orbit are illustrated in figure 6. Confocal conics are obtained by considering both M and λ as parameters in (28): varying λ for fixed M generates an ellipse, and varying M for fixed λ generates a hyperbola. The initial condition s_0, p_0 determines the parameter of the conic that is repeatedly touched. This parameter is the constant of motion; tedious algebra gives the explicit formula

$$F(s,p)=\frac{p^2-\epsilon^2\cos^2\psi(s)}{1-\epsilon^2\cos^2\psi(s)} \qquad 0<\epsilon<1 \quad (32)$$

where $\psi(s)$ is obtained from (1) and (31). (J H Hannay (private communication) points out that the constant of motion is simply interpreted as the product of angular momenta about the two foci.)

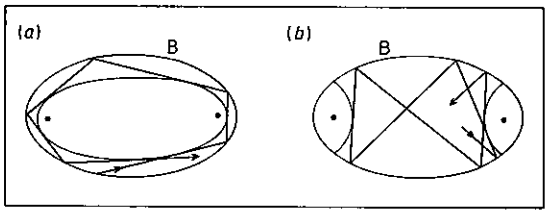

Figure 6 Orbits in an ellipse: (*a*) repeatedly touching a confocal ellipse, (*b*) repeatedly touching confocal hyperbolae.

Figure 7 shows some of the contours of $F(s,p)$; these are the invariant curves. There are two kinds of orbit. Firstly, for $|p|$ near to unity there are orbits like that in figure 6(*a*) which bounce all round B, exploring all values of s whilst repeatedly touching an ellipse. Secondly, for $|p|$ small and s close to $\frac{1}{4}L$ or $\frac{3}{4}L$ there are orbits like that in figure 6(*b*), which bounce across B, exploring a restricted range of s whilst repeatedly touching a hyperbola.

Along some of the invariant curves, motion will be periodic as in the case of the circular-billiard orbits with rational α (cf the dot orbits in figure 3(*d*)); these closed orbits are not isolated but form families filling their curves. Much more important, however, are two *isolated* closed orbits which dominate the topology of figure 7. These are the diametral two-bounce orbits. Firstly, there is the orbit along the long diameter, with $p=0$ and $s=0$ and $\frac{1}{2}L$. Applying criterion (20), and using (28) and (31), we have

$$\frac{\rho}{2R}=\frac{2x(\lambda=0)}{2R(\psi=\frac{1}{2}\pi)}=\frac{1}{\tanh^2\psi}>1 \qquad (33)$$

so that this orbit is unstable. It differs from its counterpart in the stadium (figure 4) because neighbouring orbits escape along smooth invariant curves (locally hyperbolic) rather than exploring a chaotic area. Secondly, there is the orbit along the short diameter, with $p=0$ and $s=\frac{1}{4}L$ and $\frac{3}{4}L$. Now (20) gives

$$\frac{\rho}{2R}=\frac{2y(\lambda=\frac{1}{2}\pi)}{2R(0)}=\tanh^2\psi<1 \qquad (34)$$

so that this orbit is stable, in contrast to its counterpart in the stadium (figure 4) which was neutrally stable and moreover non-isolated.

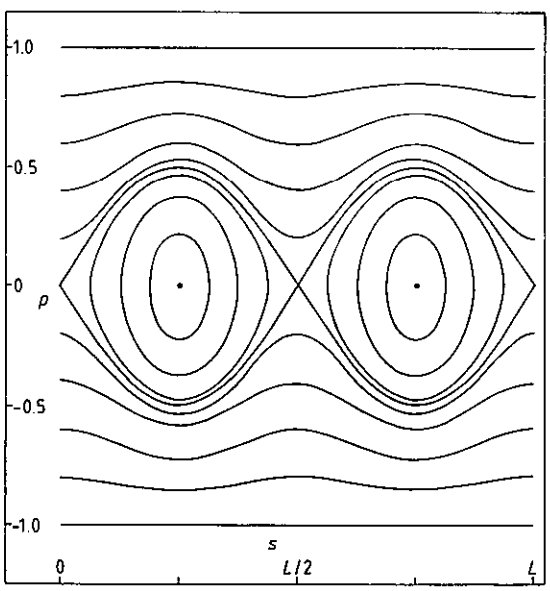

Figure 7 Ellipse billiard mapping.

7. Third deformation: ovals

So far we have encountered stable closed orbits, unstable closed orbits, marginally stable closed orbits, orbits covering smooth invariant curves and orbits filling areas chaotically. In the 'generic' case, that is for 'typical' boundaries B, all these different kinds of orbit co-exist. In this respect neither the circle nor the stadium nor the ellipse is generic. Now I shall describe a class of oval billiards that does display generic behaviour.

Recall from §2 the specification of B in terms of the function $R(\psi)$. For analysis and computation of billiard mappings this representation of curves is more convenient than customary representations (e.g. $\psi(s)$, $R(s)$, $x(s)$ and $y(s)$, $1/R$ as a function of s or ψ). R must be a periodic function chosen such that $x(\psi)$ and $y(\psi)$ are periodic too, so that B is closed. From (8), the condition for this is

$$\int_0^{2\pi} \mathrm{d}\psi R(\psi)\mathrm{e}^{\mathrm{i}\psi}=0. \qquad (35)$$

The Fourier expansion of $R(\psi)$ therefore begins with the terms involving 2ψ, and from this point of view the simplest deformation of a circle is

$$R(\psi)=a(1+\delta\cos 2\psi). \qquad (36)$$

From (8), the Cartesian coordinates are

$$\begin{aligned} x(\psi)&=a[(1+\tfrac{1}{2}\delta)\sin\psi+\tfrac{1}{6}\delta\sin 3\psi] \\ y(\psi)&=a[(-1+\tfrac{1}{2}\delta)\cos\psi-\tfrac{1}{6}\delta\cos 3\psi] \end{aligned} \qquad (37)$$

and from (1) the arc length is

$$s(\psi) = a(\psi - \tfrac{1}{2}\pi + \tfrac{1}{2}\delta \sin 2\psi) \qquad (38)$$

so that all the curves have length $L = s(2\pi) = 2\pi a$.

These equations describe a family of ovals parametrised by δ. If $0 < \delta < 1$ the ovals are smooth, with short diameter

$$2y(\pi) = a(1 - \tfrac{1}{3}\delta) \qquad (39)$$

and long diameter

$$2x(\tfrac{1}{2}\pi) = a(1 + \tfrac{1}{3}\delta) \qquad (40)$$

as shown in figures 8(*a*) and (*b*). If $\delta \geqslant 1$ the radius of curvature can vanish and the curve develops two 'swallowtails' with four cusps (figures 8(*c*) and (*d*)). For billiards only the case $0 \leqslant \delta < 1$ will be considered. The curves bear some resemblance to certain cycloids but do not appear to be identical with any of the classical ovals, despite the simplicity of their parametric equations (37).

Let us begin studying the intricate phase space structure of the map generated by oval billiards by considering the two shortest closed orbits. Just as for ellipses, these consist of diametral bounces ($N =$

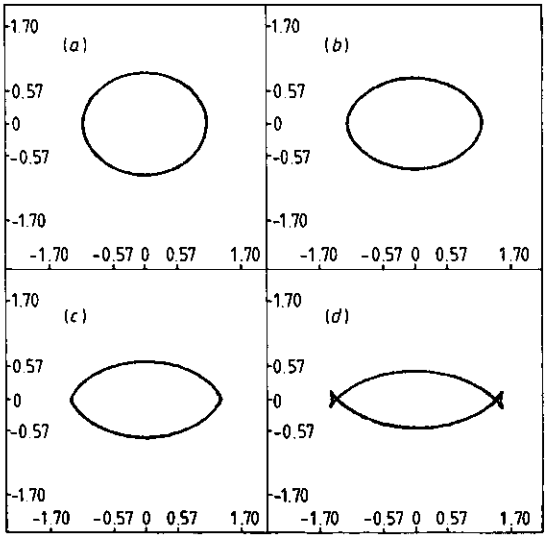

Figure 8 Ovals with (*a*) $\delta = 0.3$, (*b*) $\delta = 0.6$, (*c*) $\delta = 1.0$, (*d*) $\delta = 1.5$.

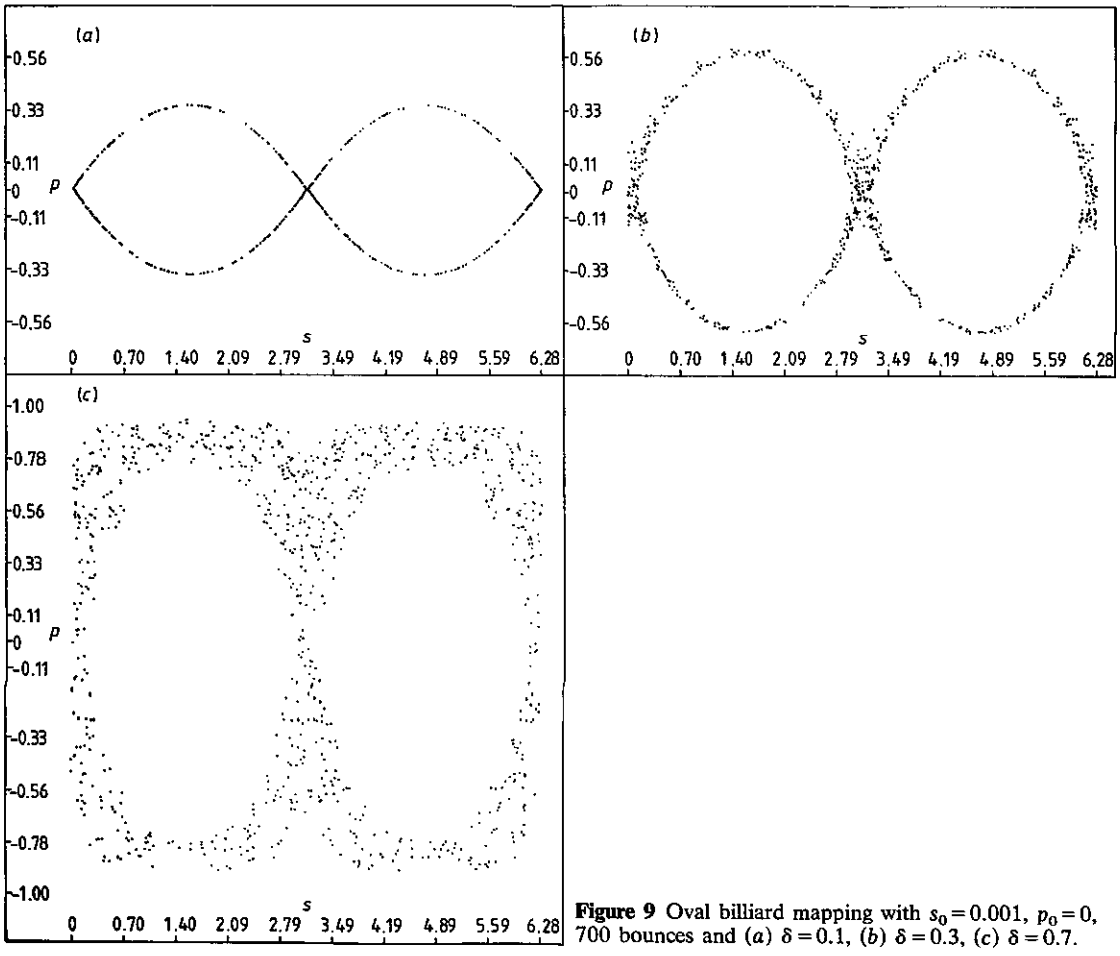

Figure 9 Oval billiard mapping with $s_0 = 0.001$, $p_0 = 0$, 700 bounces and (*a*) $\delta = 0.1$, (*b*) $\delta = 0.3$, (*c*) $\delta = 0.7$.

2) along the long and short diameters. And just as for ellipses, the 'short' orbit is stable and the 'long' orbit is unstable, for all deformations δ. This follows from the criterion (20) together with (36) (for R) and (39) and (40) (for ρ). Moreover, just as for ellipses the stable orbit is surrounded by smooth invariant curves (cf figure 7). However, motion near the unstable orbits reveals that not all the s, p plane is covered with invariant curves. To illustrate this, figure 9 shows orbits starting very close to the unstable fixed point at $s_0 = 0$, $p_0 = 0$ and followed through 700 iterations, for $\delta = 0.1$, 0.3 and 0.7. It is clear that these orbits explore chaotic areas containing and linking the unstable points. The areas grow with the deformation δ but remain localised as $n \to \infty$, in contrast to stadia where the whole plane gets filled as $n \to \infty$. (The 'dust patterns' of figure 9 are not quite symmetrical about the s axis because p_n can keep the same sign for many successive bounces before 'leaking' across the s axis through the chaos near a fixed point.)

For ellipses, only these diametral closed orbits were isolated. But ovals display the generic property that all closed orbits—infinitely many—are isolated. As an example, consider the four-bounce anticlockwise circulating orbits. In circles and ellipses these would form a continuous family with neutral stability. In ovals there are only two: one unstable and one stable. Figures 10(a) and (b) show orbits close to these for $\delta = 0.3$, as well as sketches of the orbits themselves. Use of symmetry and equations (38)–(40) show that the coordinates of the orbits are

$$\begin{aligned} &s = 0,\ \pi a \qquad p = \frac{1-\frac{1}{3}\delta}{[2(1+\frac{1}{9}\delta^2)]^{1/2}} \qquad \text{unstable} \\ &s = \tfrac{1}{2}\pi a,\ \tfrac{3}{2}\pi a \qquad p = \frac{1+\frac{1}{3}\delta}{[2(1+\frac{1}{9}\delta^2)]^{1/2}} \qquad \text{unstable} \\ &s = \pm a(\tfrac{1}{4}\pi - \tfrac{1}{2}\delta),\ \pi a \pm a(\tfrac{1}{4}\pi - \tfrac{1}{2}\delta) \\ &p = 1/\sqrt{2} \qquad \text{stable.} \end{aligned} \tag{41}$$

Now refer to figure 10(c). The chaotic area is a magnification of part of figure 10(a), and illustrates the generic feature that chaos surrounds unstable closed orbits. The loop is a magnification of part of figure 10(b), and illustrates the generic feature that

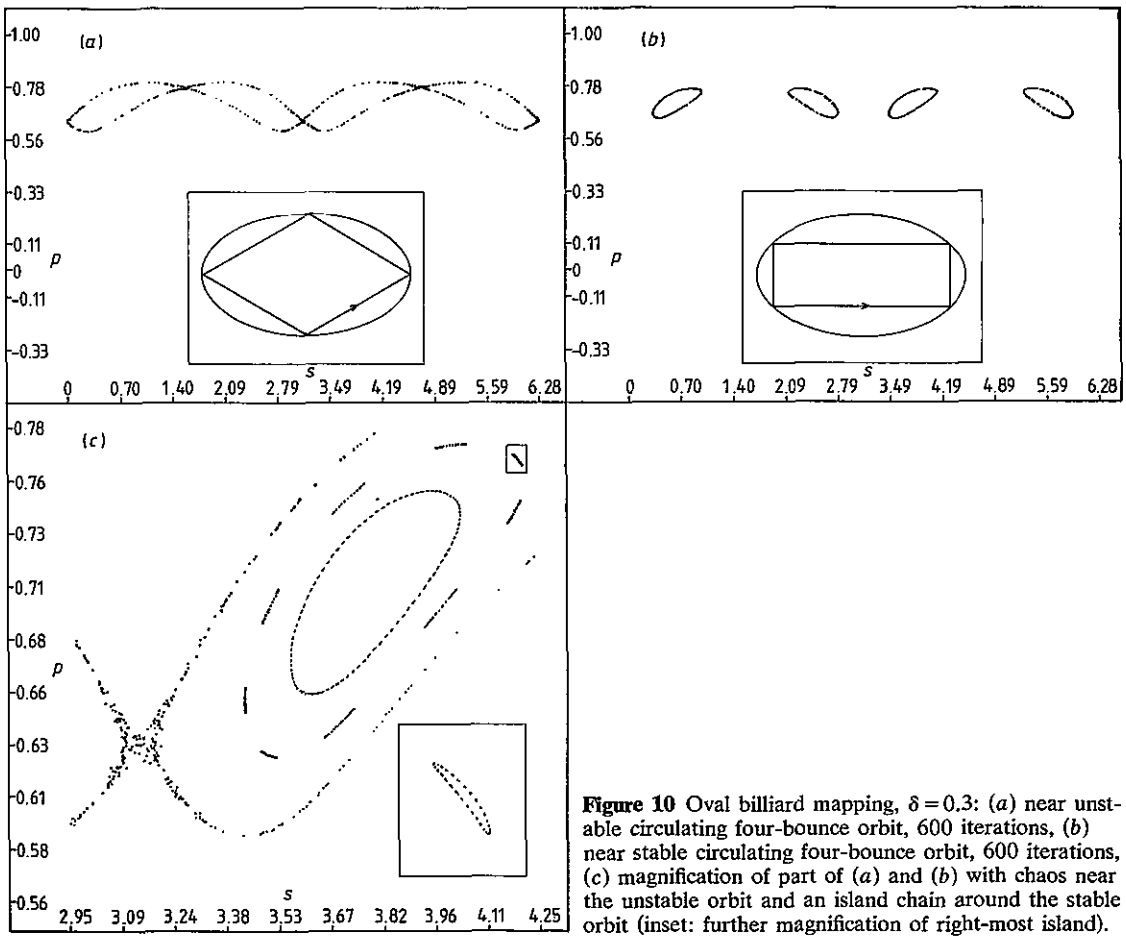

Figure 10 Oval billiard mapping, $\delta = 0.3$: (a) near unstable circulating four-bounce orbit, 600 iterations, (b) near stable circulating four-bounce orbit, 600 iterations, (c) magnification of part of (a) and (b) with chaos near the unstable orbit and an island chain around the stable orbit (inset: further magnification of right-most island).

smooth invariant curves surround stable closed orbits. Around this loop is a chain of nine elongated 'islands' (one of which is shown (inset) under still higher magnification), surrounding part of a much longer closed orbit (with $N = 36$) close to the stable four-bounce orbit. These islands illustrate a further and very important feature: the whole structure (invariant curves around stable fixed points, chaos around unstable fixed points) repeats recursively down to infinitely fine scales. (Resolving the island structure in figure 10(*c*) required high precision: it was necessary to solve the map equation (9) for ψ, to one part in 2^{25}.)

Some idea of the richness of orbital structure for generic billiards, can be got from figure 11, which is a synoptic picture made by combining 25 orbits, each followed for 200 bounces. The orbits of figure 10 are not shown, because their structure would not be fully developed in 200 iterations; but the 'band' where they lie is clearly evident near $p = 1/\sqrt{2}$. In addition, islands surrounding two stable triangular orbits ($N = 3$, circulating clockwise and anticlockwise) can be seen just above and below the central chaos, and elongated islands surrounding a stable period-8 orbit can be seen just outside the large central invariant curves.

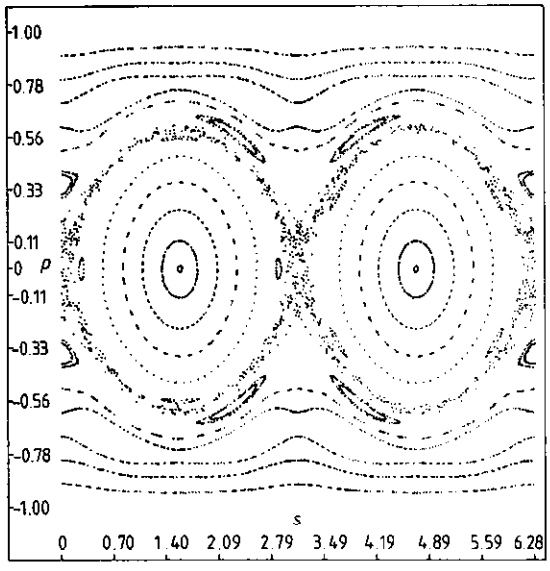

Figure 11 Oval billiard mapping, $\delta = 0.3$; 25 orbits followed through 200 bounces.

The final topic to be illustrated by oval billiards is the birth (or disappearance) of closed orbits as a parameter (in this case δ) varies. Consider the stable two-bounce orbit along the short diameter. Imagine this traversed twice, so that it becomes a diametral orbit with $N = 4$. Its stability, according to §3, depends on the matrix m_4 which is simply the square of m_2 as given by (19). The trace can be calculated to be

$$\mathrm{Tr}\, m_4 = \mathrm{Tr}(m_2^2) = 2 - 16\frac{\rho}{R}\left(2-\frac{\rho}{R}\right)\left(\frac{\rho}{R}-1\right)^2. \quad (42)$$

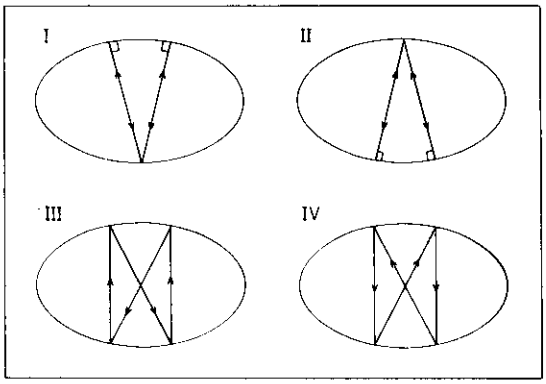

Figure 12 Oval billiards: four-bounce closed orbits near the short diameter; I and II are stable, III and IV unstable.

Equations (36) (39) and (40) give

$$\frac{\rho}{R} = \frac{2(1-\frac{1}{3}\delta)}{(1+\delta)} \quad (43)$$

from which it follows that, when $0 \leqslant \delta \leqslant 1$, $|\mathrm{Tr}\, m_4|$ never exceeds 2 so that (cf equation (16)) the orbit with $N = 4$ is never unstable. But it becomes *neutrally stable* ($\mathrm{Tr}\, m_4 = 2$) when $\rho = R$, i.e. when

$$\delta = \tfrac{3}{5}. \quad (44)$$

What does this mean?

What it means is that as δ increases through $\frac{3}{5}$, four new closed orbits with $N = 4$ split off from the basic diametral orbit. Their topologies are shown in figure 12, and should be contrasted with the four-bounce orbits in figures 10(*a*) and (*b*). It is left as an exercise for readers to show that when $\delta - \frac{3}{5}$ is small the orbits' coordinates are

$$\begin{aligned}
&\text{I} && s \approx \tfrac{1}{4}L \pm \Delta s, \quad p = 0 && s = \tfrac{3}{4}L, \quad p \approx \pm\Delta p \\
&\text{II} && s = \tfrac{1}{4}L, \quad p \approx \pm\Delta p && s \approx \tfrac{1}{4}L \pm \Delta s, \quad p = 0 \\
&\text{III} && s \approx \tfrac{1}{4}L \pm \tfrac{1}{\sqrt{3}}\Delta s, \quad p \approx \mp\tfrac{1}{\sqrt{3}}\Delta p && \\
& && s \approx \tfrac{3}{4}L \pm \tfrac{1}{3}\Delta s, \quad p \approx \pm\tfrac{1}{\sqrt{3}}\Delta p && \\
&\text{IV} && s \approx \tfrac{1}{4}L \pm \tfrac{1}{\sqrt{3}}\Delta s, \quad p \approx \pm\tfrac{1}{\sqrt{3}}\Delta p && \\
& && s \approx \tfrac{3}{4}L \pm \tfrac{1}{\sqrt{3}}\Delta s, \quad p \approx \mp\tfrac{1}{\sqrt{3}}\Delta p &&
\end{aligned} \quad (45)$$

where

$$\Delta s \equiv 8a\left(\frac{\delta - \frac{3}{5}}{6}\right)^{1/2} \qquad \Delta p = \frac{5\Delta s}{8a}. \quad (46)$$

Orbits I and II are stable, III and IV unstable.

The phenomenon just described is a particular type of 'bifurcation', which because of the ovals' high symmetry is different from the usual case where a stable orbit becomes unstable whilst 'emitting' two new stable orbits.

Figure 13 is a synoptic picture of the phase plane for $\delta = 0.65$, showing islands surrounding the stable orbit of type I, just discussed. The large chaotic area has grown out of the unstable diametral two-bounce orbit (cf figure 9). This area is bounded not only by invariant curves surrounding the stable diametral two-bounce orbit but also by invariant curves near $p = \pm 1$, indicating the existence of near-grazing orbits that circulate eternally anticlockwise or clockwise. The chaotic area is not uniformly filled. For example, the 'holes' near the stable four-bounce circulating orbit of figure 10(*b*) are clearly visible.

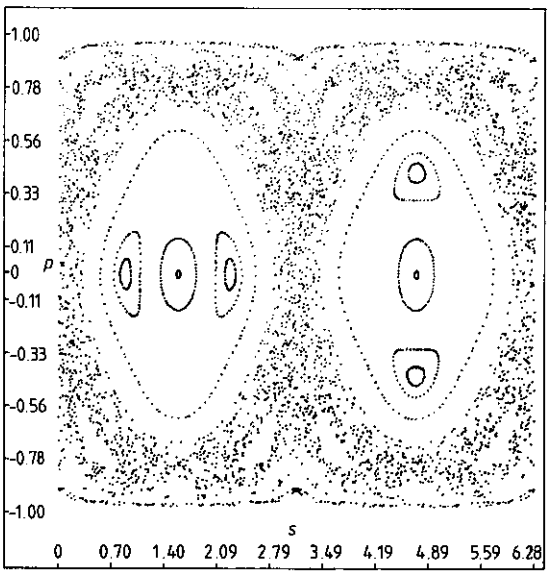

Figure 13 Oval billiard mapping, $\delta = 0.6$: 25 orbits followed through 200 bounces.

8. Brief literature guide

Several review articles describing the recent developments in mechanics but not primarily devoted to billiards have been written by physicists. A very readable elementary article by Whiteman (1977) gives a list of applications in physics, including plasma physics, celestial mechanics and statistical mechanics. Ford (1975) emphasises the connection with statistical mechanics, Berry (1978) emphasises the intermixing of regularity and chaos and describes the application to the gaps in Saturn's rings and the asteroid belt, Treve (1978) emphasises the structure of area-preserving mappings and describes applications in plasma physics, and Helleman (1980) emphasises the bifurcation of closed orbits.

Underlying the whole subject is a great deal of mathematics, lucidly presented in the book by Arnol'd (1978), and explored in detail by Abraham and Marsden (1978).

Concerning billiards, the proof that stadia are ergodic was given by Bunimovich (1974, 1979), employing concepts developed by Sinai (1970, 1979). Joyce (1975) gives a fascinating application of ergodic billiard theory to auditorium acoustics. The proof that billiards for which B is sufficiently smooth (e.g. ovals) are not ergodic (i.e. part of phase space is filled with invariant curves) was given by Lazutkin (1973); his proof required $R(\psi)$ to possess 553 continuous derivatives, but a much smaller number is probably sufficient! An early paper giving a clear account of billiard geometry is that by Poritsky (1950). Statistical properties of stadium billiards, as well as a class of ovals (different from those discussed here) which interpolate between circles and stadia and exhibit generic behaviour, are described by Benettin and Strelcyn (1978).

I have not discussed the delicate case of polygon billiards, where the absence of focusing or defocusing curved boundaries means that all closed orbits have neutral stability. The problems involved with billiards of this type are considered by Zemlyakov and Katok (1975), Hobson (1975) and Richens and Berry (1981).

Acknowledgements

I thank the Institute of Theoretical Physics at the University of Utrecht for hospitality and generous provision of computer facilities, and D Berry for assistance with computations. This work was not supported by any military agency.

Appendix 1

To show that the billiard mapping M is area preserving, it is necessary to evaluate the derivatives in (4). Referring to figure 14, it is clear that in consequence of small initial deviations δs_0, $\delta\alpha_0$ the deviation δs_1 is given by

$$\delta s_0 \sin\alpha_0 + \delta s_1 \sin\alpha_1 = \rho_{01}(\delta\alpha_0 + \delta\psi_0) \qquad (A.1)$$

where ρ_{01} is the length of the chord between s_0 and s_1, and the angles are related by (cf equation (10))

$$\delta\alpha_0 + \delta\psi_0 = \delta\psi_1 - \delta\alpha_0. \qquad (A.2)$$

To obtain these relations in terms of s and ρ, equations (1) and (2) are invoked, to give, after a little algebra,

$$\begin{pmatrix}\delta s_1\\ \delta p_1\end{pmatrix} = \begin{pmatrix}\partial s_1/\partial s_0 & \partial s_1/\partial p_0\\ \partial p_1/\partial s_0 & \partial p_1/\partial p_0\end{pmatrix}\begin{pmatrix}\delta s_0\\ \delta p_0\end{pmatrix} \equiv m_{1,0}\begin{pmatrix}\delta s_0\\ \delta p_0\end{pmatrix}. \qquad (A.3)$$

where the 'deviation matrix' $m_{1,0}$ is

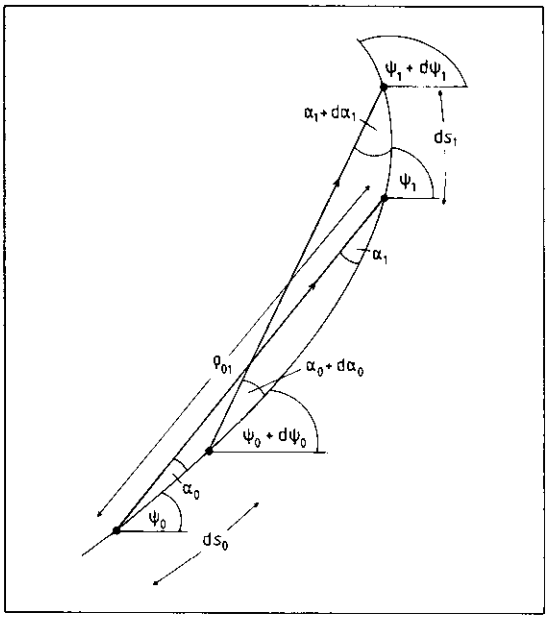

Figure 14 Geometry of deviations

$$m_{1,0}=\begin{cases} -\dfrac{\sin\alpha_0}{\sin\alpha_1}+\dfrac{\rho_{01}}{\sin\alpha_1 R(\psi_0)} & -\dfrac{\rho_{01}}{\sin\alpha_0\sin\alpha_1} \\ -\dfrac{\rho_{01}}{R(\psi_0)R(\psi_1)}+\dfrac{\sin\alpha_1}{R(\psi_0)}+\dfrac{\sin\alpha_0}{R(\psi_1)} & -\dfrac{\sin\alpha_1}{\sin\alpha_0}+\dfrac{\rho_{01}}{\sin\alpha_0 R(\psi_1)} \end{cases}. \quad (A.4)$$

This matrix has determinant unity, so that equation (4) is satisfied.

After N bounces, deviations accumulate by successive multiplication of matrices of the form (A.4). In particular, for an N-bounce closed orbit the matrix m_N defined by equation (11) is given in terms of the bounce geometry by

$$m_N = m_{0,N-1}m_{N-1,N-2}\cdots m_{3,2}m_{2,1}m_{1,0}. \quad (A.5)$$

For the special case of a diametral two-bounce orbit, (19) follows on substituting $\alpha_0=\alpha_1=\frac{1}{2}\pi$, $\rho_{01}=\rho$, $R(\psi_0)=R(\psi_1)=R$.

Appendix 2

I shall estimate the average number of iterations $N(\delta p)$ before a stadium orbit enters the phase plane regions within δp of the two s-axis line segments corresponding to the non-isolated neutrally stable short diametral closed orbits (figure 4). These two regions have area $4\eta\delta p$, and the whole phase plane has area $2L = 2(2\pi R+2\eta)$. By the ergodic property of the motion, the fraction of iterations for which the point s_n, p_n lies in one of the regions is

$$f=\frac{4\eta\delta p}{2L}=\frac{\eta\delta p}{\pi R+\eta}. \quad (A.6)$$

However, once inside this region the orbit will resonate on a zigzag path for $\eta/2R\delta p$ bounces before emerging. Therefore the average number of bounces between such resonances is

$$N(\delta p)=\frac{1}{f}\times\frac{\eta}{2R\delta p}=\frac{(\pi+\eta/R)}{2(\delta p)^2}. \quad (A.7)$$

For the stadium of figure 5(c), $\eta/R=1$ and $N=5000$, so that this theory predicts $\delta p\sim 1/50$ for the half-width of the 'excluded' regions. By measurement, δp is about 1/30 so that the theory gives the right order of magnitude for the effect.

It is probable that this 'repulsion' by non-isolated closed orbits is a general phenomenon, greatly slowing down the exploration of the phase plane in ergodic systems. The effect will be stronger if there are more families of non-isolated closed orbits.

References

Abraham R and Marsden J E 1978 *Foundations of Mechanics* (Menlo Park Ca: Benjamin/Cummings)

Arnol'd V I 1978 *Mathematical Methods of Classical Mechanics* (New York: Springer)

Benettin G and Strelcyn J-M 1978 *Phys. Rev.* A **17** 773–85

Berry M V 1978 in *Am. Inst. Phys. Conf. Proc.* No 46 ed S Jorna pp 16–120

Bunimovich L A 1974 *Funct. Anal. Appl.* **8** 254–5

—— 1979 *Commun. Math. Phys.* **65** 259–312

Ford J 1975 in *Fundamental Problems in Statistical Mechanics* vol 3, ed E G D Cohen (Amsterdam: North-Holland) pp 215–55

Helleman R H G 1980 in *Fundamental Problems in Statistical Mechanics* vol 5, ed E G D Cohen (Amsterdam: North-Holland) pp 165–233

Hobson A 1975 *J. Math. Phys.* **16** 2210–4

Joyce W B 1975 *J. Acoust. Soc. Am.* **58** 643–55

Lazutkin V F 1973 *Math. Izv. USSR* **37** 186–216

Poritsky H 1950 *Ann. Math.* **51** 446

Richens P J and Berry M V 1981 *Physica* D in press

Sinai Ya G 1970 *Russ. Math. Surv.* **25** No 2 137–87

—— 1979 in *Recent Advances in Statistical Mechanics* (Bucharest: Central Inst. Phys. Bucharest, Romania) pp 109–47

Treve Y M 1978 in *Am. Inst. Phys. Conf. Proc.* No 46 ed S Jorna 147–220

Whiteman K 1977 *Reps. Prog. Phys* **40** 1033–69

Zemlyakov A N and Katok A B 1975 *Math. Notes* **18** No 1–2 760–4